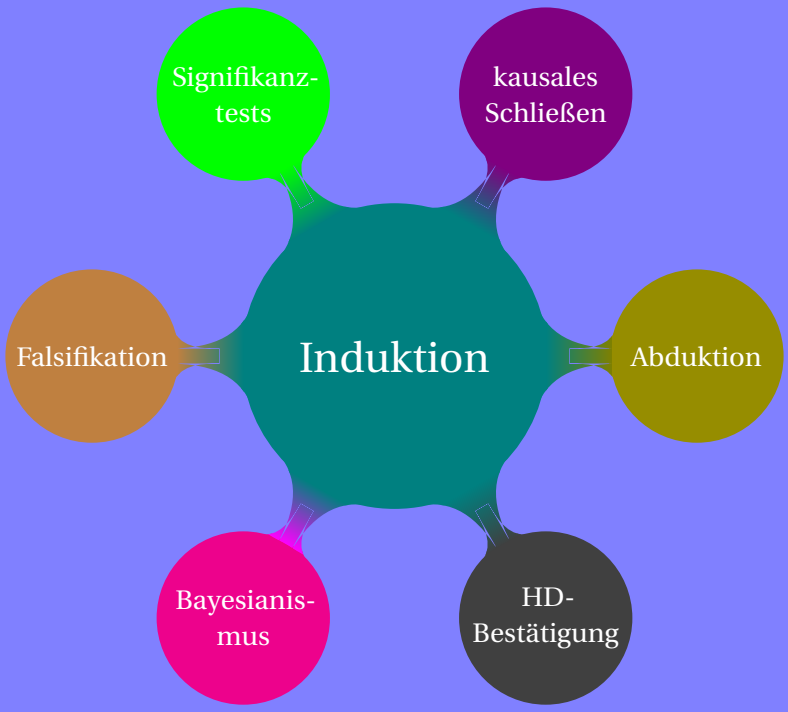
Die erkenntnistheoretischen Grundlagen induktiven Schließens



Thomas Bartelborth

Die erkenntnistheoretischen Grundlagen induktiven Schließens

Thomas Bartelborth

Die erkenntnistheoretischen Grundlagen induktiven Schließens

von Thomas Bartelborth (Leipzig)
(zweite überarbeitete und stark erweiterte Ausgabe von 2017)
URN: urn:nbn:de:bsz:15-qucosa-220168

Bemerkung zum Ziel und Format des Buches

Das vorliegende Buch stellt eine überarbeitete und deutlich erweiterte zweite Ausgabe meines gleichnamigen Buches von 2012 dar. Es wendet sich in Form eines Lehrbuchs sowohl an Anfänger wie Fortgeschrittene der Wissenschaftstheorie sowie an Wissenschaftler, die sich dafür interessieren, wann Daten eine bestimmte Theorie begründen und wie stark die Bestätigung der Theorie durch die Daten ist. Insbesondere das Kapitel 6 über Signifikanztests und ihre Probleme wurde stark erweitert.

Im Vordergrund steht dabei immer die erkenntnistheoretische Frage, ob bestimmte Begründungsverfahren die Ziele der Wissenschaften in überzeugender Weise verfolgen oder ob es dagegen substantielle Einwände gibt. Leider wird sich herausstellen, dass kein Verfahren ohne Fehl und Tadel ist, und wir sollten die Schwächen unserer Begründungsverfahren genau kennen, um sie korrekt einsetzen zu können.

Das Buch ist zunächst als Ebook im PDF-Format erhältlich, um so unentgeltlich und direkt für Lehrende und Lernende zur Einführung in das Thema Induktion und Wahrscheinlichkeit zur Verfügung zu stehen. Das Buch ist dafür optimiert, zwei Seiten auf einen A4-Bogen drucken zu können. Dazu müssen Sie z.B. im Acrobat Reader die Seiteneinrichtung auf Querformat einstellen und dann mehrere Seiten pro Blatt einstellen. Außerdem ist noch unter Seitenanzeige die Option »Deckblatt in Zweiseitenansicht einblenden« zu aktivieren, damit die richtigen Doppelseiten zusammenkommen. Nutzt man noch die Rückseiten, lässt sich das Buch dann platzsparend und trotzdem gut lesbar ausdrucken. Wer gerne mehr Platz zum Kommentieren hat, kann natürlich auch einzelne Seiten auf einen A4-Bogen drucken und dabei sogar mit Vergrößerungen arbeiten, die das Lesen erleichtern können. Damit die Seiten gut lesbar bleiben, sind die Seitenränder für innen und außen symmetrisch gestaltet.

Das Buch hat einen eher knappen Index, da Sie die Datei direkt nach Stichworten durchsuchen können, und es über ein ausführliches Inhaltsverzeichnis verfügt, dessen Einträge auch über die Lesezeichen

in der PDF-Datei wiederzufinden sind. Ich selbst nutze es meist direkt in elektronischer Form und lese es überwiegend am Bildschirm. Auch dafür scheint mir das Format gut geeignet zu sein. Für Zitate verweisen Sie bitte auf die URN-Adresse: Bartelborth, Thomas (2017): *Die erkenntnistheoretischen Grundlagen induktiven Schließens*, 2. Ausgabe, Ebook, <URN: urn:nbn:de:bsz:15-qucosa-220168>. Für alle Hinweise und Rückmeldungen zum Buch bin ich sehr dankbar (bitte einfach per email an: bartelbo@uni-leipzig.de).

Vorwort zur zweiten Ausgabe

Die Philosophie beschäftigt sich schon seit ihren Anfängen u.a. mit Fragen der folgenden Art: Wie können wir zu gut *begründeten* Erkenntnissen über die Welt gelangen und wodurch zeichnen sich *gute* Begründungen vor allem in der Wissenschaft aus? Daran anschließend können wir fragen: Wie können wir anhand wissenschaftlicher Theorien zuverlässig auf zukünftige Ereignisse schließen? Speziell für wissenschaftliche Theorien spricht man davon, dass wir sie *induktiv* erschließen oder begründen müssen, und fragt dazu nach den geeigneten *induktiven Schlussverfahren*, die das leisten können.

Leider herrscht weder unter Philosophen noch unter Fachwissenschaftlern Einigkeit darüber, welche das sind. Es konkurrieren viele Ansätze um den Titel des *besten wissenschaftlichen Begründungsverfahrens*, die jeweils unterschiedliche Vor- und Nachteile aufweisen und sich für unterschiedliche Anwendungsbereiche anbieten. Die wichtigsten Verfahren sollen in diesem Buch kritisch unter die Lupe genommen werden, wobei keines ohne größere Einwände davon kommt. In der Philosophie wird seit einigen Jahren vor allem der Bayesianismus als die Lösung der Bestätigungsfrage vor allem in der Erkenntnis- und Wissenschaftstheorie propagiert, der u.a. aus diesem Grund eine zentrale Stellung im Buch einnimmt. Außerdem stellt er den umfassendsten Einsatz von Wahrscheinlichkeiten in der Explikation induktiven Schließens dar, von dem aus andere probabilistische Ansätze wie die klassische Statistik diskutiert werden können.

Neben diesen quantitativen Verfahren lassen sich die Grundideen induktiven Schließens bereits in einigen *qualitativen Verfahren* erkennen, und dort finden wir auch schon die meisten der Probleme, mit denen fast alle anderen Verfahren ebenfalls zu kämpfen haben. Die drehen sich etwa um auftretende Paradoxien, irrelevante Schlüsse, die Induktionseigenschaft, die Unterbestimmtheit von Theorien, die Objektivität der Schlussverfahren und immer wieder um spezielle Beispiele, in denen die

Verfahren nicht das leisten, was wir uns von ihnen versprechen. Zu den qualitativen Verfahren gehören u.a. die konservative Induktion (bzw. das einfache Extrapolieren) zusammen mit ihren statistischen Varianten, die hypothetisch-deduktive Theorienbestätigung, falsifikationistische Ansätze, die Instanzenbestätigung, die eliminative Induktion und der Schluss auf die beste Erklärung (auch »Abduktion« genannt).

Die induktiven Begründungen wissenschaftlicher Theorien dienen vor allem der *Auswahl einer Theorie*, die wir dann etwa als Grundlage für unsere Handlungen akzeptieren. Gerade das abduktive Schließen wird sich dabei als eines der besten Verfahren zur Theorienwahl herausstellen, jedenfalls dann, wenn wir es in ein holistisches Konzept von Erklärungskohärenz einbetten, das bereits gewisse quantitative Präzisierungen erlaubt. Die meisten Verfahren werden gern als einfache Regeln der Theorienwahl dargestellt, aber letztlich sollten wir hier keine falschen Hoffnungen hegen, denn es gilt weiterhin Thomas Kuhns Einsicht, dass es *keinen Algorithmus* zur Auswahl einer Theorie gibt, sondern immer komplexe Abwägungsprozesse dazu erforderlich sind. Am nächsten kommt vermutlich der bayesianische Ansatz der Idee eines umfassenden Algorithmus, aber wir werden sehen, dass dabei wichtige Aspekte der Theorienwahl außen vor bleiben, wie z.B. die Erklärungskraft der Theorien. Die kuhnsche Einsicht führt uns aber keineswegs in einen radikalen Relativismus, denn der entstünde erst, wenn wir uns auch noch der Auffassung anschlössen, dass es keine klaren Ziele für die wissenschaftliche Arbeit gäbe. Doch dem ist nicht so und ich werde in Kapitel 2 die objektiven erkenntnistheoretischen Ziele der Wissenschaft genauer beschreiben.

Der bayesianische Ansatz trumpft vor allem damit auf, dass er für alle Anwendungsfälle eine quantitative Auskunft geben kann, in welchem Umfang bestimmte Daten eine Theorie stützen. Die Kritiker wenden allerdings ein, das gehe auf Kosten der Objektivität, und es sei im Bayesianismus keineswegs klar, welche Größen als Maß für die Bestätigung herangezogen werden sollten. Tatsächlich wird es sich auch in diesem Buch immer wieder erweisen, dass es keine so einfachen (quasi algorithmischen) Induktionsverfahren gibt, die eine zuverlässige Entscheidung liefern, ob wir eine Theorie akzeptieren sollten oder nicht – auch nicht im Bayesianismus.

Das gilt übrigens ebenso für die klassische Statistik, die manchmal so dargestellt wird, als ob sie ein solches Patentrezept für die Auswahl einer Theorie zu bieten hätte, und von den Anwendern gerne in diesem Sinne genutzt wird. Wir werden sehen, dass hier gleich zwei Formen von Fehlschlüssen drohen, die danach verlangen, die klassische Statistik zu ergänzen.

Außerdem werde ich u.a. dafür argumentieren, dass wir für induktive Schlüsse immer auf weitergehende (metaphysische) Annahmen über die Kausalstruktur unserer Welt sowie ein komplexeres Abwägen unterschiedlicher Kriterien angewiesen sind, die allen einfachen Rezepten einen Strich durch die Rechnung machen.

Wir kommen daher leider nicht umhin, uns zunächst den Mühen einer substantiellen Theorienbildung zu unterziehen. Erst diese Theorien können schließlich empirisch getestet werden. Gelingt das, sind wir auf ihrer Grundlage zu *begründeten* Vorhersagen in der Lage, die wir spätestens dann benötigen, wenn wir wichtige Entscheidungen darauf stützen möchten. Die substantiellen Theorien oder Modelle lassen sich nur anhand komplexer Abwägungen ihrer Zusammenhänge zu unseren Daten und anderen Theorien akzeptieren oder zurückweisen. Unsere Hilfsmittel, um kausales Wissen dabei einzubringen, sind im Bayesianismus etwa bayessche Netze und i.A. besondere kausale Schlussverfahren, die in den letzten Jahren entwickelt wurden. Dazu werden logische (boolesche) und probabilistische Verfahren vorgestellt, die diese Aufgabe übernehmen können.

Die substantielle Theorienbildung ist noch aus einem anderen Grund geboten, der in der ganzen Debatte leider oft übersehen wird. Wir suchen nach gut begründbaren neuen Theorien, weil wir uns von ihnen Ratschläge für wichtige Entscheidungen im Alltag, der Politik oder der Wissenschaft erhoffen. Dazu müssen die Theorien informative Vorhersagen gestatten. Es geht uns also nicht nur darum, irgendwelche Theorien zu bestätigen, sondern vielmehr um eine Auswahl besonders gehaltvoller Theorien. Dadurch wird das induktive Schließen überhaupt erst spannend, denn empirisch inhaltsarme Theorien über einfache mathematische Zusammenhänge oder direkt beobachtbare Tatsachen sind dagegen noch relativ leicht zu gewinnen. Dieses Spannungsverhältnis

zwischen gut begründeten, aber zugleich empirisch anspruchsvollen Behauptungen wird uns das ganze Buch über begleiten.

Es zeigen sich aber trotz aller (oft heftigen) Kontroversen auch bestimmte Gemeinsamkeiten zwischen den Verfahren. Von den qualitativen Verfahren bis hin zu den quantitativen Schlussverfahren werden uns in unterschiedlichen Variationen die sogenannten *Likelihoods* der Theorien bzgl. der Daten als ein wichtiger Maßstab für die Qualität einer Begründung der Theorien durch die Daten verfolgen. (Die Likelihood P(E|H) ist dabei die Wahrscheinlichkeit, die ein Datum E hat, wenn wir die Hypothese H einmal als wahr annehmen und auf der Basis von H die Wahrscheinlichkeit von E berechnen.) Ihre genaue Aufgabe wird am deutlichsten im abduktiven Schließen und in den bayesianischen Ansätzen, die der induktiven Logik nahestehen.

Ein Hauptziel des Buches ist es dabei, die erkenntnistheoretischen Fragestellungen und Prinzipien, auf die wir uns stützen, immer wieder in den Vordergrund zu stellen und so die Erkenntnistheorie und die Wissenschaftstheorie wieder enger zusammenzubringen. Gerade in den Problembereichen induktiven Schließens müssen wir uns auf erkenntnistheoretische Prinzipien besinnen, um unsere Lösungsvorschläge begründen zu können. Dazu ist es von großer Bedeutung, die Ziele der Wissenschaft explizit vorzustellen und die Induktionsverfahren daraufhin zu untersuchen, inwieweit sie denen gerecht werden können.

Das geschieht im Rahmen der Ausarbeitung einer bestimmten Konzeption *wissenschaftlichen Wissens*, die uns als Leitfaden für die Induktionsverfahren gelten soll. Vorrang wird in den folgenden Debatten dabei immer eine Analyse der zugrundeliegenden Ideen vor einer Analyse der eher technischen Zusammenhänge und Beweise haben. Dadurch soll das Buch als einführende und ergänzende Lektüre für alle diejenigen dienen, die induktive Schlussverfahren nicht nur in Form von »Kochrezepten« oder als komplizierte mathematische Verfahren kennenlernen wollen, sondern vor allem die Begründungen und Grenzen bestimmter Verfahren ermitteln möchten. Es steht also immer die erkenntnistheoretische Bewertung der Zusammenhänge im Vordergrund gegenüber den technischen Umsetzungen, die nur so weit verfolgt werden, wie das für ein Verstehen sowie eine Bewertung erforderlich erscheint.

Inhaltsverzeichnis

1 Probleme induktiven Schließens

Zu den wichtigsten methodischen Fragen für die (empirischen) Wissenschaften gehört die, wann wir eine bestimmte Theorie oder Hypothese (als hinreichend [empirisch] bestätigt) akzeptieren sollen. Welcher Art müssen unsere Daten sein und wie viele benötigen wir? Welcher Art ist die Bestätigungsbeziehung bzw. wie gut passen jeweils Theorien und Daten zusammen und welche Rolle spielt unser weiteres Hintergrundwissen dabei? Dieses Buch soll dazu dienen, diese Fragen weiter zu klären. Trotz einer boomenden Wissenschaft und zahlreichen Debatten zum Thema der Rechtfertigung wissenschaftlicher Behauptungen gibt es leider immer noch grundlegende Meinungsverschiedenheiten um die Frage, nach welchen Methoden oder Standards wir beim Akzeptieren von Hypothesen in der Wissenschaft verfahren sollten. Es wird also zu klären sein, was die Kriterien für gute wissenschaftliche Erkenntnisse sind, und wir werden unterschiedliche Induktionsverfahren daraufhin anschauen, inwiefern sie das Potential haben, die besten wissenschaftlichen Theorien zu finden.

Dabei soll die Frage beantwortet werden, wann eine Behauptung oder spezieller eine wissenschaftliche Theorie als *gut begründet* gelten darf und ob sich die Stärke der Begründung vielleicht sogar *quantifizieren* lässt. Solche Begründungen oder auch Schlussverfahren werden alle unter dem Stichwort der *Induktion* bzw. dem des *induktiven Schließen*s diskutiert. Dazu stelle ich konkurrierende Ansätze vor, die heutzutage als beste Verfahren für diese Aufgabe der induktiven Bestätigung von Theorien vertreten werden. Leider gibt es hier – anders als im Bereich der deduktiven Logik – kein zentrales Paradigma (wie eine induktive Logik), dem sich alle Ansätze unterordnen würden. Vielmehr stehen diese Induktionsverfahren seit Jahrzehnten in erbitterter Konkurrenz zueinander, und es ist sehr schwer, den Überblick zu behalten oder einen Sieger auszumachen. Ich werde dafür argumentieren, dass der *Schluss auf die beste Erklärung* zusammen mit einer *Kohärenztheorie*

der epistemischen Rechtfertigung den übergeordneten Rahmen abgeben kann, innerhalb dessen sich die anderen Ansätze anordnen und ihre Verdienste und Probleme bewerten lassen.

Ausgangspunkt ist also das Grundproblem der Erkenntnistheorie, wann wir über eine *gute Begründung* für eine Meinung verfügen, d.h., was *gute Hinweise auf die Wahrheit* einer bestimmten Überzeugung sind und wann sie vorliegen. Das ist auch die zentrale Frage für die Annahme wissenschaftlicher Theorien bzw. Hypothesen. Wenn wir in der Wissenschaft eine Theorie akzeptieren, so muss sie durch Beobachtungen und die Resultate von Experimenten gut begründet sein und darf keinesfalls willkürlich akzeptiert werden. Dabei betrachten wir hier die Frage der Datenerhebung meist schon als geklärt und fragen uns nur noch, welche Daten in welchem Maß für welche Theorien sprechen, wobei ich mit *Daten* hier manchmal Rohdaten meine, die direkt unsere Beobachtungen wiedergeben, aber manchmal auch bereits stärker interpretierte Beobachtungen. Solche abgeleiteten Daten setzen meist schon erste Induktionsschlüsse voraus.

Wir werden daneben darauf zu sprechen kommen, dass zum Akzeptieren von Theorien mehr gehört als gute Begründungen. Das ganze Unternehmen Wissenschaft wird erst dadurch spannend, dass wir von den Theorien bestimmte Leistungen erwarten. Wahre Theorien, die uns nicht viel über die Welt sagen, sind relativ leicht zu finden und zu begründen. Man denke z.B. an viele relativ triviale Aussagen wie: (1) Zwei Liter Wasser sind mehr als ein Liter Wasser, (2) 2+3 = 5 oder (3) vor mir steht ein weißer Tisch etc. Stattdessen suchen wir aber nach Theorien, die gehaltvoll und informativ sind (uns möglichst viele Phänomene erklären können) und trotzdem noch gut begründbar sind. Das ist schon schwerer zu erzielen und beschreibt erst die anspruchsvolle Aufgabe, die uns in der Wissenschaft erwartet. Unsere Induktionsverfahren sollen es gerade gestatten, solche Aufgaben der Theorienwahl durchzuführen und nicht nur nach irgendwelchen leicht zu begründenden Behauptungen Ausschau zu halten.

Die besseren tatsächlichen wissenschaftlichen Theorien sind in der Regel gehaltvoll und gehen somit deutlich über den Rahmen unserer Beobachtungen hinaus. Sie behaupten etwa, dass an allen Orten im Universum zu allen Zeiten ein bestimmtes Gravitationsgesetz gilt. Oder

sie postulieren unbeobachtbare Teilchen, die wir bestenfalls mühsam indirekt erschließen können wie Neutrinos. An diesen Stellen wird allerdings die Frage besonders brisant: Wie gut sind unsere Gründe für derartig gehaltvolle Annahmen? Empiristen und speziell Bayesianer scheinen diese Problematik manchmal aus den Augen zu verlieren, wenn sie nur danach fragen, welche unserer Theorien die wahrscheinlichste bzw. die am besten begründete ist, ohne dabei in Rechnung zu stellen, wie gehaltvoll oder erklärungsstark die Theorien jeweils sind. Popper scheint es auf der anderen Seite oft nur um den Gehalt der Theorien und nicht mehr um ihre Begründbarkeit oder Begründetheit durch empirische Daten zu gehen, zumal er induktive Begründungen generell ablehnt. Doch dazu später mehr.

Im Folgenden unterscheide ich nicht strikt zwischen induktiven *Schlüssen* und induktiven *Rechtfertigungen*. Man kann es so sehen, dass bei *Schlüssen* aus vorliegenden Daten unser Schlussverfahren die Hypothesen erst erzeugt und damit zugleich ein Entdeckungsverfahren für neue Hypothesen darstellt. Doch dem Entdeckungskontext von wissenschaftlichen Hypothesen soll hier nicht viel Raum eingeräumt werden. Er gehört eher in die Psychologie der Wissenschaften. Das Hauptthema in diesem Buch ist dagegen, wie bereits vorliegende Hypothesen gerechtfertigt werden können. Doch dieser Rechtfertigungskontext kann nicht immer strikt vom Entdeckungskontext getrennt werden (die induktiven Schlüsse auf solche neuen Hypothesen dürften zugleich eine erste Rechtfertigung der Hypothesen bieten). Außerdem bietet es sich oftmals an, aktiver vom *Schluss auf Hypothesen* zu sprechen, obwohl man vor allem ihre nachträgliche *Rechtfertigung* meint. Diese ist zumindest das Ziel des Buches, während die Tatsache, dass das abduktive Schließen darüber hinaus manchmal sogar Hypothesen generieren kann, bestenfalls ein Randthema bleiben wird. Daher bitte ich den Leser, meine etwas anschaulichere Redeweise vom *induktiven Schließen* im Zweifelsfall einfach als *induktives Rechtfertigen* zu verstehen (vgl. auch Kap. 2).

Wir grenzen induktive Rechtfertigungen außerdem von *deduktiven Schlüssen* oder *Begründungen* ab. Deduktive Begründungen sind vor allem irrtumssichere Begründungen: Wenn wir ihre Prämissen bereits sicher wissen, dann sind wir uns auch der deduktiven Konklusionen

sicher. Wenn wir also schon sicher wissen, dass alle Raben schwarz sind, dann können wir ebenso sicher sein, dass der nächste Rabe, den wir sehen werden, schwarz ist. Der Schluss selbst ist hier deduktiver Natur. Wenn unsere Ausgangsannahmen wahr sind, dann muss genauso unsere Schlussfolgerung wahr sein. Doch in der Wissenschaft sind wir normalerweise nicht in einer so komfortablen Situation. Wissenschaftliche Aussagen sind häufig schwer nachweisbare Allbehauptungen. Es geht uns z.B. darum, zunächst die Aussage zu begründen, dass tatsächlich alle Raben schwarz sind. Doch schon eine so kleine Theorie wie unsere Rabentheorie lässt keine sicheren empirischen Begründungen zu. Selbst wenn wir sehr viele Raben auf ihre Farbe hin untersuchen, können wir nicht völlig ausschließen, dass irgendwann noch ein weißer Rabe geboren wird und unsere Rabentheorie umstößt. Der Schluss von vielen beobachteten schwarzen Raben auf unsere Rabentheorie ist daher nur ein *induktiver Schluss* bzw. die Rabentheorie wird durch unsere Daten nur induktiv begründet. Sogar wenn wir uns bei den Daten ganz sicher sind, können wir nicht mit entsprechender Sicherheit auf die Rabentheorie schließen. Man sagt auch: Induktive Schlüsse sind *gehaltserweiternd* (im Hinblick auf die Prämissen) und das bringt ein Irrtumsrisiko mit sich, das deduktive Schlüsse nicht aufweisen.

Wissenschaftliche Behauptungen sind vielfach noch deutlich riskanter als unsere kleine Rabentheorie und besagen etwa, dass alle Materie aus unsichtbar kleinen Teilchen zusammengesetzt ist, die sich nach sehr seltsamen Gesetzen richten. Hier ist die epistemische Kluft zwischen den Daten und den entsprechenden Theorien noch einmal deutlich größer. Es handelt sich ganz klar um induktive Begründungen. Die folgenden Beispiele beleuchten dazu weiter, was unter *Induktion* verstanden werden soll.

Übrigens wird selbst das *deduktive Schließen* oft zum Rechtfertigen eingesetzt, statt zur Generierung neuer Annahmen. Wir können aus vorgegebenen Prämissen zahllose deduktive Schlüsse ziehen (etwa mit Hilfe der Oder-Abschwächung), von denen die meisten für uns irrelevant erscheinen. Deshalb ist die Fragestellung in vielen Fällen eher die, ob sich bestimmte Annahmen (die wir schon vorher auf den Tisch legen) dann aus bestimmten Prämissen ableiten bzw. durch diese Prämissen deduktiv

begründen lassen oder nicht. Das hindert uns auch im deduktiven Fall nicht daran, von *Schlüssen* zu reden.

Von besonderer Bedeutung sind in der Wissenschaft *kausale Behauptungen*, und die sind besonders schwer zu begründen. So behaupten Wissenschaftler etwa: »Papilloma-Viren verursachen Gebärmutterhalskrebs« oder »Senkungen der Leitzinsen beleben die Wirtschaft.« Hierbei geht es nicht nur darum zu zeigen, dass Instanzen der genannten Typen von Ereignissen nacheinander auftreten, sondern vor allem darum nachzuweisen, dass die ersteren tatsächlich die zweiteren *herbeigeführt* haben. Das verlangt nach ganz speziellen Daten.

Eine erste einfache Regel für das induktive Schließen (genannt Nicods Kriterium oder Regel: NR) besagt, dass zumindest die *einzelnen Instanzen* einer wissenschaftlichen Allaussage diese Aussage ein Stück weit begründen. Für unsere Rabenhypothese H: »Alle Raben sind schwarz« betrachtet man typischerweise alle *schwarzen Raben* als seine positiven Instanzen. Jeder gefundene schwarze Rabe wäre demnach eine erste induktive Bestätigung (und Raben, die sich nicht als schwarz erweisen, eine Widerlegung) unserer Rabenhypothese.

Doch selbst die simple Regel (NR) ist nicht unumstritten. Betrachten wir z.B. die Hypothese H*: »Alle Menschen sind kleiner als 2,50m.« Stoßen wir nun auf mehrere Menschen in Sibirien, die alle bereits die Größe von 2,48 m erreicht haben, so handelt es sich zwar noch um Instanzen von H*, aber wir würden intuitiv nicht mehr behaupten wollen, dass H* dadurch gestärkt wird. Im Gegenteil, wir würden vermutlich eher annehmen, dass es dann bald auch Menschen mit 2,50 m Größe geben wird.

Hier spielt unser Hintergrundwissen hinein, wonach es höchst wahrscheinlich ist, dass die Grenze für das Größenwachstum beim Menschen nicht scharf definiert ist und wir daher durchaus Menschen mit 2,50m Größe erwarten können, wenn sich erst einmal viele Menschen finden lassen, die nur kurz darunter liegen. Statt also H* zu bestätigen, schwächen *viele* neue Instanzen vom Typ »die Größe ist 2,48m« unsere Hypothese zunehmend. Es zeigt sich hier erstmals, wie unsere einfachen Induktionsverfahren von allgemeineren kausalen Annahmen darüber abhängen, wie unsere Welt funktioniert. Empiristen versuchen den Einfluss solcher nach ihrer Meinung metaphysischen Annahmen immer

wieder zu negieren und suchen weiterhin nach einfachen Induktionsverfahren (vgl. auch Bartelborth 2004). Dieser Einfluss findet sich in vielen Induktionsverfahren bis hin zur klassischen Statistik wieder. Ein wichtiges Thema des Buches wird daher sein, dass wir diese Einflüsse des Hintergrundwissens nicht übersehen dürfen, obwohl sie unsere Induktionsverfahren komplexer machen.

Damit haben wir das erste Problem der Induktion beschrieben, genauer anzugeben, nach welchen Regeln man induktive Schlüsse oder induktive Begründungen vollziehen sollte. Dabei wird in diesem Text der Schluss auf die beste Erklärung eine zentrale Rolle übernehmen. Das andere Problem ist, diese Regeln gegen einen Induktionsskeptiker zu verteidigen. David Hume hat generelle Argumente dafür vorgetragen, dass kein Induktionsverfahren selbst wieder in rationaler Weise gerechtfertigt werden kann. Zum Hauptthema des Buches, nämlich der ersten Frage, werde ich zunächst die einfachste Form der Induktion (die konservative Induktion) schildern und dann kurz zur zweiten Frage das Rechtfertigungsdilemma beschreiben, in dem wir uns befinden. Danach werde ich mich den unterschiedlichen Ansätzen der Induktion widmen und zeigen, inwiefern der Schluss auf die beste Erklärung sie unter einem Dach vereinigt.

1.1 Beispiele für Induktionsschlüsse

Induktionsschlüsse vollziehen wir in allen Bereichen unseres Lebens. Wenn wir in einen Fahrplan schauen, in dem steht, dass unser Zug um 16 Uhr eintreffen soll, dann schließen wir daraus vermutlich darauf, dass der Zug um ca. 16 Uhr eintreffen wird. Die Deutsche Bahn liefert freilich gute Beispiele dafür, dass es sich hierbei nicht um einen deduktiven Schluss handelt. Viele Schlüsse vollziehen wir automatisch und ohne erst darüber nachzudenken. Kommen unsere Gäste nass von draußen herein, werden wir annehmen, dass es draußen regnet. Es sei denn, wir haben weiteres Hintergrundwissen, wonach draußen gerade eben noch strahlender Sonnenschein herrschte und wir wissen, dass unser Nachbar die Funktionsweise seines Rasensprengers noch nicht sehr gut versteht. Sehe ich jemanden auf eine Straßenbahn zulaufen, nehme ich an, dass

er sie noch erreichen möchte etc. Wenn wir genau hinschauen, erkennen wir, wie unser Alltag schon von induktiven Schlüssen durchsetzt ist. Viele können sogar lebenswichtig sein. Wenn jemand im Straßenverkehr blinkt, nehmen wir an, dass er abbiegen wird, und verlassen uns vielleicht darauf. Sagt uns der Arzt, dass unsere Gallensteine unbedingt entfernt werden müssen, vermuten wir, dass durch eine entsprechende Operation unsere Überlebenschancen am größten sein werden usf.

Vor Gericht werden solche Schlüsse meist etwas expliziter gezogen und hoffentlich sorgfältiger begründet. Es gibt gewisse Indizien (Fingerabdrücke, DNA-Spuren, Zeugenaussagen etc.), die auf einen bestimmten Verdächtigen als Täter hinweisen; und wenn die eine bestimmte Stärke erreichen, werden wir den Angeklagten schuldig sprechen. In all diesen Fällen ist klar, dass es sich um nicht-deduktive Schlüsse handelt und dass es auch bei noch so guten Begründungen ein Irrtumsrisiko gibt. Leider ist es oft sehr schwer, das genauer zu beziffern. Auch im Zivilrecht sind natürlich induktive Schlüsse zu ziehen. Aus bestimmten Aussagen oder Verträgen schließen wir etwa darauf, was der Wille der beteiligten Parteien war und welche Art von Vertrag damit zustande gekommen ist.

Am ausführlichsten und genauesten sollten unsere Begründungen in der Wissenschaft ausfallen. Sie ist daher unser Paradebeispiel und die Zielvorstellung für das induktive Schließen. Nichtsdestoweniger gibt es unter Wissenschaftstheoretikern die übereinstimmende Ansicht, dass Wissenschaftler ebenfalls Menschen sind. Das soll bedeuten, dass Wissenschaftler letztlich *dieselben Begründungsverfahren* für intuitiv plausibel halten und darauf vertrauen, die wir auch im Alltag anwenden. Natürlich werden sie diese zu präzisieren versuchen und z.B. die mathematische Statistik zum Einsatz bringen, um so gediegenere Begründungen abzuliefern als der Alltagsmensch, doch die Grundintuitionen sind dieselben. Unsere Induktionsverfahren können wir daher oft an relativ einfachen Beispielen testen, um uns nicht gleich in komplizierten wissenschaftlichen Debatten zu verlieren.

Physiker schließen etwa aus der 3°K Hintergrundstrahlung auf den länger zurückliegenden Urknall, Paläontologen aus Versteinerungen auf die Existenz der Dinosaurier und ihre Eigenschaften, die Medizin aus der Verbreitung bestimmter Krankheiten darauf, dass es sich um Infektionskrankheiten handelt, Biologen aus der Zweckmäßigkeit unserer

Organe auf die dahinterstehende Evolutionsgeschichte, Historiker aus bestimmten Dokumenten auf die Motive der beteiligten Personen der jeweiligen Geschichte, Psychologen aus den Ergebnissen bestimmter Tests auf Charaktereigenschaften bestimmter Versuchspersonen, Soziologen aus bestimmten Eigenschaften von sozialen Institutionen auf deren zugrundeliegende möglicherweise versteckte Zielsetzung usf. Das sind typische Schlüsse auf die beste Erklärung, wie wir sie schon aus dem Alltag kennen. Das Verfahren werden wir uns noch genauer ansehen.

Leider sind wir Menschen nicht besonders geschickt, wenn es um Überlegungen mit Wahrscheinlichkeiten geht. Ein inzwischen recht bekanntes Beispiel belegt sehr schön, wie schwer wir uns dann mit einfachen und im Prinzip überschaubaren Fällen induktiven Schließens tun, nämlich das sogenannte *Ziegenproblem*. Das Beispiel und seine Lösung sind mittlerweile weit verbreitet, aber es verursacht praktisch bei allen, die es noch nicht kennen, große Überraschung. Es zeigt für mich auch wie vorsichtig wir mit unseren Intuitionen in Fällen induktiven Schließens umgehen müssen und am Anfang haben sich sogar Lehrstuhlinhaber der Statistik davon genauso in die Irre führen lassen wie die Laien.

Für die Anwendung der Statistik ist nämlich das größte Problem, dass wir erst ein geeignetes (Urnen-) Modell für eine solche Situation finden müssen. Ganz kurz zu der Geschichte, die man an vielen Stellen nachlesen kann. In einer Gameshow kann der Kandidat eine von drei Türen auswählen. Hinter einer davon steht ein Ferrari (das wurde etwa vor der Show ausgelost) und hinter beiden anderen jeweils eine Ziege. Der Kandidat wählte eine Tür aus (sagen wir Tür 1). Dann tritt der Moderator in Aktion. Seine Aufgabe ist es, von den verbleibenden zwei Türen (hier also 2 und 3) eine Tür aufzumachen, hinter der in jedem Fall eine Ziege steht. Nehmen wir an, er öffnet nun Tür 2. Dann wird der Kandidat vor die Wahl gestellt, nun bei seiner ersten Tür zu bleiben oder zur letzten noch verbliebenen Tür (in unserem Falle Tür 3) zu wechseln.

Die meisten meinen, dass es nun gleichgültig sei, ob wir wechseln oder nicht. Der Ferrari wird hinter einer von beiden Türen sein, aber wir haben keine Hinweise auf eine der beiden Türen und daher hat nun jede eine Wahrscheinlichkeit von 0,5, dass dort der Ferrari zu finden ist. Doch tatsächlich (und das lässt sich sogar empirisch testen) ist die

Wahrscheinlichkeit für den Ferrari hinter Tür 1 nur 1/3, während sie für einen Ferrari hinter Tür 3 inzwischen auf 2/3 angestiegen ist. Der Moderator hat uns mit seiner Wahl von Tür 2 einen Hinweis auf Tür 3 gegeben. Er sagt sozusagen: *Wenn hinter den beiden verbliebenen Türen überhaupt ein Ferrari steht, dann ist er hinter Tür 3 zu finden.* Der Vordersatz dieses konditionalen Hinweises ist immerhin in ca. 2/3 der Fälle erfüllt. Also erhält man bei häufiger Wiederholung des Spiels mit der Wechselstrategie alle Ferraris, die hinter den beiden verbleibenden Türen stehen und das sind im Durchschnitt 2/3 der Ferraris. Die Lösung wird an vielen Stellen (etwa bei Wikipedia) genauer erläutert (vgl. a. Kap. 5.6.1). Mich interessiert hier nur, wie schnell wir einen solchen Fall falsch beurteilen können und wie vorsichtig wir daher beim induktiven Schließen vorgehen müssen, um die richtigen Verfahren dafür zu finden. Wir übersehen in diesem Beispiel gerne, dass wir aus dem Öffnen der Tür 2 durch den Moderator tatsächlich einen verwertbaren, konditionalen Hinweis auf Tür 3 erhalten.

Das mag zunächst genügen, um ein erstes Licht auf die Vielfalt und Allgegenwart unserer Induktionsschlüsse zu werfen. Dabei können – wie schon gesagt – Teile unseres Hintergrundwissens eine wichtige Rolle spielen, die in unseren nur erwähnten Beispielen nicht ausgeführt wurde. Außerdem geht es oft darum, dass nicht nur *ein* Datum unsere Konklusionen begründet, sondern wir (gerade vor Gericht und in der Wissenschaft) über eine Reihe von Indizien für eine Annahme verfügen. Dabei kommen aber sogleich Schwierigkeiten ins Spiel, denn manche Indizien mögen zwar in eine Richtung (etwa gegen den Angeklagten) weisen, aber andere *zugleich* in die Gegenrichtung (sind also entlastend für den Angeklagten). Dann müssen wir eine Gewichtung und Verrechnung der Indizien vornehmen, wofür uns oft präzise Verfahren fehlen.

1.2 Historische Vorläufer der Induktionsdebatte

Historisch gab es zwar schon in der Antike erste Methodenreflexionen, aber der Beginn der systematischeren Erforschung und Anwendung dieser Methoden wird häufig auf Francis Bacons *Novum Organon Scientiarum* von 1620 gelegt. Darin stellt Bacon (1561–1626) bereits

Regeln für das Experimentieren auf und widmet sich vor allem der Ausschaltung typischer Fehler, die dabei auftreten können. Außerdem schildert er schon, wie wir beim Experimentieren bestimmte Faktoren gezielt und kontrolliert verändern und dadurch erkennen können, auf welche anderen Faktoren sie einen Einfluss haben (vgl. Carrier 2006). Selbst das Konzept eines *experimentum crucis* (eines Entscheidungsexperiments) findet sich bereits bei Bacon. Ein solches Experiment soll entscheiden, welche von zwei Theorien die richtige ist, indem wir nach einer speziellen Schlussfolgerung aus den Theorien suchen, in der sich die beiden Theorien unterscheiden und dann überprüfen, welche der beiden Theorien dort Recht hat.

Noch bekannter sind jedoch vermutlich die Regeln von John Stuart Mill (1806–1873) zur Ermittlung von Kausalbeziehungen in seinem *System of Logic* von 1865. Intuitiv sind sie heute noch in Gebrauch und letztlich die Grundlage für praktisch alle Verfahren zum Auffinden von Ursachen, von denen einige genauer im Kapitel 7 beschrieben werden. Die Grundideen sind recht plausibel, doch sie wirken ziemlich idealisierend.

Mill verwies darauf, dass wir nach bestimmten allgemeinen Mustern in der Natur suchen, die wir für Zwecke der Vorhersage, der Erklärung und der Begründung kontrafaktischer Aussagen verwenden können. Die Muster, die sich dabei bewähren, betrachten wir als Naturgesetze. Auch das deduktiv-nomologische Erklärungsschema findet seinen ersten Verfechter in Mill (vgl. Wilson 2007). Insbesondere aber glaubte er, dass es für jeden Typ von Ereignis ein kausales Naturgesetz gibt, das wir auch entdecken können. Dazu gab er induktive Schlussregeln an, die er schon als *Eliminationsmethoden* verstand, wonach wir möglichst alle bis auf eine Hypothese eliminieren sollten (*System of Logic*: Buch 2, Kap. 9).

Bei den Regeln für die Ermittlung von Ursachen ist zunächst Mills *Regel der Übereinstimmung* (»method of agreement«) zu nennen. Wenn zwei Ereignistypen immer wieder Instanzen aufweisen, die kurz aufeinanderfolgen, spricht das für eine *Kausalbeziehung* zwischen ihnen. Folgen auf unseren lustigen Weinabend immer morgendliche Kopfschmerzen, machen wir vermutlich den Wein als Verantwortlichen dafür aus. Diese Vorgehensweise wird heute noch als wichtiges Indiz für Ursachen betrachtet, obwohl wir eigentlich wissen sollten, dass sich aus Korrelationen nicht so einfach Kausalbeziehungen ableiten lassen. Trotzdem hören

wir in den Medien immer wieder, dass etwa Schokolade Herzinfarkte verhindert, wobei meist nur in Massenuntersuchungen bestimmte Korrelationen zwischen Schokoladengenuss und Herzinfarkten festgestellt wurden. In Kapitel 7 wird das genauer verfolgt. Mill sah schon die Probleme, die sich etwa daraus ergeben, dass wir es mit einer Vielzahl von Faktoren zu tun haben. An diesen Weinabenden waren wir etwa immer mit denselben Leuten zusammen. Also könnten auch diese Treffen die Auslöser der Kopfschmerzen sein.

Mill ergänzte die Regel daher durch die noch wichtigere *Regel des Unterschieds* (»method of difference«), die letztlich in allen Verfahren des kausalen Schließens wiederzufinden ist. Wir schließen typischerweise: Wenn sich zwei Ausgangssituationen S_1 und S_2 nur in einem Faktor F unterscheiden, der in S_1 vorliegt und in S_2 nicht, und E dann in S_1 auftritt, aber in S_2 nicht, so ist F zumindest eine (Teil-) Ursache von E. Das ist heute noch unser Ansatz für viele *kontrollierte Experimente*, für die wir genau diese Konstellation für eine *Versuchsgruppe* und die dazugehörige *Kontrollgruppe* herbeiführen wollen. Die Personen der Versuchsgruppe erhalten etwa ein Medikament X und die der Kontrollgruppe nicht. Wenn hierbei nur die Versuchspersonen aus der Versuchsgruppe gesunden, schließen wir, es müsste am Medikament gelegen haben, jedenfalls dann, wenn es uns (etwa durch eine Zufallsauswahl der Gruppen) gelungen ist, die anderen relevanten Faktoren gleich auf die beiden Gruppen zu verteilen.

Auch das Ausbleiben der Wirkung E kann natürlich in entsprechenden Fällen einen wichtigen Hinweis darauf geben, dass F eben doch nicht den entscheidenden Unterschied ausmacht und daher keine Ursache von E darstellt. Doch dieser Schluss ist meistens eher fragwürdig, weil etwa nicht die erforderlichen Randbedingungen vorlagen, unter denen F seine Wirkung entfaltet. Um zu zeigen, dass F Ursache von E ist, versuchen wir also in der Differenzmethode aufzuzeigen, dass F einen *Unterschiedsmacher* für E in bestimmten Situationen darstellt. Das ist ein wesentlicher Hinweis auf seine kausale Wirksamkeit. Für probabilistische Verursachung müssen wir allerdings größeren Fallzahlen betrachten. Das wichtigste Problem der Methode ist natürlich der Nachweis, dass die Situationen S_1 und S_2 in allen anderen Faktoren, die ebenfalls Einfluss auf den Effekt E haben könnten, gleich (wir werden auch sagen *homogen*)

sind. Das ist in der Praxis oft nur schwer nachweisbar und kann die jeweiligen Ergebnisse in Frage stellen.

Martin Carrier (2006, 31 ff.) beschreibt dazu eine schöne Debatte zwischen Louis Pasteur und Felix Pouchet um 1860 über die Frage, ob es so etwas wie *Urzeugung* geben könne, bei der aus nicht lebendem Material lebendige Zellen entstehen. Ist also nicht lebendes Material eine Ursache für lebendige Zellen? Pouchet war ein Vertreter der Urzeugung und konnte zeigen, dass in abgekochten Proben von Heu nach der Zugabe von Luft Mikroorganismen auftraten, und hielt das für eine Urzeugung aus dem sterilisierten Heu.

Pasteur hielt aber die Luft für den Überträger der Mikroorganismen. Pouchet wiederholte dann das Experiment unter Quecksilberabschluss, so dass nur frisch erzeugter steriler Sauerstoff an das Heu gelangte. Trotzdem traten wieder Mikroorganismen auf. Das sprach nach der Regel des Unterschieds gegen Pasteur, denn mit und ohne sterilisierte Luft traten jeweils Mikroorganismen auf, obwohl sie, wenn Pasteur Recht behalten hätte, nur im zweiten Fall auftreten dürften.

Dagegen konnte Pasteur jedoch kontern und nachweisen, dass es nun das Quecksilber für den Luftabschluss war, das den Effekt erzeugte, weil es mit Mikroorganismen kontaminiert war. Indem er auch das Quecksilber noch erhitzte und zeigte, dass dann keine Mikroorganismen mehr auftraten, argumentierte er gegen die Urzeugung. Auch dabei bemühte er wieder das Unterschiedsprinzip. Der einzige Unterschied war hier das erhitzte oder nicht-erhitzte Quecksilber und das führt zu Mikroorganismen oder keinen. Die Erhitzung bzw. das erhitzte oder nicht-erhitzte Material war hier also ein Unterschiedsmacher für das Auftreten lebendiger Zellen. Wir erkennen in dem Beispiel schon einige der Anwendungsmöglichkeiten der Differenzmethode, die wir in Kapitel 7 ausführlicher untersuchen werden.

Mills dritte Regel der *begleitenden Veränderung* (»method of concomitant variations«) bezieht sich vor allem auf quantitative Größen. Verändere ich die eine Größe A, ändert sich die andere Größe B. Spreche ich dem Wein stärker zu, stellen sich am nächsten Tag auch stärkere Kopfschmerzen ein. Das ist ein gutes Indiz dafür, dass gerade der Weingenuss Ursache meiner Kopfschmerzen ist. Das sind die beobachtbaren Bestandteile der sogenannten *interventionalistischen Kausalitätsauffassung*, nach der

A genau dann Ursache von B ist, wenn ich durch eine Änderung der Größe A eine Änderung der Größe B herbeiführen kann. Das ist gerade heutzutage wieder eine sehr prominente Vorstellung von Kausalität, auf die wir in Kapitel 7 noch zurückkommen werden. Jedenfalls können wir hier erkennen, dass die Grundideen des Experimentierens und der Ermittlung von Kausalbeziehungen durchaus schon früher expliziert wurden und vermutlich schon viel eher intuitiv angewandt wurden.

William Whewell (1794–1866) in *The Philosophy of the Inductive Sciences* (1847) ist der dritte Urvater der modernen Debatte um das induktive Schließen. Er stützte sich explizit auf Bacon und betont in seiner Debatte mit dem eher biederen Empiristen Mill die idealen Elemente in unserem Wissen, bei denen wir durch geeignete Begriffsbildung – oder wir würden heute auch sagen *geeignete Modellbildung* – zu neuen Theorien gelangen, die über eine konservative Induktion hinausgehen. Er sprach dazu von »colligation« oder »consilience«, das wir heute vielleicht am besten als begriffliche und theoretische Vereinheitlichung verstehen können, durch die es gelingt, verschiedene Phänomene unter ein Naturgesetz zu subsumieren. Whewell nennt das die fundamentale Antithese des Wissens, dass hier immer zwei Dinge zusammenkommen, nämlich Wahrnehmungen und Ideen oder Begriffe. Ohne diesen zweiten Anteil kommen wir nicht zu interessanten Kausalgesetzen (vgl. Snyder 2006).

So konnte Kepler die einzelnen Daten zur Position des Mars, die Tycho Brahe gesammelt hatte, durch sein Konzept einer elliptischen Bahn in einen größeren Rahmen setzen. Damit eine Theorie dann als bestätigt gelten kann, muss sie unbekannte Tatsachen erfolgreich vorhersagen, und zu einer gewissen »consilience« sowie Kohärenz führen. Dabei ist mit »consilience« gemeint, dass hier Tatsachen zusammengeführt werden, die zu unterschiedlichen Phänomenen gehören. Ein schönes Beispiel dafür ist Newtons Gravitationsgesetz, durch das so verschiedene Phänomene wie Planetenbewegungen, schiefe Würfe, Pendelbewegungen etc. kausal vereinheitlicht werden können. Wenn uns solche Vereinheitlichungen gelingen, ohne dass wir dazu ad-hoc Anpassungen unserer Hypothesen vornehmen müssen, dann steigert das die Gesamtkohärenz unseres Wissenssystems und stellt daher eine weitere Bestätigung dieser Hypothesen dar (vgl. Snyder 2006). Auch der Wissenschaftstheoretiker

Malcolm Forster hat in verschiedenen Artikeln (etwa 1988), die u.a. auf seiner Homepage zu finden sind, Whewells Idee von »consilience« und die »colligation of facts« an weiteren Beispielen erläutert und argumentiert dafür, dass wir gerade von Whewells Auseinandersetzung mit Mill heute etwas lernen können.

1.3 Grundbegriffe der Bestätigung

Ehe wir uns speziellen Begründungsverfahren zuwenden, möchte ich einige allgemeine begriffliche Klärungen und Unterscheidungen vornehmen sowie einige Abkürzungen verabreden, die uns im Folgenden begleiten werden. Ganz allgemein ist mit »B(H:E)« jeweils gemeint, dass die Hypothese H durch das Datum E bestätigt wird. Mit dieser Notation soll noch nichts darüber gesagt sein, was unter Bestätigung genauer zu verstehen ist, sondern es beschreibt zunächst nur unseren intuitiven Zugang zur Begründung. Natürlich erwarten wir von jeder Begründungskonzeption, dass klare Fälle von wissenschaftlichen Begründungen durch die Konzeption auch als solche eingestuft werden, sonst haben wir ein Problem, dem wir weiter nachgehen müssen. Zunächst spricht eine solche »Anomalie« jedenfalls gegen die vorgeschlagene Konzeption von induktiver Begründung. Wenn es darauf ankommt, können wir auch das Hintergrundwissen K explizit mit einbringen B(H:E;K).

Im konkreten Fall kann B(H:E) etwa bedeuten, dass H ein kleines Stück weit durch E bestätigt wird – im Sinne einer *inkrementellen Bestätigung* – oder dass H im *absoluten* Sinn bestätigt wird, so dass wir nun auch über Gründe verfügen, H zu akzeptieren. Die *Probabilisten*, die solche Beziehungen durch Wahrscheinlichkeiten ausdrücken, würden das z.B. wie folgt beschreiben: Wenn wir mit P(H|E) die Wahrscheinlichkeit bezeichnen, die H erhält, wenn wir bereits E voraussetzen bzw. neu kennengelernt haben, dann lassen sich die beiden Konzepte von Bestätigung wie folgt explizieren:

(1) inkrementelle Bestätigung: $P(H|E) > P(H)$
(bzw. äquivalent dazu: $P(H|E) > P(H|\neg E)$)

(2) absolute Bestätigung: $P(H|E) > k > 1/2$

Demnach erhöht bei der inkrementellen Bestätigung das Vorliegen von E die Wahrscheinlichkeit von H um einen gewissen Betrag gegenüber der Situation, in der wir E noch nicht wissen, aber damit muss die neue Wahrscheinlichkeit von H noch keineswegs hoch sein. Sie kann so immer noch bei sehr kleinen Werten liegen, was bedeutet, dass die Wahrscheinlichkeit für die Negation der Hypothese P(¬H) = 1–P(H) deutlich höher sein kann als P(H). Wenn wir die Wahrscheinlichkeit als eine Art von *Plausibilitätsgrad* auffassen, wäre demnach non-H immer noch sehr viel plausibler als H selbst. Das ändert sich erst bei der absoluten Bestätigung, die verlangt, dass P(H) eine bestimmte Schwelle überschreitet, die sinnvollerweise größer als 0,5 sein sollte. Das muss natürlich noch nicht bedeuten, dass wir H demnach akzeptieren sollten. Wir werden sehen, dass eine entsprechende Abtrennungsregel nicht so leicht zu begründen ist.

(3) Abtrennungsregel:
Bei hinreichender Bestätigung können wir H akzeptieren.

Dazu kommen Vorstellungen von *komparativer Bestätigung*. Zum Beispiel können wir nach Ansicht der Likelihoodisten immer nur sagen, dass ein Datum E eine Theorie H_1 mehr bestätigt als eine Theorie H_2, ohne dabei etwas über eine absolute Bestätigung der einen oder anderen Theorie auszusagen.

(4) komparative Bestätigung: $B(H_1,H_2{:}E)$ (etwa: $P(E|H_1) > P(E|H_2)$)

Probabilisten können das z.B. so deuten, dass die Wahrscheinlichkeitserhöhung, die H_1 durch E erfährt, größer ist als die, die H_2 durch E erfährt. Oft wird das auch so verstanden, dass die Likelihood von H_1 relativ zu H_2 größer ist gegeben E, d.h. $P(E|H_1) > P(E|H_2)$. (Diese etwas seltsame, aber immer noch übliche Redeweise werde ich meist zugunsten der intuitiveren Redeweise vermeiden, dass in diesem Fall E die höhere Likelihood hat bei Vorliegen von H_1 gegenüber dem Vorliegen von H_2.)

Außerdem können wir ebenfalls noch von differentieller Bestätigung sprechen. Dabei geht es darum, dass ein Datum E_1 unsere Hypothese H stärker bestätigt, als es das Datum E_2 vermag:

(5) differentielle Bestätigung: $B(H:E_1,E_2)$ (etwa: $P(E_1|H) > P(E_2|H)$)

Auch das wird oft anhand von Likelihoods gedeutet, wonach $P(E_1|H) > P(E_2|H)$ die entsprechende Beziehung zum Ausdruck bringt. Schurz (2006) spricht allgemeiner von der *Likelihood-Intuition*, wonach wir die rechtfertigende Wirkung eines Datums E für eine Hypothese H vor allem an der Likelihood P(E|H) festmachen. Je höher die ausfällt, umso größer ist auch die rechtfertigende Wirkung von E für H. Das ist eine sehr naheliegende Idee, die in vielen Ansätzen zu finden ist und verschiedene weitere Explikationen aufweist. Damit ist ein erster begrifflicher Rahmen vorgegeben, den wir im weiteren Verlauf der Debatte mit konkreten inhaltlichen Beziehungen ausfüllen müssen.

1.4 Die Daten

Die Daten werden im Folgenden meist als bereits gegeben und frei von Irrtümern betrachtet, aber zunächst sollen doch einige Überlegungen angestellt werden, die diese Idealisierung hinterfragen. Im Rahmen einer fundamentalistischen Erkenntnistheorie wie der empiristischen werden typischerweise Beobachtungsaussagen als basale Aussagen akzeptiert, die keiner weiteren Rechtfertigung bedürfen. Auf sie stützen wir dann alle weiteren Annahmen oder Theorien, aber diese haben eigentlich keine rechtfertigenden Rückwirkungen auf die basalen Aussagen selbst. Man könnte sagen, dass hier *epistemische Rechtfertigung* als eine Einbahnstraße verstanden wird. Trotzdem möchte man meist einräumen, dass diese Beobachtungsaussagen nicht infallibel sind. Dann fragt man sich allerdings, wie wir die fehlerhaften basalen Aussagen aufdecken können und wie sie sich vielleicht sogar von wahren basalen Aussagen unterscheiden lassen. Dabei spielen in aller Regel doch wieder unsere Theorien über die Welt eine wichtige Rolle. Sie sagen uns, in welchen Situationen wir relativ zuverlässige Beobachter sind und welche Irrtumsquellen es andererseits gibt. Warum sollten wir dann nicht immer nach entsprechenden Rechtfertigungen der Beobachtungsaussagen fragen dürfen?

Meines Erachtens beschreiben Kohärenztheoretiker diese Situation besser als die *erkenntnistheoretischen Fundamentalisten*. Wenn etwa die

folgende Beobachtungsmeinung in mir auftritt: (*) »Albert Einstein tritt plötzlich aus dem Nichts geradewegs auf mich zu und beglückwünscht mich zu meinem letzten Buch«, so sollte ich diese Meinung nicht gleich für bare Münze nehmen und mich gebauchpinselt fühlen. Meine Überlegungen sollten vielmehr sein: 1. Albert Einstein ist tot und Tote können nicht wieder auferstehen, also kann es sich nicht um Albert Einstein gehandelt haben. 2. Objekte mittlerer Größe entstehen nicht einfach so aus dem Nichts, also kann es sich nur um eine Sinnestäuschung handeln. 3. Es ist schon recht seltsam, dass sich Einstein gerade für mein letztes Buch so interessiert haben sollte. Das spricht wiederum gegen die Wahrheit von (*). Ich muss also die spontan in mir auftretende Beobachtungsüberzeugung mit meinen anderen Überzeugungen über die Welt konfrontieren und so einer genaueren Bewertung unterziehen.

Aber es könnten auch noch andere Überlegungen ins Spiel kommen, die doch für (*) sprechen: 4. Wie kann ich einer solch lebhaften Sinnestäuschung unterliegen, ohne dass besondere Umstände vorliegen? Also muss doch etwas dran sein an (*). Wir sehen an diesem Beispiel, dass wir sogar unsere Wahrnehmungsüberzeugungen durchaus mit Hilfe anderen Wissens über die Welt in Frage stellen oder begründen müssen, wobei immer gewisse Theorien darüber, wie die Welt funktioniert, mit im Spiel sind (vgl. dazu auch Bartelborth 2015).

Daher sollte ich jede Beobachtungsüberzeugung wie: »Vor mir steht ein weißer Tisch« entsprechend begründen. Tritt sie spontan in mir auf, sollte ich zu ihrer Rechtfertigung anführen, dass nach meiner bisherigen Kenntnis ich ein zuverlässiger Beobachter solcher Objekte mittlerer Größe bin, wenn keine besonderen Bedingungen vorliegen, die einen Irrtum erklärbar machen und diese Ansicht auch sonst in mein Hintergrundwissen kohärent hineinpasst (vgl. Bartelborth 1996, Kap. IV.B). Da nach meinem Wissensstand solche Irrtumsquellen tatsächlich nicht vorliegen, ist meine Beobachtungsüberzeugung also begründet. Damit sind solche Beobachtungsüberzeugungen jedoch nicht mehr wirklich basal bzw. fundamental für unsere weiteren Überzeugungen. Doch diesen Aspekt werden wir im Folgenden nicht weiter verfolgen, da die meisten Ansätze eher fundamentalistisch konzipiert sind und ich hier nur die Frage verfolgen möchte, wann bestimmte Aussagen, die wir

als Daten akzeptieren, andere Behauptungen wie etwa wissenschaftliche Theorien stützen.

Ein dabei immer wieder genanntes Problem ist, dass unsere Daten vielleicht keine neutralen Schiedsrichter für unsere Theorien abgeben können, weil sie selbst schon immer durch Theorien infiziert sind. Daher möchte ich zumindest einige wenige Überlegungen zur Theorienbeladenheit oder *Theorienabhängigkeit* unserer Beobachtungen anstellen. Im schlimmsten Fall sähe das so aus, dass ein Vertreter einer Theorie T_1 in einem bestimmten Phänomen E_1 erblickt, was seine Theorie T_1 stützt, während ein anderer, der Theorie T_2 vertritt, dort E_2 wahrnimmt, was seine Theorie stützt, obwohl die Theorien und Daten einander eigentlich ausschließen.

Dafür sind mir allerdings keine wirklich überzeugenden Beispiele bekannt. Natürlich kann es immer zu entsprechenden Interpretationsproblemen mit den *Rohdaten* kommen. Wenn Galileo durch seine Fernrohre zum Jupiter blickte, konnte er daneben die Monde des Jupiter sehen, während seine Gegner dort nur weiße Flecken erblickten, die sie als Linsenfehler interpretierten (vgl. Chalmers 1994). Die *Rohdaten* sind hier wohl in bestimmten hellen Flecken beim Schauen durch das Fernrohr zu sehen. Galileo konnte für seine Interpretation punkten, indem er darauf verwies, dass sich diese Flecken in systematischer Weise um den Jupiter bewegen, so wie das seine Theorie für bestimmte Monde des Jupiter vorhersagte, während das für Linsenfehler wohl kaum zu erwarten wäre. Diesen Fall würde ich aber eher so beschreiben, dass aus den gemeinsamen (oder ähnlichen) Rohdaten unterschiedliche induktive Schlüsse gezogen wurden, die dann als Daten zur Überprüfung weiterer Theorien dienten. Jedenfalls lassen sich solche Streitfälle oft auflösen, indem wir auf grundlegendere Beschreibungen unserer Wahrnehmungen zurückgehen.

Es gibt unter diesem Stichwort aber noch eine ganze Reihe weiterer Phänomene, die alle eine ausführlichere Debatte verdient hätten. Einige sollen zumindest kurz genannt werden (vgl. dazu etwa Schurz 2006, 57 ff.; Carrier 2006, Kap. 3). Da ist zunächst die Tatsache, dass unsere Erfahrung theoriegeleitet ist. Bestimmte Beobachtungen und Experimente unternehmen wir nur, weil wir bereits bestimmte Theorien dazu im Hinterkopf haben. Daltons Atomtheorie von 1808 ließ uns erst auf

die genauen Verhältnisse der Gewichte einzelner Stoffe schauen und feststellen, dass es ganzzahlige Verhältnisse bestimmter Größen sind. Doch das besagt natürlich nicht, dass diese Beobachtungen dadurch theoretisch infiziert im Sinne des schlimmsten Falles sind. Als neutrale Schiedsrichter können sie trotzdem dienen. Jede Seite wird natürlich speziell die für sie sprechenden Daten herbeischaffen und muss aber auch mit denen der anderen Seite umgehen können.

Speziell können theoretische Terme und komplexe Messverfahren weitere Probleme mit sich bringen, für die etwa Carrier (2006) einige schöne Beispiele analysiert. Aber das soll hier nicht mein Thema sein. Wir arbeiten mit der idealisierenden Annahme weiter, dass unsere Daten solide und sicher sind und wir uns nur noch fragen müssen, welche Theorien sie dann stützen. Das ist die Hauptfragestellung für die Problematik des induktiven Schließens.

1.5 Die konservative Induktion

Bevor wir zu Humes Problem der Induktion kommen, möchte ich einige einfache Induktionsverfahren skizzieren, die dem humeschen Problem mehr Anschaulichkeit verleihen. Wir können sie unter dem Stichwort der konservativen Induktion zusammenfassen. Konservative Induktionsschlüsse versuchen aus den bisherigen Erfahrungen schlicht auf die Zukunft zu *extrapolieren*. Sie schreiben einfach bestimmte Entwicklungen fort und sind in dem Sinne konservativ, dass sie dabei *keine neuen Begriffe* und damit auch keine neuen Entitäten einführen, die nicht schon in der Beschreibung der bisherigen Beobachtungen enthalten wären.

1.5.1 Extrapolationen

Typischerweise extrapolieren wir im Alltag aus unseren bisherigen Erfahrungen auf zukünftige Ereignisse. Wenn wir bisher immer wieder Kopfschmerzen nach heftigem Rotweingenuss bekamen, werden wir das beim nächsten Mal ebenfalls erwarten. Wenn unsere Treppe uns bisher immer gut getragen hat, betreten wir sie und ähnliche Treppen immer wieder ohne ungute Gefühle. Auch in der Wissenschaft extrapolieren

wir so. Wenn sich bisher immer alle Metalle, die wir erhitzt haben, daraufhin ausdehnten, so erwarten wir das genauso für den nächsten Fall oder schließen sogar: Alle Metalle, die erhitzt werden, dehnen sich aus. Das nennt man die *konservative Induktion*. Sie ist die einfachste Art des induktiven Schließens und heißt wie gesagt *konservativ*, weil wir dabei nur auf solche Vorhersagen oder Aussagen schließen, die *dieselben Begriffe* benützen, wie die, die wir schon zur Beschreibung unserer Daten eingesetzt haben. Wir schließen hier also nicht auf unbeobachtbare Dinge hinter den Phänomenen, die vielleicht als Erklärung dieser Phänomene dienen könnten, sondern nehmen nur eine einfache Extrapolation vor, nach dem Motto »*more of the same*« bzw. »*weiter so*«. Die einfachste Form dieser Schlüsse ist daher:

(EX) Extrapolation
Wir beobachten Fa und Fb und Fc und ..., und schließen auf: ∀x(Fx)

Das Verfahren kommt in unterschiedlichen Varianten vor. Als universeller Schluss wie oben in dem Beispiel oder auch – weit weniger anspruchsvoll – als *spezielle Inferenz*, in der wir nur darauf schließen, dass das nächste Stück Metall, das wir untersuchen, sich auch ausdehnen wird, wenn es erhitzt wird. Das kann noch mit dem Ausdruck »wahrscheinlich« qualifiziert werden, wie wir weiter unten sehen werden.

(SEX) spezielle Extrapolation
Fa und Fb und Fc und ..., und schließen auf: Fd (für ein neues d)

Die Extrapolationen sind sicher sehr basale Formen induktiven Schließens, die in unserem Alltag eine große Rolle spielen. Sie treten so auf, als ob wir relativ *direkt* und fast mechanisch von Daten auf bestimmte Hypothesen oder zumindest weitere Fälle schließen könnten und z.B. weiteres Hintergrundwissen dabei keine Rolle spielen würde. Zumindest werden die Verfahren oft so dargestellt. Eine meiner Thesen in dieser Arbeit ist es, dass induktive Schlüsse eigentlich immer *dreistellig* sind und dem *Hintergrundwissen* eine wesentliche Rolle dabei zukommt und wir auch nur auf dem Weg über substantielle Theorien in die Zukunft extrapolieren. Das ist für die konservativen Induktionen nur meist versteckt. Wir finden es aber auch für die Formen des Extrapolierens.

Das wird dann deutlich, wenn wir nicht mehr so einfach bereit sind, dem obigen Schema zu folgen.

Zunächst einmal ist das Schema in gewisser Weise unvollständig, denn wir beziehen uns mehr oder weniger explizit beim Extrapolieren bereits auf eine bestimmte *Grundmenge* oder bestimmte *Art* G von Objekten oder Situationen vertreten durch ein Prädikat G oder eine entsprechende Eigenschaft G, die wir für einander ähnlich halten, so dass wir innerhalb dieser Gruppe bestimmte Eigenschaften extrapolieren können. In unseren Beispielen etwa auf Menschen in normalen Situationen, Metalle oder Treppen normaler Bauweise. Dann erhalten wir das ausführlichere Schema:

(EX) Extrapolation (allgemeine)
Ga & Fa und Gb & Fb und Gc & Fc und ..., dann gilt:
$\forall x(Gx \rightarrow Fx)$ (»Alle G sind F.«)

Und für den speziellen Fall erhalten wir:

(SEX) Extrapolation (spezielle)
Ga & Fa und Gb & Fb und Gc & Fc und ...
und Gd, dann gilt: Fd (für ein neues Objekt d)

Das ist so zu verstehen, dass wir eine endliche Zahl von Objekten untersucht haben, die jeweils G waren, wobei alle davon auch die Eigenschaft F aufwiesen, woraus wir schließen, dass alle G-Objekte auch F-Objekte sind bzw., dass das nächste G-Objekt d, das sich nicht unter den bisher untersuchten Objekten befindet, ebenfalls wieder ein F-Objekt sein wird.

Damit verbinden wir normalerweise bereits die Annahme einer gewissen *Homogenität* in der Gruppe G speziell bezüglich der Eigenschaft F. Haben wir etwa fünf Pfirsiche untersucht und alle hatten einen Kern, so schließen wir schnell, dass wohl alle Pfirsiche einen Kern haben. Wir nehmen typischerweise von bestimmten Obstsorten an, dass sie sich in Bezug auf solche Eigenschaften *relativ homogen* verhalten, also möglichst entweder alle einen Kern haben oder alle keinen Kern enthalten. Wir vermuten nämlich, dass im Hintergrund auch ein kausaler Mechanismus steht, der die Mitglieder der Art G mit dieser

Eigenschaft verknüpft, da die Kerne eine wichtige biologische Funktion bei der Fortpflanzung der Pfirsichbäume haben. Allerdings wird diese Annahme in neuerer Zeit immer mehr aufgeweicht, da wir um spezielle Züchtungserfolge etwa bei Weintrauben wissen, die es nun auch ohne Kern gibt.

Entfällt diese *Homogenitätsannahme*, extrapolieren oder projizieren wir jedenfalls nicht mehr so einfach und sollten das auch nicht tun. Habe ich bisher fünf Philosophen kennengelernt und alle hatten einen Bart, so werde ich nicht sogleich schließen, dass wohl alle Philosophen einen Bart haben werden. Fing der Nachname der ersten vier Kinder einer Schulklasse, die wir kennengelernt haben, mit einem »B« an, so werden wir nicht gleich schließen, dass es sich um eine reine »B-Klasse« handelt. Wenn wir genauer hinsehen, ließen sich viele Muster in unserer Umwelt finden, die wir nicht extrapolieren. Allerdings fallen die uns normalerweise kaum auf, weil wir immer schon gezielt nach den Mustern suchen, von denen wir annehmen, wir dürften sie extrapolieren.

Allerdings »finden« wir solche Muster auch manchmal zu schnell. Ein Nobelpreisträger der Wirtschaftswissenschaften sagte einmal in einer Fernsehsendung dem Sinne nach: »Wann immer wir in der Vergangenheit mehr Öl benötigt haben, wurden auch neue Ölfelder entdeckt. Das wird auch weiterhin der Fall sein.« Doch dieser Schluss kommt uns in der nicht weiter qualifizierten Form kaum akzeptabel vor. Ein anderer Diskutant antwortete darauf: »Aber die Erde ist doch rund.« Wir wissen jedenfalls, dass durch die Endlichkeit der Erde unsere Ressourcen letztlich begrenzt sein müssen und daher die Extrapolation des Nobelpreisträgers in dieser allgemeinen Form nicht zulässig ist.

Oder denken wir an einen »Optimisten«, der aus dem 30. Stockwerk eines Hochhauses springt und man hört ihn nach 20 Stockwerken Fluges noch rufen: »Na bitte, es läuft doch ganz prima.« Wir wissen in solchen Fällen, dass wir aufgrund bestimmter *Endlichkeitsannahmen* so nicht einfach extrapolieren sollten. Nur weil bisher das Einsparen der Wartung für unsere Bremsen zu keinen negativen Folgen geführt hat, dürfen wir nicht einfach annehmen, dass das auch ein stabiles Muster für die Zukunft abgibt. *Endlichkeitsannahmen* sind für uns wichtige Bestandteile des Hintergrundwissens, die unsere Extrapolationen beschränken können.

Auch die Anzahl der bisherigen Beobachtungen ist zumindest für die Qualität der Extrapolation von Bedeutung, wobei die Anzahl umso höher sein sollte, je größer die Heterogenität der Grundmenge G ausfällt. Werden diese Zusatzanforderungen an die konservative Induktion nicht beachtet, spricht man gern von vorschnellen Verallgemeinerungen. Doch das wird noch deutlicher für die einfachen statistischen Extrapolationen.

1.5.2 Statistische Extrapolationen

Ähnlich einfache Schlüsse finden sich in den statistischen Extrapolationen, die ganz analog zu den obigen Extrapolationen verlaufen. Haben wir bisher beobachtet, dass von den mit Lassa-Fieber erkrankten Menschen ca. 15% gestorben sind, so schließen wir darauf, dass 15% aller Menschen mit Lassa-Fieber daran sterben werden.

(StatEX) statistische Extrapolation
$h_n(F/G) = r$, dann gilt: $P(F|G) = r$

Dabei ist mit der relativen Häufigkeit $h_n(F/G)=r$ gemeint, dass unter den ersten n G-Objekten gerade k F-Objekte zu finden waren (mit k/n=r), während P(F|G) die Wahrscheinlichkeit bezeichnet, dass ein G-Objekt auch ein F-Objekt ist, die wir an dieser Stelle einfach als die relative Häufigkeit der F-Objekte in der Gesamtheit der G-Objekte verstehen wollen. Auch dazu können wir wieder eine spezielle Extrapolation finden, die allerdings schon etwas ungewöhnlicher ist:

(SStatEX) spezielle statistische Extrapolation
$h_n(F/G) = r$ und Gd, dann gilt: $P(Fd) = r$ (für eine neues Objekt d).

Neben diesen einfachen Schlussformen finden sich auch andere Regeln wie die Laplacesche Regel

(LR) Laplacesche Regel
$h_n(F/G) = k/n$ und Gd, dann gilt: $P(Fd) = (k+1)/(n+2)$

Diese Regel liefert gerade für kleine n etwas andere Werte als die spezielle statistische Extrapolation. Die Laplacesche Regel verrechnet hier die beobachteten Daten bzw. relativen Häufigkeiten mit einer

Indifferenzeinschätzung, nach der beide Möglichkeiten Fd und ¬Fd dieselbe Wahrscheinlichkeit aufweisen. Für n=0 starten wir damit und behalten dann diese Einschätzung noch für die nächsten n ein Stück weit bei, gehen dann allerdings für größere n zur einfachen statistischen Extrapolation über.

All diesen Schlussformen ist gemeinsam, dass sie von der Kenntnis einer Stichprobe ausgehen und etwas über die Grundgesamtheit sagen wollen oder zumindest über ein weiteres Objekt aus dieser Gesamtheit. Damit das gelingen kann, muss die Stichprobe *repräsentativ* sein für die Grundgesamtheit. Das bedeutet, dass in der Stichprobe die untersuchten Eigenschaften möglichst genauso verteilt sein müssen, wie in der Grundgesamtheit. Wollen wir etwa testen, ob das Lassa-Fieber G zum Tode F führt und wir nehmen an, dass es sich um einen deterministischen Vorgang handelt und dass alle Menschen in der Grundgesamtheit M genau dieselben Faktoren aufweisen, die weiterhin Einfluss auf das Auftreten von F haben können, so genügt eine Einerstichprobe, um das zu entscheiden. Das wäre der Fall völliger Homogenität. Sind aber die anderen für F kausal relevanten Faktoren f_1,...,f_n (d.h. die Faktoren, die zumindest in einer Situation bzw. Faktorenkonstellation zu F oder ¬F beitragen) nicht völlig gleichmäßig verteilt, benötigen wir in unserer Stichprobe eine entsprechende Verteilung dieser Faktoren. Idealerweise sollte jeder Faktor f_i in der Stichprobe S mit derselben relativen Häufigkeit auftreten wie in M. Dann lässt sich an der Stichprobe S erkennen, was G im Hinblick auf F in der Grundgesamtheit für kausale Folgen hat.

Man versucht das entweder dadurch zu erreichen, dass wir nach den bekannten Anteilen der f_i ganz bewusst auch die Stichprobe auswählen oder durch eine *Zufallsstichprobe*, bei der man hofft, dass durch die zufällige Auswahl der Stichprobenelemente der entsprechende Effekt erreicht wird. Das erste Verfahren hat den Nachteil, dass die relevanten Faktoren bekannt sein müssen sowie ihre Verteilung in M. Das zweite Verfahren hat den Nachteil, dass nur mit einer gewissen Wahrscheinlichkeit die entsprechende Repräsentativität der Stichprobe gewährleistet ist. Manchmal bezeichnet man bereits eine richtige Zufallsstichprobe schlicht als repräsentativ, doch das greift etwas zu kurz, denn es geht uns letztlich darum, dass wir aus den Eigenschaften der Stichprobe auf die

entsprechenden Eigenschaften der Grundgesamtheit schließen dürfen, und das ist bei einer Zufallsstichprobe nicht immer gewährleistet. Je größer die Zufallsstichprobe ist, umso größer ist allerdings natürlich der Anteil der Zufallsstichproben (bzw. die Wahrscheinlichkeit für solche Stichproben), die der Grundgesamtheit gleichen. Dass auch die *Randomisierung* nicht immer ausreicht, um darauf aufbauend Schlüsse zu ziehen, werden wir im Kapitel 7 sehen.

Das Grundproblem der Extrapolation ist, dass wir zwischen sinnvoll extrapolierbaren Mustern in bestimmten Situationen und solchen Mustern bzw. Situationen, in denen das nicht der Fall ist, unterscheiden müssen. Irgendwelche Muster finden sich in jeder Datenreihe, doch die wollen wir nicht unbedingt extrapolieren. So könnten Personen auf ein Merkmal X hin untersucht worden sein und alle Personen mit X hatten auch ein e im Vornamen oder hatten dasselbe Sternzeichen usf. Doch derartige Korrelationen würden wir nicht weiter verfolgen. Welche Muster wir beachten oder weiterverfolgen bzw. in die Zukunft oder auf neue Fälle projizieren, wird vor allem durch unser kausales Hintergrundwissen darüber mitbestimmt, welche kausalen Mechanismen in unserer Welt möglich sind und welche wir zumindest gegenwärtig für zu utopisch halten.

Ein besonders wichtiges Induktionsverfahren, das in Kapitel 5 ausführlicher besprochen wird, soll hier schon genannt werden. Es handelt sich um den statistischen Syllogismus, bei dem wir aus bestimmten Kenntnissen über die Grundgesamtheit auf die Eigenschaften der Elemente einer kleinen Stichprobe schließen.

Statistischer Syllogismus: Ist die relative Häufigkeit von Fs in der Grundgesamtheit G gerade r und wir betrachten ein a aus G, von dem wir keinen Grund zu der Annahme haben, dass seine Auswahl speziell mit dem Haben der Eigenschaft F verknüpft ist, dann sollte unsere (subjektive/logische) Wahrscheinlichkeit P(Fa) = r sein.

Dieses Schlussverfahren ist eine wesentliche Grundlage für die induktive Logik und wird aber auch sonst an vielen Stellen im Bereich des induktiven Schließens als Teil des Schlussverfahrens erforderlich (vgl. Franklin 2001).

Wir schließen dabei aus Kenntnissen über die Grundgesamtheit auf Einzelfälle. Wenn wir bereits wissen, dass die relative Häufigkeit von Todesfällen bei Lassa-Fieber 15% beträgt, so können wir damit für einen neuen Fall von Lassa-Fieber vorhersagen, dass er mit 15% Wahrscheinlichkeit zum Tode führen wird, wenn wir keine weiteren Kenntnisse über den Fall haben, die dem entgegenstehen würden.

Wenn wir z.B. für unsere eigene Situation etwas aus statistischen Daten lernen wollen, sind wir immer auf einen entsprechenden Schluss angewiesen. Entwickeln etwa sehr viele Raucher einen Lungenkrebs, dann ist das nur in dem Fall aufschlussreich für mich, wenn ich auch den Schluss von einer solchen Statistik auf meinen konkreten Fall ziehen darf, und der zumindest prima facie begründet, dass das Rauchen genauso mein Lungenkrebsrisiko erhöht. Für das *Lernen aus der Erfahrung* ist der statistische Syllogismus daher praktisch unverzichtbar.

Doch auch dieser sehr wichtige Schluss (auf den wir uns ebenfalls an vielen Stellen in der klassischen Statistik stützen müssen) ist nicht unproblematisch. Er scheint nur dann wirklich statthaft zu sein, wenn die Grundgesamtheit eine geeignete Referenzklasse für unseren Einzelfall darstellt, d.h., wenn die Elemente der Grundgesamtheit, die wir heranziehen, im Wesentlichen dieselben für den Tod relevanten Begleitfaktoren aufweisen, die unser Einzelfall aufweist. Handelt es sich etwa um eine schwangere Frau, liegt die Letalität von Lassa-Fieber wesentlich höher als 15%. Unsere bisherige Grundgesamtheit wäre dafür also keine geeignete Referenzklasse, sondern vielmehr die Menge der schwangeren Frauen mit Lassa-Fieber. Aber vielleicht gibt es sogar noch weitere relevante Faktoren, die wir berücksichtigen müssen. Oder wir verstehen ihn eher im Sinne der epistemischen Wahrscheinlichkeit als ein Maß für unsere Unkenntnis und schließen so, wenn wir über den nächsten Patienten mit Lassa-Fieber schlicht keine weiteren Informationen haben. Dabei können wir für den Schluss auch unterschiedliche Varianten formulieren, wie das etwa bei Schurz (2008) geschieht. Wir werden ihn später in folgendem Sinnen verteidigen: Nur wenn wir kein weiteres relevantes Hintergrundwissen haben, ziehen wir die bisherige relative Häufigkeit als besten Schätzwert für das fragliche Ereignis heran. Dann würde nämlich ein anderer Wert ein Missachten der Daten und eine

Voreingenommenheit bzgl. bestimmter Ergebnisse darstellen. Das wird in Kapitel 5 weiter diskutiert.

1.6 Induktion und Metaphysik

Die bisherigen Induktionsregeln scheinen das induktive Schließen als relativ einfaches Extrapolieren aus den bisherigen Daten zu beschreiben. Doch dieser Eindruck kann unser tatsächliches induktives Schließen nicht richtig beschreiben. Es ist vielmehr eng verknüpft mit unseren kausalen Hypothesen darüber, welche Mechanismen in unserer Welt am Werke sind und welche Gesetze bzw. nomischen Muster ihnen zugrundeliegen. Bayesianer und logische Induktivisten betonen dagegen immer wieder, dass das induktive Schließen bzw. Rechtfertigen keine Beziehungen zu unserer Metaphysik (damit sind hier z.B. unsere sehr allgemeinen Annahmen darüber gemeint, wie die Welt funktioniert) aufwiese. Es ginge wie in der deduktiven Logik schlicht um eine Frage von inferentiellen Beziehungen zwischen Aussagen oder rein erkenntnistheoretische Beziehungen, die unabhängig von unseren metaphysischen Annahmen seien.

Schwächen ihrer Konzeption (etwa im Bayesianismus) versuchen sie sogar manchmal so zu kaschieren. Sie sagen dann etwa: Wie in der deduktiven Logik sei der Bayesianismus nur ein neutraler Schiedsrichter für die Beziehung zwischen Prämissen und Konklusion. Man könne daher nicht zu viel von ihm erwarten. Die Schlüsse könnten nur so gut sein wie die Prämissen. Doch von einer wissenschaftlichen Methodologie erwarten wir in der Regel mehr. Sie soll konkretere Hinweise geben, wo genuine Begründungen unserer Theorien vorliegen und wo die Begründungen weniger überzeugend sind. Außerdem geht unser Hintergrundwissen in das induktive Schließen mit ein, wie wir oben bereits gesehen haben. Das induktive Schließen ist *nichtmonoton* im Unterschied zum deduktiven Schließen, d.h. beim induktiven Schließen kann eine neu hinzukommende Prämisse aus einem plausiblen Schluss einen unplausiblen machen. Daher müssen wir beim induktiven Schließen immer all unsere relevanten Kenntnisse bzgl. der Konklusion berücksichtigen. Das ist für deduktive Schlüsse nicht erforderlich, denn

sie sind monoton. Vor allem gehen – anders als in der deduktiven Logik – die induktiven Schlüsse über den Gehalt der Prämissen hinaus und sollten auch eine entsprechende Projizierbarkeit aufweisen, auf die wir gleich zu sprechen kommen werden.

Über das Hintergrundwissen fließen schließlich auch unsere Metaphysik bzw. unsere kausalen Rahmenannahmen in unsere Schlüsse mit ein. Eine These des Buches ist, dass wir daher unsere Metaphysik viel ernster nehmen müssen, wenn wir die Idee akzeptieren, dass unsere Begründungsbeziehung dreistellig ist und dass unser Hintergrundwissen und insbesondere unser kausales Hintergrundwissen dabei eine wichtige Rolle spielt.

Wir können das etwa im Umfeld der oben eingeführten *Nicodschen Regel* (NR) diskutieren, nach der eine Instanz einer Generalisierung H ≡ ∀x(Rx→Sx) also eine Aussage A ≡ Ra&Sa die Generalisierung H bestätigt: B(H:A). Diese Regel sieht recht plausibel aus, stößt aber auch schnell an ihre Grenzen, wie wir bereits in der Einleitung gesehen haben. Das sollte das Beispiel der Hypothese: »Alle Menschen sind kleiner als 2,50 m« zeigen. Menschen, die 2,48 m groß sind, stellen demnach zwar Instanzen der Hypothese dar, aber unser Verständnis der Situation besagt dann eher: Wenn es schon so große Menschen gibt, dann ist es auch sehr wahrscheinlich, dass irgendwann Menschen von 2,50 m Größe geboren werden.

Auch unsere alltäglichen Extrapolationen verlaufen keineswegs so einfach, wie gern suggeriert wird. Selbst wenn *Krake Paul* die Ergebnisse der deutschen Mannschaft bei der WM 2010 mehrfach korrekt vorhergesagt hat, werden wir ihm für das nächste Spiel keine Vorhersagekraft einräumen wollen, denn nach unserem besten Wissen gibt es keinen kausalen Mechanismus, der es bewerkstelligen könnte, dass Paul in die Zukunft schauen kann oder ein besserer Fußballexperte sein könnte als alle menschlichen Experten. Unsere beste Erklärung für die bisherigen Vorhersageleistungen von Krake Paul ist schlicht und einfach, dass er bisher Glück hatte. Das aber ist keine Grundlage für ein weiteres Vertrauen in Paul oder sollte es jedenfalls nicht sein.

Ob wir also ein bisheriges Muster weiter extrapolieren, hängt natürlich von unserem allgemeinen Modell der Welt ab. Ist in dem kein Platz für einen perfekten Vorhersager der Zukunft, sollten wir auf entsprechende

Vorhersagen nichts geben. Sie passen nicht in den kausalen Rahmen, mit dem wir unsere Welt beschreiben. Das erkennen wir immer wieder in den Fällen, in denen unsere einfachen Extrapolationen mit unserem weiteren Hintergrundwissen in einen deutlichen Konflikt geraten. Wir werden das auch in der Wissensdebatte wiederfinden. Es geht immer darum, eingehende Informationen im Lichte unseres bisherigen Wissens zu beurteilen. Wir nehmen sie nicht einfach für bare Münze und folgen keinen simplen Extrapolationsregeln.

Wenn wir bisher acht Soziologen mit Bart kennengelernt haben, schließen wir noch nicht sofort, dass der nächste Soziologe ebenfalls einen Bart aufweisen wird. Das liegt daran, dass wir im Normalfall nicht annehmen, dass unsere Beobachtungen auf ein zugrundeliegendes *nomisches Muster* (vgl. Bartelborth 2007, Kap. III.5 und s.a. Kap. 3) verweisen – noch nicht einmal ein psychologisches Muster. Das hieße nämlich, dass speziell Soziologen psychisch so disponiert wären, sich ein besonderes Aussehen zu geben, indem sie sich einen Bart wachsen lassen. Doch unser Hintergrundwissen spricht nicht dafür, dass wir von acht Soziologen mit Bart einen derartigen Schluss ziehen sollten. Und dann sollten wir auch keine entsprechenden Vermutungen über den nächsten Soziologen anstellen.

Diese Zusammenhänge finden wir ebenso in komplexeren Beispielen wieder. Wenn wir in der Vergangenheit immer wieder beobachten konnten, dass nach dem Genuss von Schokolade ein Migräneanfall folgte, so dürften wir – wie in ähnlich gelagerten Fällen – zunächst schließen, nach dem nächsten Schokoladengenuss wird uns wohl wieder die Migräne erwischen. Doch wie gut begründet diese Annahme tatsächlich ist, hängt von weiteren Annahmen unseres Hintergrundwissens ab. Implizit beruht sie auf der Vermutung, dass es sich um einen kausalen Zusammenhang handelt, nach dem Schokolade Migräneanfälle *auslöst*.

Sollte diese kausale Annahme falsch sein, steht unsere Vermutung auf tönernen Füßen. Handelt es sich etwa um ein rein zufälliges Zusammentreffen dieser zwei Ereignistypen, dann spricht eine weitere Aufnahme von Schokolade nicht dafür, dass abermals ein Migräneanfall eintritt. Oder, wenn die Medizin Recht hat, dass im Vorfeld der Migräne als erste Vorstufe des Anfalls ein Heißhunger auf Schokolade auftritt, dann ist unser Induktionsschluss ebenfalls nicht unbedingt gerechtfertigt. Esse

ich etwa nur, weil mir jemand eine besonders schmackhafte Schokolade anbietet, ist kein folgender Migräneanfall zu erwarten. Esse ich hingegen, weil ich gerade einen Drang zu Schokolade verspüre, dann habe ich einen guten Grund, einen folgenden Migräneanfall zu befürchten. Man sieht die komplexe Abhängigkeit unserer Extrapolationen und Vorhersagen von kausalen Hintergrundannahmen.

Unsere ersten intuitiven Induktionsüberlegungen und eine wichtige These meiner Arbeit – auf die ich immer wieder zurückkommen werde – lässt sich etwas ausführlicher so zusammenfassen:

> **Intuitives Induktionsverfahren**: Wenn in der Vergangenheit Instanzen von B immer auf Instanzen von A folgten und wir anhand unseres Hintergrundwissens Grund zu der Annahme haben, dass es sich dabei um einen stabilen Zusammenhang – eine Art von (kausalem) nomischen Muster zwischen A und B – handelt bzw. ein kausaler Mechanismus im Hintergrund diesen Zusammenhang stiftet, dann liefert uns ein weiteres Auftreten von A einen Grund, an ein weiteres Auftreten von B zu glauben (man sagt auch: A ist projizierbar in Bezug auf B).

Kausale Zusammenhänge oder kausale Ähnlichkeiten spielen für unsere Schlüsse also eine wesentliche Rolle. Wenn wir sechs Exemplare einer exotischen Frucht untersucht haben (nennen wir sie »Pflirsche«) und alle hatten einen großen zentralen Kern, dann werden wir das auch von der nächsten Pflirsche erwarten oder sogar gleich von allen Pflirschen. Das liegt daran, dass wir die Exemplare einer Frucht als Exemplare einer *natürlichen Art* betrachten, die in bestimmten Aspekten wie dem Aufweisen eines Kerns typischerweise einander sehr ähnlich sind. Die Entstehung dieser Früchte und die Zugehörigkeit zu der Art ist nach unserem bisherigen Wissen kausal mit dem Auftreten des Kerns verknüpft. Hier findet sich also ein entsprechendes kausales Muster. Wir wissen allerdings aus Züchtungserfolgen, dass auf diese Annahme kein hundertprozentiger Verlass ist. Doch damit müssen wir beim *induktiven* Schließen immer rechnen.

Wenn wir hingegen sechs Philosophen kennengelernt haben und alle hatten eine graue Hose an, werden wir keineswegs sogleich erwarten,

dass auch der nächste Philosoph wieder graubehost sein wird. Das würden wir nur erwarten, wenn wir weiteres Hintergrundwissen oder zumindest entsprechende Hinweise in dieser Richtung hätten, etwa derart, dass es eine entsprechende Kleiderkonvention unter Philosophen gäbe. Im Normalfall würden wir dagegen denken, dass es sich um einen bloßen Zufall handelt und wir beim nächsten Philosophen nur mit der für Akademiker üblichen Quote wieder eine graue Hose erwarten dürften. Philosophen sind insbesondere im Hinblick auf ihre Kleidung keine natürliche Art, deren Elemente einander im Hinblick auf diese Eigenschaft kausal ähnlich wären. Daher würden wir hier keine entsprechend einfache Extrapolation vornehmen, sondern eher vermuten, dass nur ein Zufall vorlag.

Man kann das zugespitzt so formulieren: *Um in die Zukunft extrapolieren zu können, müssen wir aus unseren bisherigen Erfahrungen zuerst auf ein zugrundeliegendes kausales nomisches Muster schließen und können erst mit seiner Hilfe auf künftige Ereignisse schließen.* Welche Muster dabei einen vermutlich gesetzesartigen Charakter haben, wird schon anhand weiteren Hintergrundwissens zu ermitteln sein. Daher setze ich kein Vertrauen in einfache lineare Extrapolationen. Sie sind nur brauchbar, wenn entsprechende Muster dahinter stehen. Dieser Zusammenhang erklärt zugleich viele Regeln für das induktive Schließen. Das hat schon damit zu tun, dass unser Wissen oft eine spezielle konditionale Form hat. Wissenschaftstheoretiker und Logiker identifizieren die oft vorschnell mit dem materialen Konditional, doch das greift zu kurz. Das lässt sich schon in einfachen Alltagsbeispielen erkennen.

1.6.1 Konditionales Wissen

Unser Wissen über die Welt, nach dem wir im Alltag und in der Wissenschaft suchen, hat oft eine konditionale Form. Es besagt etwa: »Wenn ich eine bestimmte Menge Alkohol trinke, habe ich einen schlimmen Kater« oder »Wenn ich ein Stück Metall erhitze, wird es sich ausdehnen« oder »Wenn ich bei einer Lungenentzündung Antibiotika nehme, dann werde ich *meistens* schnell danach gesund«. Diese Form des Wissens hilft uns auch dabei, in bestimmten Situationen zu entscheiden, was zu tun ist, wenn wir ein bestimmtes Ergebnis herbeiführen möchten.

Unser wissenschaftliches Wissen soll es uns erlauben, aus bisher beobachteten Fällen oder Situationen auf neue Fälle oder neue Situationen zu schließen, die wir noch nicht beobachtet haben. Es soll uns sagen, was passieren würde, wenn wir bestimmte Dinge täten. Oder auch was passiert wäre, wenn wir uns anders verhalten hätten. Erst dadurch wird es für uns so spannend. Die oben genannten »Wenn-dann-Aussagen« sollen dieses Wissen repräsentieren. Sie werden meistens als einfache *materiale Konditionale* im Sinne der klassischen Logik gedeutet. Mit dieser Vereinfachung werde ich später auch in der Regel arbeiten. Aber wir sollten uns zunächst fragen, ob diese Darstellung tatsächlich geeignet ist, damit unsere wissenschaftlichen Erkenntnisse, die ihnen zugedachte Rolle erfüllen können. Das scheint m.E. offensichtlich nicht der Fall zu sein, denn die wissenschaftlichen Konditionale haben einen gewissen kontrafaktischen Gehalt. Um das zu erläutern, muss ich mich ein wenig mit der Logik von Konditionalaussagen beschäftigen.

Unsere erste Frage ist also, wie das Konditional in diesen Aussagen zu verstehen ist. Jonathan Bennetts Werk *A Philosophical Guide to Conditionals* von (2003) ist inzwischen ein Klassiker zu dem Thema, und vermittelt, wie schwierig es ist, schon normale Alltagskonditionale zu verstehen. Man denke z.B. an:

(1) Wenn ich diese Schweinshaxe esse, wird mir schlecht.
oder allgemeiner:
(2) Wenn ich eine Schweinshaxe esse, wird mir schlecht.

Wie sind diese Konditionalaussagen zu verstehen? Betrachten wir zunächst den Satz (1). Hat er tatsächlich die Gestalt »A→B« mit einem materialen Konditional »→«? Das hieße dann, der Satz wäre logisch äquivalent zu dem Satz »non-A oder B«. Ich könnte dann behaupten, meine Aussage (1) wäre wahr, sobald ich auf die Schweinshaxe verzichte oder sobald mir schlecht wird. Das ist natürlich im Normalfall nicht gemeint mit meiner Aussage (1) oder entsprechend mit (2). Mit (1) möchte ich vielmehr nur eine Behauptung über die Situationen aufstellen, in denen ich die Schweinshaxe tatsächlich essen würde. Man könnte sich also der Bedeutung vermutlich schon eher durch das *subjunktive Konditional* nähern:

(1*) Wenn ich die Schweinshaxe äße (oder: essen würde), würde mir schlecht. (formal hier: »A>B«; wird auch manchmal »A $\Box\!\!\rightarrow$ B« geschrieben)

Auch dieses subjunktive Konditional ist nicht leicht verständlich. Es wird meist als kontrafaktisches Konditional bezeichnet, weil es im Deutschen oft impliziert, dass das Antezedens falsch ist, was für subjunktive Konditionale nicht unbedingt der Fall sein muss, obwohl es in unserem Beispiel naheliegend erscheint. Man denke aber z.B. an:

(3) Wenn ich das Antibiotikum einnehme, wird es mir schnell besser gehen.

Damit will ich sicher nicht zum Ausdruck bringen, dass ich das Antibiotikum nicht einnehmen werde. Ich will nur sagen, was in den speziellen Situationen passieren wird, in denen ich es tue.

Dabei muss das Konsequens auch nicht zwangsläufig auftreten, wenn das Antezedens wahr wäre. Wir können und sollten dafür durchaus Ausnahmen erlauben und eine Art von Ceteris-Paribus-Klausel einbauen. Die grundlegende Semantik für diese subjunktiven Konditionale hat schon David Lewis (1973) entwickelt, aber z.B. Hannes Leitgeb stellt (2012) eine weiterentwickelte (probabilistische) Semantik vor, die auch solche Ausnahmen erlaubt bzw. eine Ceteris-Paribus-Klausel einbaut. Das ist besonders hilfreich, wenn wir uns auf diesem Wege wissenschaftlichen Aussagen oder zumindest allgemeineren Aussagen wie (2) nähern möchten, da diese meistens mit einer solchen Klausel zu versehen sind.

Doch selbst die Lesart von (1) – (3) als subjunktiver Konditionale (A>B) ist noch nicht ganz zufriedenstellend. Es ist nicht sichergestellt, dass der behauptete inhaltliche oder kausale Zusammenhang dadurch schon angemessen erfasst wird. So folgt im Rahmen der Standardsemantik (allerdings nicht bei Leitgeb 2012) aus »A&B« das Konditional »A>B«, aber mit »A&B« ist noch nichts über den Zusammenhang zwischen A und B gesagt. Wir möchten mit (1) im Normalfall jedoch zusätzlich zum Ausdruck bringen, dass das Essen der Schweinshaxe (kausal) verantwortlich für unsere Übelkeit wäre, sollte beides auftreten. Die Aussage (1) ist dann nicht schon automatisch wahr, wenn ich die Schweinshaxe esse und mir übel wird. Das allein reicht oft noch nicht. Es

geht uns jedenfalls oft um einen engen inhaltlichen Zusammenhang, den wir mit dem Konditional ausdrücken möchten, den ich in Anlehnung an ähnliche Überlegungen bei Pollock (1976) als *nomisches Konditional* bezeichnen möchte: »A≫B«.

Das bietet sich jedenfalls spätestens für die generelle Aussage (2) an und damit auch für die meisten Behauptungen im Rahmen der Wissenschaft: ∀x(Fx≫Gx). Wenn der Ökonom behauptet, dass Preissenkungen zu einer verstärkten Nachfrage führen, möchte er inhaltliche Zusammenhänge behaupten. Allerdings sollen nicht nur direkte Ursache-Wirkungs-Zusammenhänge unter das nomische Konditional fallen, sondern auch indirekte Zusammenhänge, die etwa durch gemeinsame Ursachen bzw. kausale Mechanismen erzeugt werden: »Wenn das Barometer fällt, wird ein Sturm auftreten.« Oder indirektere Zusammenhänge wie: »Wenn etwas ein Smaragd ist, ist es grün« oder »Wenn etwas ein Pfirsich ist, hat es einen Kern.« Hier gibt es kausale Mechanismen im Hintergrund, die die Zusammenhänge stiften. Damit erhalten wir:

> **Nomische Konditionale**: ∀x(Fx≫Gx) bedeutet demnach: Wenn ein Objekt x die Eigenschaft F aufweisen würde, dann gibt es (im Hintergrund) einen gesetzesartigen kausalen Mechanismus, der dafür sorgt, dass es dann (zumindest in den meisten Fällen) auch die Eigenschaft G aufweisen würde, wenn keine speziellen Störfaktoren auftreten.

Im Grenzfall handelt es sich um echte *Naturgesetze*, die einen ausnahmslosen Zusammenhang zwischen bestimmten Größen herstellen. Die Darstellung solcher Zusammenhänge durch materiale Implikationen »∀x(Fx→Gx)« ist jedenfalls irreführend und hat schon zu verschiedenen Problemen und Paradoxien geführt, auf die wir später eingehen werden. Das erste Problem ist die Annahme, wir hätten es nur mit einer losen Sammlung von Fakten zu tun, die durch einfache Extrapolationen zu erschließen sei.

Es ist nicht leicht, das nomische Konditional genauer zu beschreiben. Es gibt keinen einfachen Kalkül dafür, aber es ist in jedem Fall »variably strict« wie man im englischen sagt. Das heißt, es ist kein striktes Konditional der Art »Es ist notwendig, dass aus F G folgt«. Solche Konditionale wären nämlich monoton (wenn a≫b, dann gilt a&c≫b),

was für die variabel strikten Konditionale nicht gilt. Sie haben eine variable Art von Notwendigkeit (die entsprechende materiale Implikation muss nicht in allen erreichbaren möglichen Welten gelten, sondern nur in den nähergelegenen). Sie sind meistens *nicht-monoton* und damit automatisch auch nicht transitiv. Das kennen wir bereits von normalen kontrafaktischen Konditionalen. Man betrachte etwa die Aussagen:

(1) Wenn es gestern sonnig gewesen wäre, wäre ich Wasserski gelaufen.
(2) Wenn es gestern sonnig gewesen wäre und ich vorgestern einen Unfall gehabt hätte, wäre ich Wasserski gelaufen.

Auch wenn (1) wahr wäre, könnte (2) sehr wohl falsch sein. Doch trotz dieser Probleme, einen schönen logischen Kalkül für »≫« zu erhalten, müssen wir uns weiter mit dem nomischen Konditional auseinandersetzen, weil es eine Grundform des konditionalen Wissens darstellt. In diesem Buch begnüge ich mich aber mit den obigen Hinweisen und einem ersten intuitiven Verständnis dessen, was wir mit einer gesetzesartigen Konditionalaussage meinen.

Damit haben wir zugleich die Form der Muster bestimmt, die erforderlich sind, damit wir induktiv schließen dürfen bzw. aus bestimmten Daten extrapolieren können. Das ist nämlich nur dann sinnvoll, wenn wir auch zugleich darauf schließen dürfen, dass unsere Daten Instanzen eines solchen nomischen Konditionals darstellen. Das können wir uns an weiteren Beispielen ansehen.

1.6.2 Projizierbarkeit und Induktionseigenschaft

Wir alle müssen aus unseren Beobachtungen und anderen Erfahrungen aus der Vergangenheit lernen bzw. aus den gemachten Beobachtungen Schlüsse ziehen, um unsere zukünftigen Entscheidungen auf das so Gelernte stützen zu können. Die Frage stellt sich aber, wann wir aus bestimmten Beobachtungen etwas schließen dürfen und welcher Art das so gewonnene Wissen ist. Diesen Fragen möchte ich zunächst für einfache Fälle von konservativer Induktion vor allem im Hinblick auf wissenschaftliches Wissen weiter nachgehen. Viele Autoren behaupten, dass diese einfachen Induktionsverfahren für unser induktives Schließen besonders grundlegend sind, und es dafür entsprechend einfache

Schlussregeln gibt, die nicht auf weitere metaphysische Annahmen wie etwa kausale Annahmen angewiesen sind. Das war auch immer der Traum der Empiristen, dass wir vielleicht sogar eine weitgehend annahmenfreie *induktive Logik* erstellen können, die die einfachen Lernprozesse beschreibt und begründet, ohne dass wir gleich in eine inhaltlich substantiellere Theorienbildung einsteigen müssten.

Diese Annahme scheint mir falsch zu sein, was in den letzten Abschnitten bereits anhand erster Beispiele begründet wurde, und ich möchte an dieser Stelle ein anderes Bild des Lernens aus der Erfahrung dagegenstellen. Vereinzelt sind schon Vorstöße in dieser Richtung zu finden, wie etwa in Norton (2003), aber das Gesamtbild wird doch durch die Vorstellung einer einfachen Induktion geprägt. Die geschilderte empiristische Grundidee findet sich an vielen Stellen im Aufbau unseres Wissen wieder u.a. im Bayesianismus, aber ebenso in der klassischen Statistik. Den Spezialfall der konservativen Induktion, die aus einfachen Beobachtungen im Sinne einer einfachen Extrapolation ohne weitere Annahmen zu unterschreiben schließt, möchte ich hier abkürzend als *lineare Induktion* oder als *lineares Schließen* bezeichnen. Es versucht, ohne größere Umwege über komplexeres Hintergrundwissen zu einfachem konditionalem Wissen zu gelangen. Nach Ansicht von Hawthorne & Fitelson (2010, Anm. 8, S. 272) hat die Bestätigungsbeziehung auch keine entsprechenden Verbindungen zu metaphysischen Konzepten: »Confirmation is a logical or epistemic relation, which may or may not align neatly with metaphysical relations like causation or law-likeness.«

Demgegenüber wird hier die Idee vertreten, dass wir nur dort induktiv aus den Daten etwas erschließen können, wo wir sinnvollerweise annehmen dürfen, dass ein nomisches Konditional unsere Daten erklärt.

Fortsetzen möchte ich meine Analyse mit einer Reihe von einfachen Beispielen von hypothetischen Daten und der Leser sollte sich in jedem Einzelfall fragen, was wir daraus schließen dürfen:

(1) Alle 5 Philosophen, die ich bisher gesehen habe, trugen braune Schuhe.
(2) Alle bisher untersuchten Früchte der Sorte X hatten einen großen Kern.
(3) Alle bisher untersuchten Früchte der Sorte X waren von Maden befallen (/ hatten eine beschädigte Stelle).

(4) Alle bisher untersuchten Proben von Wachs schmolzen bei 41C°.
(5) Alle bisher untersuchten Proben von Kupfer schmolzen bei 1360C°
(6) Alle (bisher) untersuchten Raben sind schwarz.
(7) Alle untersuchten Objekten, die Franz gehörten, hatten als dritten Buchstaben in ihrer Bezeichnung ein »a«. (z.B. Toaster)
(8) Bisher waren alle Smaragde grün (/graun).
(9) Ein Heiler, der sich auf Gott beruft, war mehrfach erfolgreich.
(10) Ein (angeblicher) Hellseher hat mehrfach korrekt vorhergesagt, welche Lottozahl gezogen wurde.
(11) Metalle dehnten sich aus, wenn sie erhitzt wurden.
(12) Starke Raucher bekamen (bisher) zu in 90% der Fälle Lungenkrebs.
(13) Nach Penicillin ging häufig die Lungenentzündung weg.
(14) Es ist beim Roulette 20-mal rot gekommen.
(15) Wenn es regnet, wurde die Straße nass.

Was können wir jeweils aus den Daten lernen? Können wir die Entwicklung zumindest für Einzelfälle fortschreiben (in die Zukunft projizieren) oder sogar auf ein allgemeineres Muster schließen bzw. dieses zumindest durch die bisherigen Daten (schwach) begründen? Auf das hier sichtbar werdende Problem hat uns vor allem Nelson Goodman mit seinem Prädikat »grue« (hier übersetzt zu »graun«) gebracht – auf das wir später noch genauer eingehen werden –, das keine *projizierbaren* Muster beschreibt. Viele der hier genannten Muster sind ebenfalls nicht projizierbar, aber andere schon, und z.T. ist die Fortschreibbarkeit der Muster auch graduell verschieden. Wir können jedenfalls in unserer Umwelt viele solcher nicht projizierbaren Muster finden.

Zunächst zu meiner etwas unübersichtlichen Liste: Im Falle (1) der Philosophen mit braunen Schuhen sollten wir keine weiteren Schlüsse auf die Schuhe des nächsten Philosophen oder allgemeine Muster der Art (Philosoph > trägt braune Schuhe) ziehen, weil uns unser Hintergrundwissen sagt, dass im Bereich Kleidung nur für bestimmte Berufsgruppen normalerweise eine spezielle Kleidung zu erwarten ist und ansonsten die betreffenden Merkmale eher heterogen verteilt sind. Das beobachtete Muster ist also wohl nur zufällig. Es steht jedenfalls vermutlich kein projizierbares gesetzesartiges Muster im Hintergrund, auf das wir uns stützen könnten. Ähnlich sieht das für viele weitere Beispiele aus.

Anders ist das für (2), (5) oder (11). Wir vermuten in diesen Fällen, dass wir auf die nächsten Beispiele extrapolieren dürfen und unser Hintergrundwissen legt nahe, dass wir es in derartigen Fällen nicht nur mit einem zufälligen Muster zu tun haben, sondern dass es einen gesetzesartigen kausalen Zusammenhang gibt, der im Hintergrund steht und diese Zusammenhänge vermittelt bzw. stabil hält, wenn überhaupt ein Zusammenhang vorliegt. Auch hier gilt wieder, dass die Muster, die wir vermuten bzw. teilweise beobachten können, keineswegs materiale Implikationen sind, sondern eher durch subjunktive Konditionale bzw. nomische Konditionale dargestellt werden können.

Liegen nach allem, was wir wissen, keine solchen kausalen Mechanismen vor, die Antezedens und Konsequens unserer Konditionale miteinander verbinden, haben wir auch keinen Grund, anhand unserer Daten auf sie zu schließen und haben auch keinen Grund, für den nächsten Fall anhand dieser Zusammenhänge zu extrapolieren. Die Fälle (1), (7), (10) und (14) sind z.B. problematisch und unser Hintergrundwissen sagt uns, dass sie keine guten Daten für eine Extrapolation darstellen, weil wir es für unwahrscheinlich halten, dass sich dahinter ein stabiles kausales Muster verbirgt, dass einen solchen Schluss erst begründen könnte.

Ob wir ein projizierbares Muster annehmen dürfen, hängt natürlich ebenfalls von der Anzahl der beobachteten Instanzen ab. Gerade für die Situationen, in den wir gemäß unserem bisherigen Hintergrundwissen nicht davon ausgehen können, dass überhaupt gesetzesartige Muster vorliegen, genügen kleine Anzahlen von Instanzen nicht, um induktiv schließen zu dürfen. Sollten die Anzahlen allerdings sehr groß werden, müssen wir unsere ursprüngliche Annahme, dass kein nomisches Muster im Hintergund steht, noch einmal überdenken.

Sind unsere Muster vermutlich nicht projizierbar, sollten unsere Induktionsverfahren allerdings auch keine *Induktionseigenschaft* für sie aufweisen. Unser Induktionsverfahren weist die Induktionseigenschaft für ein Muster M auf, wenn es vorsieht, dass durch das Auftreten bestimmter Instanzen von M auch andere Instanzen des Musters als besser bestätigt einzustufen sind. Unsere Induktionsverfahren müssen dann aber so komplex gestaltet sein, dass sie darauf reagieren, ob die vorliegenden Muster projizierbar sind oder nicht. Das ist in der Praxis nur sehr schwer umzusetzen, zumal wir keine einfachen Kriterien für Projizierbarkeit

angeben können. Diese Anforderung an unsere Induktionsverfahren wird daher wohl immer eine informelle Zusatzforderung bleiben, für die wir nur fragen können, wie sie sich am besten in das jeweilige Induktionsverfahren einbeziehen lässt. Am einfachsten gelingt das beim sogenannten *Schluss auf die beste Erklärung*, denn wir werden in Kapitel 4 sehen, dass zu einer guten wissenschaftlichen Erklärung immer auch nomische Muster dazugehören (vgl. Bartelborth 2007).

1.6.3 Das erste Problem der Induktion

Wir haben also gelernt, dass wir schon im Rahmen einer konservativen Induktion nur in bestimmten Fällen extrapolieren sollten, nämlich in den Fällen, in denen wir in unseren Daten plus Hintergrundwissen Gründe für die Annahme sehen, dass wir Instanzen eines stabilen Musters in unserer Welt vor uns haben. Das können wir auch als allgemeineres Problem formulieren, das ich als das *erste Problem der Induktion* bezeichnen möchte, das bereits dann auftritt, wenn wir noch nicht den radikalen Humeschen Induktionsskeptiker diskutieren:

> **Das erste Problem der Induktion**: In welchen Fällen und auf welche Weise können uns Daten über bestimmte Objekte oder Situationen begründete Informationen über andere Objekte oder Situationen liefern?

Dieses Problem findet sich in entsprechender Form für alle unsere Induktionsverfahren wieder und stellt das Hauptthema des ganzen Buches dar. Denken wir uns eine sehr kleine Welt mit nur drei Objekten a, b, c und einem (möglicherweise komplexeren) Prädikat F. Die Frage ist nun, wenn ich herausfinde, dass Fa gilt, was sollte mir das über die beiden anderen Objekte b und c sagen und inwiefern kann es überhaupt über die beiden anderen Objekte Informationen liefern? Die konservative Induktion scheint anzunehmen, dass wir darin einfach einen ersten (schwachen) Grund erblicken dürfen, dass auch Fb und Fc gilt. Viele Induktionsverfahren, die wir im Folgenden genauer untersuchen werden, gehen aber über die konservative Induktion hinaus und setzen zu diesem

Zweck gleich auf die Bestätigung bestimmter komplexerer Hypothesen. Ist das vielleicht der entscheidende Lösungsweg für unser Problem?

Meine Beispielhypothese sei in meiner kleinen Welt gerade $H \equiv \forall x(Fx)$. Gemäß dem hypothetisch-deduktiven Bestätigungsansatz wird H nun durch unser Datum Fa bestätigt, weil Fa aus H deduktiv folgt. Das überträgt sich in entsprechender Weise auf den bayesianischen Ansatz. Aber liefert uns das tatsächlich schon gute Hinweise darauf, dass Fb oder Fc gilt? Wenn nichts Weiteres hinzukommt, ist das nicht der Fall. Vielmehr haben wir zunächst eher einen Fall von »content cutting« vor uns, wie das Earman (1992, 98) oder Schurz (2014, 321) nennen.

Ich möchte das etwas anschaulicher als »*deduktive Teilbestätigung*« bezeichnen. Die besagt, dass zwar ein Teil der Hypothese $H \equiv Fa \,\&\, Fb \,\&\, Fc$ bestätigt wird, nämlich der erste Teil der Konjunktion – und dieser sogar deduktiv –, aber über die anderen damit noch nichts gesagt wurde. Sie wurden nicht mitbestätigt und auch im bayesianischen Ansatz ist ohne weitere Annahmen die Wahrscheinlichkeit der restlichen Teile nicht erhöht worden. Daher liegt keine genuin *induktive Bestätigung* vor, sondern eben nur eine deduktive Teilbestätigung der Hypothese, die uns für das erste Induktionsproblem nicht weiterhilft. Der Rest der Hypothese H wird zunächst jedenfalls nicht wirklich *mitbestätigt*.

Die Hypothese H könnte eine völlig zufällige Ansammlung von Einzeltatsachen darstellen. Dann wird H zwar insgesamt dadurch bestätigt, dass eine deduktive Teilbestätigung stattfindet, aber mit induktivem Schließen hat das nichts zu tun. Auf diesem Wege gewinnen wir jedenfalls keine zufriedenstellende Antwort auf die oben gestellte Frage. Zumindest können also irgendwelche beliebigen Generalisierungen uns nicht weiterhelfen. Wir benötigen zumindest wieder *nomische* Generalisierungen. Erst wenn die beste Erklärung für unsere Daten darin besteht, dass sie alle Instanzen eines nomischen Konditionals sind, dürfen wir auch für andere Fälle, in denen das Antezedens des Konditionals erfüllt ist, begründet annehmen, dass auch das Konsequenz des Konditionals wahr ist. Unser Muster, das oben nur durch ein Prädikat F angegeben wurde, hätte dann auch wieder eine komplexere konditionale Struktur und es wären mindestens zwei Prädikate involviert. Diese Idee lässt sich so zusammenfassen:

Instanzenbestätigung: Eine Instanz Fa&Ga bestätigt nur ein nomisches Konditional ∀x(Fx≫Gx) *induktiv*, während es für eine materiale Implikation ∀x(Fx→Gx) nur eine *deduktive Teilbestätigung* darstellt.

Tatsächlich neigen wir dazu, überall stabile Muster zu erkennen, selbst wenn keine vorliegen. Vermutlich hat uns die Evolution dafür einen guten »Riecher« mitgegeben, Muster zu finden, damit wir unsere Zukunft jeweils erfolgreich planen können. Wir nehmen sogar dort Muster an, wo eine statistische Analyse zeigt, dass es sich eher um Zufallsprozesse handelt. Das fand man u.a. bei der Untersuchung des »Laufs« von Basketballspielern (vgl. Gilovich 1991) oder des Erfolgs von Fondsmanagern (vgl. Kahneman 2014) und in anderen Beispielen, in denen wir aus bestimmten Daten (insbesondere Erfolgen oder Misserfolgen in der Vergangenheit) auch nicht auf dementsprechende zukünftige Entwicklungen schließen dürfen, weil den Erfolgen keine stabilen Muster zugrundeliegen.

Wir alle kennen eine ähnliche Problematik z.B. vom Roulettspiel. Das werden wir im Normalfall als Zufallsprozess betrachten, aber wir neigen trotzdem zu einer fehlerhaften Extrapolation in diesem Fall, vermutlich weil wir auch von solchen Zufallsprozessen eine zu einfache Vorstellung – als eine Art von alternierendem Muster – haben. Beim sogenannten *Fehlschluss des Spielers* nehmen wir an, dass nach häufigem Rot nun als nächstes Schwarz kommen wird. Das ist natürlich nicht zu rechtfertigen. Wenn es sich um einen Zufallsprozess eines fairen Roulettes handelt, sollten wir Rot und Schwarz gleichermaßen erwarten, und wenn wir von der Zufallsprozessannahme Abstriche vornehmen möchten, sollten wir gemäß unseren Daten eher Rot erwarten.

Einige Muster sind sicherlich stärker bzw. gesetzesartiger als andere. Den Schmelzpunkt von Kupfer können wir anhand weniger Experimente entdecken, doch bei dem Gemisch Wachs (ein Beispiel von Norton 2003) sieht das schon anders aus. Hier sollte uns klar sein, dass die Daten nur schwächere Vorhersagen gestatten. Das gilt natürlich selbst für die Früchte, denn wir wissen inzwischen, wie sich die genetische Ausstattung verändern lässt.

Die Rabenparadoxie. Die Unterscheidung zwischen der materialen Implikation und nomischen Konditionalen kann auch dabei helfen, Paradoxien wie das Rabenparadox aufzulösen, auf das wir später noch genauer eingehen werden. Es geht zunächst von der Nicodschen Regel aus, wonach eine einfache Instanz »Sa & Ra« einer Hypothese H ≡ ∀x(Sx→Rx), diese immer bestätigt. Wenn unsere Hypothese H also lautet »Alle Raben sind schwarz« und unsere Instanz besteht in einem Objekt a, das sowohl schwarz wie auch ein Rabe ist, so würde demnach H zumindest ein Stück weit bestätigt. Außerdem sollten aber logisch äquivalente Hypothesen durch dieselben Daten bestätigt werden. Nun ist unsere Hypothese H äquivalent zu ihrer Kontraposition H ⇔ ∀x(¬Rx→¬Sx). Für die wäre aber ¬Sa & ¬Ra eine Instanz, die dann die Kontraposition bestätigen müsste. Damit würde aber jedes Objekt, das nicht-schwarz und ein Nicht-Rabe wäre, zugleich unsere ursprüngliche Rabenhypothese bestätigen. Also könnte eine rote Hose oder ein weißes Küchengerät oder ein brauner Schuh unsere Rabenhypothese bestätigen.

Das kommt uns hoffentlich unsinnig vor und sollte daher zurückgewiesen werden. Leider haben fast alle Bestätigungskonzeptionen mit diesem Problem zu kämpfen und die meisten beißen letztlich in den sauren Apfel, dem auch zuzustimmen. Bayesianer betrachten es schon als Lösung, dass die nicht-schwarzen Nicht-Raben zumindest keine so starke Bestätigung der Rabenhypothese bieten, wie die schwarzen Raben selbst. Meines Erachtens resultiert der Fehler aber vor allem schon aus einem fehlerhaften Nicodschen Prinzip. Nicht jede Instanz bestätigt jede Generalisierung. Nur nomische Konditionale können so bestätigt werden: N ≡ ∀x(Sx≫Rx). Die sind aber im Normalfall nicht äquivalent zu ihrer Kontraposition: N* ≡ ∀x(¬Rx≫¬Sx). Was aber noch schwerer wiegt: N* stellt selbst sicher kein echtes nomisches Konditional mehr dar und kann daher nicht durch seine Instanzen bestätigt werden. Damit tritt das Paradox nicht mehr auf.

In einer ähnlichen Richtung hat schon Quine (1969) nach einer Lösung für das Paradox gesucht. Für ihn sind nur die Generalisierungen projizierbar und damit bestätigbar, die mit Prädikaten für natürliche Arten formuliert sind. Nicht-schwarz und Nicht-Rabe sind keine solchen Prädikate und die entsprechenden Konditionale sind daher nicht durch ihre Instanzen zu bestätigen. In der hier vertretenen Konzeption werden

dagegen immer Paare von Eigenschaften ausgezeichnet, die einen speziellen kausalen Zusammenhang aufweisen, wonach das Auftreten der einen Eigenschaft im Normalfall einen kausalen Mechanismus in Gang setzt, der zum Auftreten der anderen Eigenschaft führt. Das muss selbst für *natürliche Arten* nicht immer der Fall sein, selbst wenn wir schon einige Fälle kennen, in denen sie zusammen aufgetreten sind.

Holistische Induktion. Führt uns dieses Vorgehen sogleich in den Skeptizismus? Wenn wir Daten nur im Lichte weiteren Hintergrundwissens auswerten können oder wenn wir immer gleich auf substantiellere Hypothesen schließen müssen, könnte uns das überfordern. Natürlich wird das induktive Schließen dadurch nicht gerade erleichtert. Doch ein solches eher holistisches Vorgehen ist eben erforderlich und keinesfalls so problematisch wie die grundsätzliche Problematik des Humeschen Problems, das wir im nächsten Abschnitt ansprechen werden. Außerdem würde eine dogmatische Setzung, dass einfach alle Prädikate für alle Arten projizierbar seien, keinen rationalen Ausweg aus unserer Problematik darstellen. Wir müssen also holistischer verfahren und müssen die Daten sogleich mit substantielleren Theorien über die Welt konfrontieren und können nicht nur isolierte materiale Allaussagen betrachten.

Slogan zum induktiven Schließen: Es gibt kein gut begründetes induktives Schließen ohne substantielle Hypothesenbildung.

Das ist eine These meiner Arbeit, dass es keine einfache lineare Extrapolation gibt, sondern wir immer auf die Mitbestätigung nomischer Muster angewiesen sind, wenn wir induktiv schließen möchten. Wir formulieren mehr oder weniger explizit zumindest kleine Hypothesen (etwa in Form nomischer Muster) und die sind auf ihre Kohärenz mit unserem weiteren Hintergrundwissen zu überprüfen. Erst wenn Daten und Hintergrundwissen bestimmte (kleine) Hypothesen stützen, geben uns diese dann die Möglichkeit, zu extrapolieren. Man denke dazu an das Muster von Schokolade und Migräneanfällen oder andere der genannten Beispiele.

Natürlich würden wir uns alle ein einfacheres Verfahren wünschen, in dem die Daten direkt extrapolierbar sind oder uns auf andere Weise

direkt etwas über noch nicht beobachtete Teile der Welt sagen, aber die Beispiele sollten belegen, dass solche Verfahren nicht wirklich sinnvoll sind. Dazu lassen sich zu leicht Prädikate oder Arten angeben, für die keine Projizierbarkeit gegeben ist. Die erscheinen uns meist als nicht wirklich natürlich, aber das ist nur eine Redeweise dafür, dass wir in diesen Fällen andere Erklärungen der Daten vermuten als die durch die »unnatürlichen Muster«. Sind die Prädikate relativ zur betrachteten Art nicht projizierbar, so sollten wir dafür jedenfalls keine Induktionseigenschaft unserer Induktionsverfahren annehmen.

Es gibt dazu eine Reihe von Ideen, wie wir herausfinden können, ob ein stabiles kausales Muster vorliegt, die wir im Folgenden kennenlernen werden. Erste einfache Regeln finden wir schon in unseren Extrapolationen. Möchte ich etwa wissen, wie stabil das Muster ist, dass Wasser bei 100°C anfängt zu kochen, so werde ich die Zuverlässigkeit dieser These nach folgenden Regeln bewerten: Wir benötigen zum einen möglichst *viele Instanzen*, da sie am ehesten einen Hinweis darauf geben, dass hier ein entsprechendes Muster vorliegt. Besonders wichtig ist zum anderen, dass die *Bedingungen variieren*, unter denen ich das Wasser erhitze, denn nur so kann ich erkennen, ob es sich um ein bloß lokales Muster handelt, das nur in einem sehr eingeschränkten Bereich gilt oder um ein globaleres Muster. Außerdem werde ich mich fragen, ob es *andere Erklärungen* für die beobachteten Phänomene (das kochende Wasser) gibt bzw., ob sich diese anderen Erklärungen womöglich ausschließen lassen. Damit gelange ich bereits zur eliminativen Induktion oder auch zum Schluss auf die beste Erklärung, die wir später untersuchen werden.

Dafür sind wir oftmals gezwungen, zunächst weitere *Kausalbeziehungen* aufzuklären, ehe wir auf bestimmte Daten spezielle Prognosen stützen können. Aus einem Zusammenhang wie dem zwischen gelben Fingern und der Erhöhung der Wahrscheinlichkeit für Lungenkrebs dürfen wir nicht einfach schließen, dass beim nächsten Menschen mit gelben Fingern ebenfalls die Lungenkrebswahrscheinlichkeit erhöht ist. Wir wissen vielmehr um die dahinterstehenden kausalen Zusammenhänge, nach denen der entscheidende Aspekt ist, ob die gelben Finger durchs Rauchen verursacht wurden oder auf eine andere Weise. Die gelben Finger sind nur ein (nicht immer zuverlässiger) *Indikator* für

das Rauchen und nur das Rauchen selbst führt zu einer Erhöhung der Lungenkrebswahrscheinlichkeit.

Mindestens das eine haben wir nun gelernt. Das induktive Schließen oder auch das induktive Begründen ist *dreistellig*. Wir stützen uns dabei auf weiteres Hintergrundwissen und das besteht unter anderem aus Hypothesen über kausale und gesetzesartige Beziehungen. Ganz ohne substantielle Annahmen erhalten wir keine begründeten induktiven Schlüsse. Genau das wird besonders vom Schluss auf die beste Erklärung berücksichtigt. Eine Hypothese der Form »A führt zu B« wird demnach nur dann durch entsprechende Daten bestätigt, wenn die Hypothese die Daten auch *erklärt*. Nur für solche Fälle dürfen wir vom Vorliegen von A auf ein Vorkommnis von B schließen. Für Erklärungen sind aber im Normalfall gesetzesartige Kausalbeziehungen erforderlich. Die nächste Frage ist dann natürlich, wann wir spezielle Hinweise dafür haben, dass eine solche Kausalbeziehung vorliegt. Darauf werden wir insbesondere im Kapitel 7 eingehen. Vorher werden wir meist annehmen, dass wir zumindest bereits über grundlegende Vorstellungen von möglichen Kausalbeziehungen in unserer Welt verfügen, auf die wir uns in weiteren induktiven Schlüssen stützen dürfen.

1.7 Das zweite Problem der Induktion: Humes Induktionsproblem

Leider ist sogar die einfache konservative Induktion selbst nicht leicht zu rechtfertigen. David Hume (1888: I.III.VI) hat ein allgemeines Argument gegen jede Form von Rechtfertigung eines Induktionsprinzips (IP) vorgetragen. Dieses Humesche Induktionsproblem möchte ich hier als *zweites Problem der Induktion* diskutieren. Es ist noch grundlegender als das erste Problem und stellt das induktive Schließen in jeder Form in Frage.

Humes Induktionsproblem: Wie können wir irgendein induktives Schlussverfahren oder Induktionsprinzip (IP) begründen? Jede Begründung von (IP) müsste z.B. aus den bisherigen Erfolgen von

(IP) auf seine allgemeine Gültigkeit schließen und würde so selbst einen induktiven Schluss darstellen. Diese Begründung wäre dann aber *zirkulär* und damit erkenntnistheoretisch wertlos.

Nehmen wir an, wir hätten beobachtet, dass die Objekte a, b und c jeweils die Eigenschaft P haben und schließen mit irgendeinem Induktionsprinzip (IP) darauf, dass das nächste Objekt d, das wir beobachten werden, wieder P aufweisen wird. Damit dieses Verfahren nicht willkürlich ist, müssen wir auch (IP) selbst begründen, denn (IP) könnte schließlich eine recht verrückte Form annehmen. Es könnte besagen, dass wir an Montagen immer darauf schließen dürfen, dass Pd gilt und an anderen Tagen auf ¬Pd. Doch woher wissen wir, dass unsere normale konservative Induktion besser funktioniert? Welche Gründe habe ich etwa, für die Zukunft anzunehmen, dass die konservative Induktion erfolgreiche Vorhersagen liefern wird? Um dieses wie immer geartete Prinzip (IP) selbst zu rechtfertigen, müssten wir eine Begründung für die Annahme von (IP) finden.

Diese Begründung kann selbst nur deduktiv oder induktiv sein. Gäbe es eine deduktive Begründung für (IP), so gäbe es damit auch eine deduktive Begründung für unser Argumentationsziel P(d) aus unseren Daten. Das wäre aber ein gehaltserweiternder Schluss und solche können nicht deduktiv sein, denn das sagt uns die deduktive Logik. Also kann es nur eine induktive Begründung für (IP) geben. Doch wie soll die aussehen? Hier droht ein Rechtfertigungszirkel, jedenfalls dann, wenn wir etwa argumentieren, dass wir (IP) deshalb für wahrheitsförderlich halten, weil es sich in vielen Fällen in der Vergangenheit bewährt hat; denn das Verfahren hätte dort zu bestimmten erfolgreichen Vorhersagen geführt. Dann würden wir in unserer Begründung wiederum die konservative Induktion (IP) anwenden und so mit Hilfe von (IP) gerade (IP) selbst rechtfertigen, obwohl wir doch erst wissen wollen, ob wir (IP) tatsächlich anwenden dürfen. Das sieht zirkulär aus und wirkt deshalb nicht sehr vertrauenswürdig.

Man könnte natürlich stattdessen ein anderes Induktionsprinzip (IP*) speziell zur Rechtfertigung von Induktionsprinzipien erster Stufe verwenden, um (IP) anhand seiner vergangenen Erfolge zu rechtfertigen. Doch

dann müssten wir zuerst (IP*) rechtfertigen und dort wiederholt sich unser ursprüngliches Problem. Natürlich könnten wir dafür ein Meta-Meta-Induktionsprinzip (IP**) zum Einsatz bringen, das wir aber wieder erst rechtfertigen müssten etc. Hier geraten wir in einen unendlichen Regress, den die meisten Erkenntnistheoretiker auch nicht für besser als einen Zirkel halten. Das wäre also kein Ausweg.

Ist der vorgeführte Zirkelschluss aber tatsächlich ein bedenklicher Zirkel, wenn man darin dasselbe Prinzip zur Rechtfertigung heranzieht, das man erst rechtfertigen möchte? Ja, denn damit stützen wir uns bereits auf das, was erst zu begründen wäre. Da jedes Argument höchstens so stark ist wie die schwächste Prämisse, haben wir damit noch nichts gewonnen für die Frage, wie gut nun (IP) begründet wurde.

Und es kommt sogar noch schlimmer. Auch ein Vertreter der sogenannten »revolutionären Induktion« könnte sein Verfahren in entsprechender Weise begründen. Der Vertreter der revolutionären Induktion (RI) sagt im Unterschied zur konservativen Induktion: »It's time for a change.« Wenn bisher alle Objekte P aufwiesen, so wird das nächste wohl nicht mehr P sein. Wir kennen dieses Verfahren aus dem sogenannten Fehlschluss des Spielers. Wenn bisher 10-mal Rot kam, so schließen wir gern (aber fälschlicherweise), dass dann wohl als Nächstes Schwarz kommen wird. Im Normalfall ist uns jedoch klar, dass (RI) ein unsinniges Schlussverfahren darstellt. Wenn der Spieler schon vermutet, dass das Roulett nicht fair ist, d.h., dass hier eine Farbe wahrscheinlicher ist als eine andere, dann sollte er lieber auf Rot setzen, denn seine Daten sprechen sich klar für Rot aus. Hält er das Roulett aber für fair, so darf er keiner Farbe den Vorzug geben.

Leider kann der Vertreter von (RI) ganz ähnlich argumentieren, wie die Verteidiger von (PI) oben (vgl. Salmon 1975). Er wird sagen, dass man mit (RI) in der Vergangenheit fast immer schlecht gefahren ist, also (und hier wendet er (RI) auf der Metaebene an) ist zu erwarten, dass (RI) bei der nächsten Anwendung erfolgreich ist. Mit Hilfe von (RI) lässt sich so (RI) begründen. Das Ganze beweist natürlich nur, dass wir es dem Verteidiger von (RI) zu einfach gemacht haben. Niemand sollte sich in der Rechtfertigung eines Induktionsverfahrens bereits auf das Verfahren selbst stützen dürfen. Der Rechtfertigungszirkel ist also vom bösartigen

Typ und sollte nicht gestattet werden. Dann allerdings verfügen wir auch über keine Begründung von (PI) mehr.

BonJour (1998) setzt deshalb auf eine *apriorische Rechtfertigung* induktiven Schließens, die sich nur auf allgemeine Überlegungen zu Wahrscheinlichkeiten und der jeweils besten Erklärung setzt. Aber es zeigt sich dabei (vgl. Bartelborth 2001) an verschiedenen Stellen, dass er doch auf empirische Annahmen angewiesen ist und genau die wird der Skeptiker nicht akzeptieren. Es wäre wohl zu einfach, wenn wir ohne irgendeine substantielle Annahme starten und trotzdem noch begründen könnten, dass bestimmte Schlussfolgerungen auf zukünftige Ereignisse wahr sind oder zumindest eine bestimmte Wahrscheinlichkeit aufweisen. Leider scheint für das Begründen zu gelten, dass es keine »creatio ex nihilo« gibt. Um substantielle Annahmen begründen zu können, sind wir bereits auf einige substantielle Prämissen angewiesen.

Man kann das Induktionsproblem speziell auf eine *empirische Annahme* zuspitzen, die wir als *Gleichförmigkeitsannahme* bezeichnen wollen. Sie besagt, dass die Zukunft der Vergangenheit ähnelt oder dass die noch nicht untersuchten Fälle denen gleichen, die wir bereits untersucht haben. Genauer besagt die Gleichförmigkeitsannahme, dass die Objekte (oder Situationen), die einander ähnlich zu sein scheinen, sich in den meisten Fällen auch ähnlich verhalten (oder sich ähnlich weiterentwickeln). Diese Ähnlichkeit gilt dann insbesondere zwischen Objekten, die wir schon beobachtet haben und solchen, für die das nicht der Fall ist. Diese Gleichförmigkeitsannahme ist eine Voraussetzung dafür, dass wir aus unseren Erfahrungen *lernen* können. Ohne eine derartige Annahme könnten wir nicht induktiv schließen und praktisch nichts aus der Vergangenheit über die zukünftigen Entwicklungen lernen. Wasser könnte sich gestern noch zum Durststillen geeignet haben und heute stattdessen explodieren, sobald wir es trinken. Mit Sprengstoffen könnte es dagegen umgekehrt sein. Unsere Treppen könnten plötzlich weich werden, so dass wir darin versinken etc. Da wir in einer solchen Welt nicht einmal erahnen könnten, in welche Richtung sie sich entwickelt, würden induktive Schlüsse ihre Basis verlieren.

Eine solch verrückte Welt könnte eine *Humesche Welt* für uns sein, wie sie David Lewis (1994) unter dem Stichwort der *Humeschen Supervenienz* schildert. Danach gibt es nur *eine* Struktur in unsere Welt,

die sie zusammenhält, und das ist die *Raum-Zeit-Struktur*. An den Raum-Zeit-Punkten befinden sich sodann intrinsische und kategoriale Eigenschaften, die vor allem untereinander keine Wirkungen aufeinander ausüben können (vgl. auch Esfeld 2008, Kap. 5.1). Keine Instanz einer Eigenschaft bedingt die Existenz einer anderen Instanz einer Eigenschaft. Wir haben nur ein völlig zufälliges Kaleidoskop von Eigenschaften vor uns. Einen Flickenteppich, bei dem ein Flicken nichts darüber sagen kann, wie die anderen Flicken beschaffen sind, denn alle Kombinationen sind möglich und keine ist genuin wahrscheinlicher als eine andere. Es gibt auch keine tatsächlichen Naturgesetze oder kausale Muster, die wir aufdecken könnten. Eine *Kausalstruktur* wäre eine weitere Struktur, die man in einer Humeschen Welt, wie sie Empiristen gern annehmen, nicht vorfindet.

Empiristen wie Lewis haben allerdings einen Ersatzbegriff für gesetzesartige Zusammenhänge eingeführt, den sie auch als *Naturgesetz* bezeichnen. Aber es sind in der Humeschen Welt *Gesetze* für David Lewis nichts anderes mehr als zufällig auftretende Regularitäten, die es gestatten, unsere Welt am besten zu systematisieren. Sie werden durch die Systematisierung aller Ereignisse, die für uns am einfachsten und informativsten erscheint, wenn wir die Welt im Ganzen betrachten, ausgezeichnet. Die lassen sich aber dann erst am Ende der Welt im nachhinein ermitteln. Würden wir diese Regularitäten schon heute kennen, könnten wir natürlich einen Gewinn für unsere Entscheidungen daraus ziehen, denn immerhin müssen die Regularitäten (die auch probabilistischer Natur sein können), auch für die Zukunft gelten. Allerdings lässt sich aus heutiger Sicht noch keine begründete Prognose abgeben, welche »Gesetze« diese Systematisierung einmal leisten werden. Da die Eigenschaften bzw. ihre Instantiierungen in unserer Welt nicht tatsächlich in irgendeiner kausalen Weise untereinander zusammenhängen, könnte unsere Welt genauso gut so chaotisch sein, dass alle unsere bisherigen Erfahrungen uns nichts über die zukünftigen Geschehnisse sagen.

Auch wenn die Welt also bisher sehr regulär für uns erschien, liefert das in einer Humeschen Welt nicht den geringsten Grund zu der Annahme, dass es so bleiben wird. Haben wir dagegen Glück, und es gilt zufälligerweise die Gleichförmigkeitsannahme, so würde unser

induktives Schließen in einer Humeschen Welt ganz zufällig erfolgreich sein. Die Humesche Welt könnte aber jederzeit viele Überraschungen für uns bereit halten und es gäbe in ihr keine systematischen Gründe dafür, aus der Vergangenheit auf die Zukunft zu schließen. Insbesondere böten zufällige Regularitäten auch keine Handhabe, gezielt in die Natur einzugreifen und bestimmte Veränderungen herbeizuführen. Es gäbe eben keine echten Gesetze bzw. kausalen Muster, die wir aufdecken könnten. Genau die sucht die Wissenschaft aber und möchte mit ihrer Hilfe unsere Umwelt verstehen und in sie eingreifen. Auch Erklärungen und Prognosen stützt sie auf die Annahme solcher gesetzesartigen Muster.

Dass Empiristen ernsthaft dafür argumentieren, dass unsere Welt eine solche Humesche Welt sei, zeigt schon, dass die Annahme genuiner nomischer Muster keineswegs selbstverständlich und zwangsläufig und bestimmt nicht a priori wahr ist. Der Humesche Induktionsskeptiker kann sich also zu Recht darauf berufen, dass wir sie erst begründen müssten, wenn wir unsere Extrapolationen oder komplexere induktive Schlussverfahren begründen möchten. Dabei tritt natürlich wieder das *Zirkelproblem* auf. Wenn ich etwa zeigen kann, dass eine Gleichförmigkeitsannahme in der Vergangenheit galt, kann ich erst anhand eines induktiven Schlusses begründen, dass das auch für die Zukunft der Fall sein dürfte. Wir kommen also auch deshalb nicht zu einer sauberen (zirkelfreien) Begründung unserer Induktionsverfahren, da sie sich alle auf eine bestimmte empirische Gleichförmigkeitsannahme stützen müssen, die selbst nicht zirkelfrei begründet werden kann.

In der Wissenschaft gehen wir aber davon aus, dass wir aus unseren Erfahrungen etwas für die Zukunft lernen können. Außerdem nehmen wir an, dass das nicht nur ein Wunder ist, sondern dass dem systematische Zusammenhänge zugrundeliegen. Dabei stützen wir uns meist auf weitere substantielle Annahmen, nach denen bestimmte Naturgesetze bzw. nomische Muster (s. a. Kapitel 3), die in der Vergangenheit zu beobachten waren, auch weiterhin gelten. Nur dann ist das ganze Unternehmen Wissenschaft überhaupt sinnvoll und kann Früchte tragen. In den nächsten Kapiteln werden wir daher diese Annahme ebenfalls akzeptieren und auf ihrer Grundlage nach den geeigneten Induktionsverfahren suchen. Eine substantielle Gleichförmigkeitsannahme

wird somit zu einer Form von Präsumtion (einer Voraussetzung) der wissenschaftlichen Arbeit. Es muss etwas vorhanden sein, wonach wir in der Wissenschaft suchen können, sonst würde sich das ganze Unternehmen Wissenschaft nicht lohnen. Eine solche substantielle Gleichförmigkeitsannahme könnte dann ungefähr so lauten:

Substantielle Gleichförmigkeitsannahme (GA): Die meisten Geschehnisse in unserer Welt werden (in der Vergangenheit wie der Zukunft) von Naturgesetzen oder zumindest von nomischen Mustern bestimmt, die wir im Prinzip erkennen können, so dass wir für viele Eigenschaften und Entwicklungen in unserer Welt eine gewisse Gleichförmigkeit erkennen können, die es uns gestattet, erfolgreiche Vorhersagen für noch nicht beobachtete Fälle abzugeben. (Zum Begriff der nomischen Muster s. Kapitel 3 und Bartelborth 2007)

Diese substantielle Gleichförmigkeitsannahme wird der Skeptiker natürlich in Zweifel ziehen und fragen, wie wir sie wohl begründen können – besonders im Hinblick auf die Zukunft. Selbst wenn unsere Welt in der Vergangenheit von bestimmten Naturgesetzen regiert wurde, muss das nicht heißen, dass das ebenso für die Zukunft gilt und dass es sich in der Zukunft um dieselben Naturgesetze handeln muss. Das können wir wiederum nicht zirkelfrei begründen. Aber wir können einen anderen Weg beschreiten, den bereits Hans Reichenbach (1938) gewählt hat. Wir können vielleicht nicht begründen, dass uns die konservative Induktion überwiegend zu wahren Konklusionen führen wird, aber wir können vielleicht zeigen, dass sie immer noch das Beste ist, was wir in unserer Situation der Unwissenheit bezüglich (GA) tun können. Wie sähe denn unsere Alternative aus?

Der radikale Skeptiker kann uns nur den Ratschlag geben, uns einfach jeder Annahme zu enthalten. Doch wir haben zwei Ziele für unsere Erkenntnis, die beide möglichst weitgehend zu erfüllen sind. Das eine ist, möglichst keine falschen Aussagen zu akzeptieren. Das erfüllt der Skeptiker natürlich hundertprozentig. Aber das andere Ziel ist es, möglichst viele gehaltvolle Einsichten zu gewinnen, wie die Welt funktioniert, um unsere Umwelt verstehen zu können und zu unseren Gunsten

eingreifen zu können. Ohne das zweite Ziel wenigstens ein Stück weit zu erfüllen, ist das ganze Unternehmen Erkenntnistheorie eigentlich für uns wertlos. Das zweite Ziel ließe sich allein auch leicht erfüllen, indem wir einfach alle Aussagen akzeptieren, aber das wäre offensichtlich ebenfalls wertlos. Die beiden Ziele der Erkenntnistheorie sind zwar theoretischer Art, aber trotzdem stehen dahinter natürlich praktische Absichten, die uns zeigen, dass jedes Ziel für sich allein zwar leicht zu erfüllen wäre, aber das resultierende Überzeugungssystem völlig wertlos bliebe. Es geht also darum, das geeignete Maß zu finden, mit dem wir mit einer gewissen Vorsicht bestimmte Aussagen akzeptieren, um beide Ziele möglichst weitgehend erfüllen zu können. Der Skeptiker lässt uns hier nicht weiterkommen, weil er in Frage stellt, ob jemals eine Aussage besser begründet werden kann als eine andere, speziell, wenn es um die Zukunft geht.

Wir können ihm nun entgegenhalten: Wenn wir uns an die skeptische Strategie halten, haben wir keine Chance, das zweite Ziel auch nur ansatzweise zu erfüllen. Unser Überzeugungssystem wird für uns für immer leer und damit wertlos bleiben. Wählen wir dagegen eine andere Strategie, mit der wir wenigstens einige Aussagen (speziell über die Zukunft) akzeptieren, behalten wir zumindest eine gewisse *Chance*, doch noch ein epistemisch wertvolles Überzeugungssystem zu erhalten. Die Größe der Chance können wir natürlich nicht beziffern, sie ist aber immerhin eine logische Möglichkeit. Es scheint dann rational zu sein, diese Chance zu ergreifen und so nach einer Strategie zu suchen, die nach einem bestimmten Verfahren Aussagen zum Akzeptieren auswählt. Dazu kann es natürlich unterschiedliche Auswahlverfahren geben.

Hier kommt unsere Annahme (GA) ins Spiel. Es gibt zum einen die Möglichkeit, dass sie für größere Bereiche unserer Welt gilt. Dann bieten sich die konservative Induktion und vielleicht noch ähnliche Verfahren an, um die Chance möglichst gut zu nutzen. Sollte (GA) allerdings völlig falsch sein und unsere Welt scheint sich – zumindest was die Zukunft angeht – völlig regellos und chaotisch zu verhalten, dann haben wir allerdings wohl nur noch blindes Raten als Verfahren und werden vermutlich kaum weit kommen in der Erfüllung unserer zwei Ziele. Demnach bleibt uns vernünftigerweise nur übrig, darauf zu setzen, dass (GA) im Wesentlichen richtig ist und uns dann auf entsprechende

Induktionsverfahren zu stützen. Das gibt uns die beste Chance zu einem wahren und gehaltvollen Überzeugungssystem zu gelangen.

Ist (GA) dagegen völlig falsch, werden wir mit dieser Strategie nicht viel verlieren, weil die anderen Strategien kaum mehr für unsere beiden Ziele erreichen können. Daher scheint es rational zu sein, zwar die skeptischen Argumente anzuerkennen, aber trotzdem darauf zu setzen, dass (GA) gilt und wir aus der Vergangenheit etwas für die Zukunft lernen können, als bloß zu resignieren und schlicht davon auszugehen, dass ein solches Lernen niemals möglich sei. Diesem erkenntnistheoretischen Optimismus werde ich daher im Folgenden vertrauen (vgl. Bartelborth 2001).

In einer ähnlichen Richtung argumentiert viel ausführlicher und raffinierter Schurz (etwa in 2008a) mit weitergehenden spieltheoretischen Mitteln. Er versucht darin zu zeigen, dass ein Induktionsverfahren, das einfach bei den besten Vorhersagern lernt, zumindest erfolgreicher ist (was Vorhersagen angeht) als die Konkurrenten. Auch das sollte ein Grund sein, sich auf solche Verfahren zu stützen, selbst wenn wir den Skeptiker so natürlich nicht widerlegen können. Eine bessere Antwort auf den radikalen Induktionsskeptiker scheint es leider nicht zu geben. Mit der Herausforderung durch den radikalen Skeptiker werden wir also wohl weiter leben müssen. Im Folgenden werden wir aber gute Nerven zeigen und uns von einem radikalen Skeptiker nicht mehr weiter irritieren lassen. Dazu werden wir ihn vorläufig einfach beiseite stellen und uns der Frage zuwenden, welche Induktionsverfahren sich denn gegenüber einem normal kritischen Verstand begründen lassen, der zumindest davon ausgeht, dass wir im Prinzip aus der Erfahrung lernen können.

1.8 Wichtige Ergebnisse des Kapitels

Zu den wichtigen Ergebnissen des einführenden Kapitels gehört vor allem, dass das induktive Schließen ein sehr schwieriges Geschäft ist. Für die zwei Fragen der Induktion: »Wie sollen wir aus bestimmten Daten induktive schließen bzw. extrapolieren?« und »Wie können wir solche Schlüsse überhaupt rechtfertigen?«, gibt es keine einfachen Antworten.

Der ersten Frage ist das weitere Buch gewidmet. Ein bedeutsames Resultat ist für mich dabei vor allem, dass wir für Induktionsschlüsse immer schon auf weiteres Hintergrundwissen darüber angewiesen sind, an welchen Stellen wir auf kausale Zusammenhänge oder nomische Muster und Konditionale schließen dürfen bzw. welche Eigenschaften in unserer Welt vermutlich projizierbar sind und welche nicht. Das können wir natürlich nicht bereits vor unserem anderen Wissen über die Welt herausfinden, sondern immer nur zugleich mit unseren einfacheren Annahmen über die Welt als begründete Hypothesen formulieren.

Das induktive Schließen ist dadurch ein *holistisches Unternehmen*, das nur funktioniert, wenn wir letztlich bereit sind, substantielle Hypothesen darüber, wie die Welt funktioniert, aufzustellen und dann zu testen. Diese Hypothese über die Kausalstruktur unserer Welt und die Naturgesetze in ihr müssen dabei soweit gehen, dass strenge Empiristen jedenfalls z.T. von metaphysischen Hypothesen sprechen würden und es deshalb ablehnen würden, sich damit zu beschäftigen. Wir werden aber auch im Folgenden immer wieder erkennen, dass das induktive Schließen nicht als ein rein empiristisches Unterfangen – als eine Art von Algorithmus, der aus Daten Hypothesen generiert oder zumindest ihre Begründung erzeugt – durchzuführen ist. Diesen Aspekt werden wir manchmal (in der Hitze der Debatte um bestimmte Aspekte der diskutierten Ansätze) aus den Augen verlieren, aber wir sollten ihn zumindest immer im Hinterkopf behalten.

1.9 Der weitere Plot des Buches

Nachdem wir einfache Extrapolationen und ihre begrenzten Einsatzmöglichkeiten für das induktive Schließen kennengelernt haben, wird es in Kapitel 2 um wissenschaftlichen Fortschritt gehen und dann ein Abstecher in die Wissensdebatte folgen. Es geht darum, eine Zielvorstellung speziell für wissenschaftliches Wissen zu entwickeln, die wir im Folgenden als Leitbild für unsere Suche nach Begründungsverfahren betrachten können. Die vorgelegte Wissenskonzeption formuliert ideale Anforderungen an wissenschaftliches Wissen, die meist kaum erreicht werden können, wird gleichwohl aber als ideales Ziel der Wissenschaften

hilfreiche Dienste leisten. Insbesondere besagt sie, dass wir für Wissen alle *relevanten Unterminierer* auszuschalten haben, und damit sind – wie wir sehen werden – in der Wissenschaft vor allem die *Konkurrenztheorien* zu unserer Hypothese gemeint. Das wird ein wichtiges Leitmotiv sein, das wir durch alle Begründungsansätze verfolgen werden. Ansätze, die das nicht leisten können, dürfen wir getrost als defizitär ansehen. Letztlich muss jede Konzeption induktiven Schließens uns eine Antwort auf die Frage nach der Rolle der Konkurrenten einer Theorie geben.

Den Ausgangspunkt der weiteren Induktionsdebatte in Kapitel 3 wird dabei Poppers *Falsifikationismus* bilden. Popper versuchte eine Antwort auf Humes skeptische Einwände zu finden, indem er die klassischen Formen induktiver Bestätigung zu umgehen hofft und ganz auf *Falsifikationen* von Theorien setzt. Der Fortschritt der Wissenschaften besteht dann darin, dass sie allmählich die falschen Theorien eliminieren und sich so der Wahrheit nähern. Das scheint den Gedanken, dass wir die Konkurrenz zu eliminieren haben, schon in vorbildlicher Weise umzusetzen. Doch leider kommt auch Popper letztlich nicht darum herum, sich auf *direkt positiv begründete* Annahmen zu stützen, da uns Falsifikationen allein unseren Zielvorstellungen in der Wissenschaft ebenfalls nicht wirklich näher bringen. Daher werden wir seine Idee der Falsifikationen in den Rahmen der *eliminativen Induktion* einbringen. Dort besitzen Falsifikationen zusätzlich einen positiv bestätigenden Aspekt im Hinblick auf die noch verbleibenden Theorien, aber nur, wenn wir vorher bereits eine vollständige (endliche) Liste von konkurrierenden Theorien absegnen, worin Popper uns nicht folgen würde. Falsifikationen deuten in diesem Rahmen indirekt auf ganz bestimmte Theorien hin, indem sie die direkten Konkurrenten eliminieren. Diese Vorgehensweise werden wir mehr oder weniger deutlich in allen vollständigen Konzeptionen induktiver Rechtfertigung wiederfinden.

Neben diesem indirekten Schließen müssen wir ebenso die *direkten Bestätigungen* einer Theorie durch ihre erfolgreichen Vorhersagen berücksichtigen. Das geschieht in seiner grundlegenden Form anhand der *hypothetisch-deduktiven Theorienbestätigung*. Dort zählen die Daten positiv für eine Theorie, die sich aus der Theorie (zusammen mit bestimmten Hilfsannahmen) ableiten lassen und die sich dann mit

unseren tatsächlichen Beobachtungen decken. Viele Wissenschaftler nennen das als ihre Methode, ihre Theorien abzusichern.

Leider weist der hypothetisch-deduktive Ansatz in seiner Grundform eine ganze Reihe von Problemen auf. Eines der Hauptprobleme ist, dass die deduktive Ableitung eines Datums aus Theorie plus Hilfsannahmen noch wenig über einen inhaltlichen Zusammenhang zwischen der Theorie und dem Datum sagt. Um das zu verbessern, wurde u.a. vorgeschlagen, mit Relevanzlogiken zu arbeiten. Ein anderer Weg ist der Übergang zum umfassenderen Schluss auf die beste Erklärung, der alle bisherigen Ideen in sich vereint. Danach muss eine Theorie vor allem zu der *Erklärung* eines Datums beitragen, damit wir sagen können, dass sie durch das Datum bestätigt wird. Die Theorie, die die meisten Daten am besten erklärt, wird durch diese Daten dann auch am meisten gestützt.

Ähnlich wie bei der eliminativen Induktion benötigen wir für den *Schluss auf die beste Erklärung* oder das abduktiven Schließen zunächst ebenfalls eine möglichst vollständige Liste potentiell erklärender Theorien, die wir anschließend durch Eliminationen langsam verkleinern müssen (vgl. Kapitel 4). Diese Eliminationen müssen nun aber nicht unbedingt anhand von logischen Inkonsistenzen erfolgen, sondern es genügt, wenn eine Theorie bestimmte Daten partout nicht erklären kann. Auch der geforderte inhaltliche Zusammenhang zwischen Daten und Theorien wird durch die Erklärungsbeziehungen deutlich. Sie zeigen zudem ein weiteres Ziel der Wissenschaften auf. Es geht uns nicht nur darum, gut bestätigte Theorien zu finden, die ihre Konkurrenten aus dem Feld schlagen, sondern unsere Zieltheorien müssen zugleich eine hohe Erklärungskraft besitzen, damit sie für unser Verständnis der Welt und ein mögliches Eingreifen außerdem noch hilfreich sind. Das erkenntnistheoretische Ziel wird dabei u.a. sein, grundlegende Kausalstrukturen unserer Welt aufzudecken.

Dieser abduktive Ansatz zeigt schon erste Vergleichsmöglichkeiten in den Bestätigungen verschiedener Theorien durch bestimmte Daten auf. Die Vergleichsmöglichkeiten hängen vor allem davon ab, dass wir genauer explizieren, was die Gütekriterien für Erklärungen sind. Außerdem wird deutlich, dass Theorien auch mit anderen Theorien, die wir akzeptieren, zusammenhängen bzw. gestützt werden, und wir benötigen

eine holistische Bewertung der epistemischen Gesamtverdienste unserer Theorie, die all diese Zusammenhänge mit einbezieht.

Das geschieht im Rahmen einer Konzeption von *Erklärungskohärenz*, die es gestattet, alternative Szenarien im Hinblick auf ihre Kohärenz zu vergleichen. Die Elimination von Theorien erfolgt dann in zwei Stufen: Zunächst werden die ganz unplausiblen Theorien ausgesondert, deren Akzeptanz zu weitergehenden Inkohärenzen in unserem Überzeugungssystem führen würde. Danach werden an sich plausible Theorien ausgeschlossen, die bestimmte Erklärungsanomalien aufweisen und im letzten Schritt erfolgt dann ein kleinteiliger Vergleich der Erklärungsstärke der noch verbliebenen Theorien. Diese Form von Abduktion stellt einen bedeutsamen Rahmen dar, in dem wir auch die weiteren Ansätze betrachten werden.

Als wichtigster Konkurrent des direkten Schlusses auf die beste Erklärung ist das bayesianische Updateverfahren und seine Verwandten zu nennen, das wir im Kapitel 5 kennenlernen werden. Hier wird zunächst jeder Theorie ein Plausibilitätsgrad bzw. eine Wahrscheinlichkeit zugewiesen und diese im Lichte neu hereinkommender Daten nach einem festen Verfahren jeweils neu angepasst. Diese Vorgehensweise bietet eine Vielzahl neuer Einsichten und scheint auf den ersten Blick nur wenig mit der eliminativen Induktion zu tun zu haben. Doch an einigen Stellen sind wir wieder auf eine vollständige Liste der Konkurrenten und ihre schrittweise probabilistische Falsifikation angewiesen. Der Bayesianismus bietet zugleich einen guten Ausgangspunkt, um die Besonderheiten der klassischen Testtheorie mit ihren Signifikanztests besser zu verstehen, denen wir uns in Kapitel 6 zuwenden.

In allen Schlussformen wird immer wieder deutlich, dass wir meist nach kausalen Zusammenhängen suchen; denen wird daher noch einmal speziell das Kapitel 7 gewidmet. Für einfache Kausalzusammenhänge können wir direkte Schlussverfahren angeben, mit denen sie erschlossen werden können. Dazu werden vor allem zwei Verfahren vorgestellt, bei denen das eine von einem deterministischen Hintergrund ausgeht, während das andere Verfahren auch probabilistische Zusammenhänge modellieren kann. Beide Verfahren sind moderne Nachfahren des millschen Differenztests, bieten aber viele neue Erkenntnisse für dieses

basale Anliegen der Wissenschaft und bilden somit den würdigen Abschluss eines Buches über induktives Schließen.

2 Erkenntnistheoretische Zielvorstellungen

Bevor ich mich den spezielleren Induktionsverfahren zuwenden kann, sind einige allgemeinere erkenntnistheoretische Fragen zu klären, die zugleich Auskunft über den Sinn und Zweck von Induktionsschlüssen geben. Zunächst können wir (s.o.) zwischen *Induktionsschlüssen* und *induktiven Rechtfertigungen* bzw. *Begründungen* unterscheiden. Wenn wir von Schlüssen sprechen, haben wir damit typischerweise den Fall vor Augen, dass wir aus bestimmten Prämissen eine neue Schlussfolgerung nach gewissen Regeln ableiten, wie das etwa in der deduktiven Logik der Fall ist.

Das mag auch in manchen Fällen gelingen, aber oft haben wir es in der Wissenschaft mit Fällen zu tun, die eher in den *Rechtfertigungskontext* als in den *Entdeckungskontext* fallen. Das heißt, wir können nicht einfach neue Aussagen also speziell neue Hypothesen aus den Daten ableiten, sondern nur zu bereits vorhandenen Hypothesen entscheiden, ob die vorgelegten Daten oder Prämissen dazu passen und diese Hypothesen stützen oder eben nicht. Wie wir auf neue Hypothesen kommen, ist nicht das primäre Geschäft der wissenschaftstheoretischen Induktionsdebatte. Ich werde zwischen diesen beiden Redeweisen nicht strikt unterscheiden und meine oft die *epistemische Rechtfertigung*, wenn ich von Induktionsschlüssen spreche. Es geht dann mehr um die logische Beziehung zwischen Prämissen und Konklusionen und nicht um ein Entdeckungsverfahren für die Konklusionen.

Allerdings werden wir daneben einige Verfahren kennenlernen, die für das Aufdecken neuer Hypothesen zumindest Anhaltspunkte zu bieten haben oder sie sogar im Normalfall herleiten können. Dazu haben wir bereits die konservative Induktion eingeführt, die aus bisherigen Daten schlicht zu extrapolieren versucht, oder Regressionsverfahren, die aus Daten direkt auf zugrundeliegende funktionale Zusammenhänge schließen. Auch einige Verfahren zur Aufdeckung von Kausalbeziehungen

gehen den direkten Weg, der von vorgelegten Daten selbstständig zu Kausalbehauptungen führt.

Die meisten Schlussverfahren sind eher *indirekt*. Sie gehen davon aus, dass schon bestimmte Hypothesen vorliegen und wir nur noch fragen müssen, ob diese zu den Daten passen und sogar dadurch gestützt werden. Oft ziehen wir z.B. Schlussfolgerungen aus einer Theorie und fragen uns dann, ob diese mit den beobachtbaren Tatsachen übereinstimmen. Diese Verfahren wie das gerade skizzierte hypothetisch-deduktive Schließen oder auch die Falsifikationsansätze setzen offensichtlich voraus, dass eine Theorie bereits gegeben ist, aus der sich dann Schlussfolgerungen ziehen lassen. Der Schluss auf die beste Erklärung bietet beides. Zum einen gibt er uns manchmal Hinweise auf sinnvolle Hypothesen, zum anderen lassen sich Hypothesen dadurch begründen, dass sie die besten Erklärungen unserer Daten liefern.

Dass beides nicht einfach zusammenfallen muss, zeigt das folgende Beispiel, das immer wieder in ähnlicher Form kolportiert wird. Der Wahrheitsgehalt ist für uns aber nicht wesentlich, denn es genügt, dass eine derartige Geschichte durchaus denkbar ist: Der Chemiker Kekulé, der das System der chemischen Strukturformeln entwickelt hat, beschrieb einige Jahre später, wie er im Jahre 1865 die spezielle Strukturformel für das von Faraday entdeckte Benzol fand. Während einer Reise hatte er einen Tagtraum, in dem Ketten aus Kohlenstoffatomen wie lebende Wesen herumtanzten und sich plötzlich zusammenrollten wie eine Schlange, die sich in den Schwanz beißt. Das brachte ihn auf den entscheidenden Gedanken: Das Benzolmolekül muss *ringförmig* sein. Diese Überzeugung Kekulés konnte er später anhand einer entsprechenden Strukturformel und entsprechender Daten bestätigen.

Hier haben wir einen Fall vor uns, bei dem die *Genese einer Hypothese* nicht viel zu tun hat mit der Frage ihrer *epistemischen Rechtfertigung*, denn ein Tagtraum wird kaum als gute Begründung für eine wissenschaftliche Behauptung herhalten können. Für das Erfinden der Hypothese ist vor allem die Kreativität des Wissenschaftlers gefordert, während die Aufgabe ihrer Begründung eine sorgfältige Analyse der Daten und die Anwendung von bestimmten Begründungsverfahren verlangt.

Die Induktionsverfahren lassen sich noch auf eine andere Weise unterteilen. Einige Ansätze (wie etwa der Likelihoodismus) zielen nur

darauf ab, eine *vergleichende Bewertung* zwischen zwei Hypothesen abzugeben, während andere wie der Bayesianismus oder die induktive Logik einen absoluten Bestätigungsgrad für alle Hypothesen angeben. Überhaupt gibt es Verfahren, die eher zu der *Auswahl einer Theorie* führen und solche, die stattdessen bloß quantitative *Grade der Bestätigung* anführen, ohne eine entsprechende Abtrennungsregel z, die uns sagt, unter welchen Bedingungen wir eine Theorie schließlich akzeptieren sollten. Doch dazu später mehr. Zunächst möchte ich die Zielvorstellungen für die wissenschaftliche Forschung klären. Dabei stoßen wir auf bestimmte Aspekte, die von einigen Ansätzen übersehen werden.

2.1 Die zwei Ziele der empirischen Wissenschaften

Was sind die Ziele der Wissenschaft? Oder gibt es nur eines? Das ist die grundlegende Frage, wenn wir ermitteln möchten, ob bestimmte Methoden für die Wissenschaften geeignet sind oder nicht. Man kann schon auf den Namen verweisen und zunächst antworten: Der Wissenschaft geht es um die Schaffung speziellen *Wissens*. Wissen ist dabei als höchste Form der Erkenntnis gemeint, auf die wir unsere Entscheidungen und Handlungen stützen können und auf die wir uns verlassen können, wenn wir andere Behauptungen begründen oder widerlegen möchten. Nur wenn wir so hohe Maßstäbe in der Wissenschaft anlegen, wird auch verständlich, weshalb Staaten bereit sind, für das Unternehmen Wissenschaft so viel Geld auszugeben, und warum auf die wissenschaftliche Ausbildung junger Menschen in den westlichen Ländern so viel Wert gelegt wird.

Letztlich versprechen wir uns von der Wissenschaft Antworten auf sehr viele Fragen wie etwa: Was können wir gegen bestimmte Krankheiten tun? Wie können wir günstig Energie gewinnen? Was können wir gegen Kriminalität in unserer Gesellschaft unternehmen? Wie einen wirtschaftlichen Aufschwung befördern? Was befördert das Lernen unserer Kinder am meisten? In allen derartigen Fragen sollte das wissenschaftliche Wissen letztlich auch praktische Konsequenzen haben. Zunächst wollen wir zwar einfach nur verstehen, warum bestimmte Dinge passieren, aber

oft möchten wir darüber hinaus Einfluss auf das Geschehen nehmen und dazu benötigen wir verlässliche (wissenschaftliche) Vorhersagen darüber, welche Folgen welche Maßnahmen haben würden, um dann gezielt in den Gang der Dinge eingreifen zu können.

Das wichtigste Ziel der (empirischen) Wissenschaften ist dafür das Ermitteln der Wahrheit. Die Menge der akzeptierten Aussagen einer Disziplin – nennen wir sie ihre *Akzeptanzmenge* – soll als Zielvorstellung keine falschen Aussagen, aber alle wahren Aussagen enthalten. Das sind genaugenommen dann schon zwei Ziele. Da sich beide Ziele in der Praxis nicht perfekt realisieren lassen und wir auch nicht direkt feststellen können, welche Aussagen wahr und welche falsch sind, können wir die Ziele zunächst so formulieren:

1. Ziel: Die Akzeptanzmenge soll keine falschen oder zumindest möglichst wenige falsche Aussagen enthalten.

2. Ziel: Die Akzeptanzmenge soll alle (relevanten) wahren Aussagen oder zumindest möglichst viele davon enthalten.

Natürlich geht es gerade in der Wissenschaft nicht darum, irgendwelche wahren Aussagen zu akzeptieren. Die meisten wahren Aussagen sind völlig irrelevant für die entsprechende Disziplin. Im Zentrum wissenschaftlichen Wissens stehen vielmehr ganz bestimmte Hypothesen und Theorien, die uns im Idealfall die gewünschten Erklärungen und Vorhersagen innerhalb einer Disziplin liefern können. Sie beinhalten das komprimierte Wissen der Wissenschaft, indem sie Auskunft über die grundlegenden (kausalen) Zusammenhänge in unserer Welt geben. Wir können das stark vereinfacht so ausdrücken: Relevant sind für eine empirische Disziplin vor allem die *Kausalgesetze* dieser Disziplin.

Zumindest können wir an dieser Stelle bereits gut erkennen, dass die beiden Ziele in einem Spannungsverhältnis zueinander stehen. Jedes Ziel ist für sich leicht erfüllbar. Das erste Ziel können wir für sich erreichen, indem wir einfach keine Aussagen akzeptieren, das zweite, indem wir schlicht alle Aussagen akzeptieren. Beide Strategien sind für sich genommen natürlich völlig wertlos. Spannend ist erst eine ausgewogene Berücksichtigung beider Ziele.

Bevor Wissenschaftler eine Theorie propagieren können, müssen sie zunächst Beweise dafür sammeln. (Da das Wort »beweisen« in der Logik

und Mathematik eine recht spezielle Bedeutung hat, die mehr dem deduktiven Begründen entspricht und hier gerade nicht gemeint ist, werde ich in Zukunft eher davon sprechen, dass die Wissenschaftler Belege sammeln müssen oder *Gründe für ihre Theorien* anzugeben haben u.ä.) Sind die Belege noch nicht so stark, sprechen wir eher von Hypothesen, bei stärkeren Belegen von Theorien. Da ich glaube, dass es sich hier bloß um kontinuierliche Abstufungen handelt, werde ich allerdings beide Ausdrücke synonym verwenden und zusätzlich über die Stärke der Begründungen dieser Theorien sprechen.

Weil wir jedoch im Normalfall keinen direkten Zugang zur Wahrheit unserer Theorien haben, sind wir auf diese (indirekten) Hinweise bzw. Belege für ihre Wahrheit angewiesen. Wenn genügend Belege für eine Theorie vorliegen, die Hinweise darauf geben, dass sie wahr ist, kann sie als gut begründet gelten, und sollte von der Wissenschaftlergemeinde zumindest vorläufig (etwa bis weitere Daten vorliegen), akzeptiert werden, um dann auf ihrer Grundlage Entscheidungen treffen zu können. Es geht in der Wissenschaft also zwar um Wahrheit und Wissen, aber unser bester Weg dorthin führt über gut *begründete* Theorien. Daher wird es in diesem Buch vor allem darum gehen, was *gute Gründe für eine Theorie* sind, wann sie dafür hinreichen, eine Theorie zu *akzeptieren* und welche anderen Überlegungen für die Auswahl von Theorien sonst noch relevant sind. Das erste Ziel für die Wissenschaft stellt sich damit aus der Innenperspektive des Wissenschaftlers leicht verändert beschrieben so dar, dass er Aussagen oder Theorien akzeptieren sollte, die für ihn gut begründet sind. Die Akzeptanzmenge sollte demnach nur gut begründete Theorien enthalten.

Zuletzt haben wir hauptsächlich darüber gesprochen, dass die Theorien gut begründet sein sollten, d.h., dass wir also über gute Hinweise dafür verfügen, dass diese Theorien wahr sind. Doch wir hatten gerade erkannt, dass das alleine als Ziel für die Wissenschaft nicht ausreicht. Nur die Theorien sind für uns epistemisch wertvoll, die uns zugleich *wertvolle Informationen* über die Welt liefern, etwa darüber, wie die kausale Struktur der Welt ist und wo wir ansetzen können, um in die Welt *einzugreifen*, oder schlichter Informationen der Art, dass wir *verstehen* können, warum etwas passiert ist. Würden wir auf diese Anforderung verzichten, wäre die Auswahl gut begründeter Theorien recht einfach.

Wir könnten uns im Extremfall auf logisch oder mathematisch nachgewiesene Aussagen beschränken, die uns allerdings für sich nichts über das Funktionieren der empirischen Welt sagen. Oder wir könnten uns auf die »Theorien« beschränken, die nur eine Konjunktion unserer Daten $D_1,\ldots,D_n$ darstellen: $T \equiv D_1\&\ldots\&D_n$. Solche *reinen Datentheorien* sind dann zwar durch die Daten optimal gut begründet (sogar deduktiv aus ihnen ableitbar), aber sie geben keine Vorhersagen ab und können nichts erklären. Solche »Theorien« sind wissenschaftlich völlig wertlos. Das ist ein zweiter Aspekt der Theorienwahl, der leider oft genug in der Debatte übersehen wird.

Vor allem eine konsequente Anwendung rein empiristischer Ideen droht bei diesen Datentheorien zu landen. Entscheidet etwa nur die Wahrscheinlichkeit einer Theorie im Lichte unserer Daten über die Wahl einer Theorie, so sind die Datentheorien mit Wahrscheinlichkeit eins natürlich die klaren Gewinner. Haben wir also für die *Theorienwahl* nur eine induktive Logik im Sinne von Carnap (oder eine entsprechende epistemische Wahrscheinlichkeit) zur Verfügung, so bleibt unklar, warum wir uns dabei Theorien zuwenden sollten, die eine geringere Wahrscheinlichkeit aufweisen als unsere Datentheorien. Der Empirist muss uns erklären, wie er andere Werte von Theorien angemessen berücksichtigen kann.

Das war eine wichtige Kritik von Popper an den induktiven Vorgehensweisen der Empiristen. Popper betonte gegen die Empiristen, dass wir nach *tiefen Theorien* suchen müssen (m.E. meint er damit vor allem *erklärungsstarke* Theorien), die besonders *gehaltvoll* und *kühn* sind und damit natürlich – das war ihm ganz klar – eine geringere Wahrscheinlichkeit aufweisen. Sie sollten stattdessen einen möglichst hohen Grad an *Falsifizierbarkeit* aufweisen. Da es für Popper auf der anderen Seite keine induktive Bestätigung von Theorien gibt (s. nächstes Kapitel), verlor er z.T. ganz das erste Ziel aus den Augen. Sein Konzept der Bewährung von Theorien erweist sich leider nicht als brauchbares Ersatzkonzept für den Begriff der Begründung (mehr dazu im nächsten Kapitel).

Während die Empiristen also eher nach empirisch sehr gut begründeten und damit wahrscheinlich wahren Theorien suchten, betonte Popper das andere Ziel einer möglichst kühnen Wissenschaft. Es ist

daher kein Wunder, dass die empiristisch gesinnten Ansätze eher auf der Suche nach recht beobachtungsnahen Theorien sind, während ein guter Wissenschaftler im Sinne Poppers eine phantasievolle Weiterentwicklung unserer Theorien propagiert.

Es sind somit *zwei Ziele* der Wissenschaft involviert, die einen gegenläufigen Effekt haben. Zum einen möchten wir möglichst gut begründete und sichere Theorien ohne Irrtumsrisiko; auf der anderen Seite können wir gehaltvolle und das heißt meist erklärungsstarke und tiefliegende Theorien nur gewinnen, wenn wir ein gewisses Risiko eingehen, dass wir auch daneben liegen könnten. Hier sollte eine komplexe Verrechnung stattfinden, die uns weiter beschäftigen wird.

Die zwei Ziele der Wissenschaft sind:

(1) Möglichst nur *wahre* Theorien zu akzeptieren (die letztlich Wissen darstellen sollen), und das bedeutet, nur *möglichst gut begründete* Theorien zu akzeptieren, aber auch

(2) möglichst *gehaltvolle* (erklärungs- und vorhersagestarke) Theorien zu gewinnen.

Das Spannungsverhältnis zwischen diesen beiden Zielen bestimmt das wissenschaftliche Forschen. Sobald wir mehr Aussagen (insbesondere gehaltvolle Theorien) akzeptieren, glauben wir zwar mehr zu verstehen und damit ebenfalls mehr erklären zu können, aber damit steigt automatisch unser Irrtumsrisiko an. Halten wir uns stattdessen nur an die Daten und riskieren keine weitergehende Theorienbildung (was uns die Empiristen genau genommen nahelegen), dann verspielen wir jede Chance, tiefere Einsichten in die Welt zu erhalten und werden niemals gezielt in unsere Welt eingreifen können.

Das induktive Schließen dient letztlich der Auswahl unserer besten Theorien und der möglichst besten Verwirklichung der beiden Ziele und muss daher beiden Zielen gerecht werden und damit eine komplexe Abwägung der beiden Risiken bzw. der Erkenntnisgewinne vornehmen. Jedes der Ziele ist für sich allein leicht erfüllbar (akzeptiere nichts oder akzeptiere alles) und erst in ihrer Kombination zeigt sich die wahre Aufgabe der Wissenschaften.

Man spricht von der wissenschaftlichen Forschung deshalb auch als *kontrollierter Spekulation*. Zum einen ist die kreative Phantasie der Wissenschaftler gefragt, sich komplexe theoretische Modelle dafür auszudenken, wie die grundlegenden Kausalkräfte beschaffen sind, die zu den beobachtbaren Phänomenen in unserer Welt führen, auf der anderen Seite müssen wir diese Spekulationen aber auch mit den Daten konfrontieren und diejenigen wieder verwerfen, die nicht dazu passen. Die theoretische Physik konstruiert so phantasievolle und weit entwickelte Modelle wie die Stringtheorie (nach der unsere Welt im Innersten aus kleinen »Fäden« besteht, die auf komplexe Weise in vielen Dimensionen schwingen), die so tief gehen, dass sie viele andere physikalische Theorien nun vereinigen können. Zum anderen fragt dann aber die Experimentalphysik, welche genauen Daten wir angeben können, die speziell für eine der Stringtheorien sprechen. Die Physik ist jedenfalls insgesamt sehr gut damit gefahren, die Spekulation nicht zu kurz kommen zu lassen, aber natürlich auch immer nach der empirischen Kontrolle unserer Spekulationen zu rufen.

Eine Aufgabe der Wissenschaftstheorie ist, weiter zu explizieren, nach welchen Informationen wir gezielt suchen, wenn wir von informativen und *erklärungsstarken Theorien* sprechen. Es wird darum gehen (s. Kapitel 4), *stabile Regelmäßigkeiten* in der Natur zu finden – oder sogar *Naturgesetze* –, denn nur mit ihrer Hilfe werden wir erklären können, warum gewisse Ereignisse bestimmte andere Ereignisse hervorgebracht haben. Wir müssen die *kausale Struktur* unserer Welt aufdecken, um auf der Grundlage dieser Kenntnisse *Vorhersagen* abgeben zu können und die gewünschten Veränderungen *bewirken* zu können. Das wird besonders an den Stellen deutlich, wo es darum geht zu verstehen, warum bestimmte Krankheiten oder Unfälle aufgetreten sind, die wir in Zukunft gerne vermeiden möchten.

Erst als die Menschen verstanden haben, warum die schlimmen Cholera-Epidemien in den großen Städten aufgetreten sind, waren sie in der Lage, diese durch entsprechende Hygienemaßnahmen wirksam einzudämmen. Dazu war es vor allem erforderlich, die genaue Ursachen-Wirkungskette aufzudecken. Die zugrundeliegende Infektionstheorie war zur damaligen Zeit – das Beispiel werden wir noch genauer kennenlernen – äußerst phantasievoll und damit keine sehr wahrscheinliche Theorie,

aber das epistemische Risiko auf diese Theorie zu setzen wurde eben durch den Erkenntnisgewinn aufgewogen, den die Theorie mit sich brachte vor allem in Bezug auf ihre Erklärungsstärke.

Eventuell kommt neben dem Akzeptieren einer Theorie bzw. der Theorienwahl (oder stattdessen) noch hinzu, dass wir bestimmte *Glaubensgrade* oder *Plausibilitätsgrade* von Theorien angeben können und uns nicht dafür entscheiden müssen, sie vollständig zu akzeptieren. Das macht unser Geschäft jedoch nicht immer einfacher, wird aber von manchen Ansätzen wie dem Bayesianismus bevorzugt, während z.B. die klassische Statistik ganz auf das kategorische Akzeptieren von Theorien setzt. Uns geht es dabei auch nicht nur um Wahrheit, sondern um gut *begründete* Wahrheiten, und damit meinen wir letztlich *Wissen*. Es scheint mir durchaus interessant zu sein, dieses Ziel noch weiter zu klären, selbst wenn das Ziel dadurch in noch weiterer Ferne erscheint, denn Wissen verlangt eine ganze Menge von uns. Aber selbst wenn es ein schwer erreichbares Ideal für die Wissenschaft darstellt, sollten wir doch skizzieren, worum es uns dabei geht.

2.2 Fortschritt und Wahrheitsnähe

Das Problem der zwei Ziele der Wissenschaft, die wir gegeneinander abwägen müssen, können wir auch auf anderem Wege beschreiben. Es geht uns in der Wissenschaft nicht nur darum, einen *Teil der Wahrheit* aufzuspüren, sondern wir möchten nach Möglichkeit die *ganze Wahrheit* erfahren. Das sollten wir zumindest als Zielvorstellung akzeptieren. Bei dieser Annäherung an die ganze Wahrheit kann es passieren, dass wir *Fortschritte* anhand einer Folge von teilweise falschen Theorien erzielen, die sich aber dennoch der Wahrheit annähern. Jedenfalls liegt es nahe, Fortschritte in der Wissenschaft als eine *Annäherung an die Wahrheit* zu verstehen und dabei kann es sogar passieren, dass der Übergang von einer wahren Theorie zu einer falschen Theorie trotzdem einen Fortschritt darstellt.

Was mit Annäherung an die Wahrheit gemeint ist und wieso dann solche Phänomene auftreten können, lässt sich am besten anhand einer ziemlich vereinfachten Modellwelt mit einer einfachen Sprache erläutern.

Diese Welt beschreiben wir zunächst durch singuläre Basisaussagen und approximieren Allaussagen oder Naturgesetze durch entsprechende längere Konjunktionen von singulären Behauptungen. Außerdem gehen wir davon aus, dass diese empirischen singulären Aussagen jeweils etwas Neues über die Welt zu sagen haben und logisch voneinander unabhängig sind. Wenn unsere Modellwelt etwa durch die Basis-Aussagen $r_1,\ldots,r_n$ entsprechend beschrieben wird, dann gibt es eine komplexe Gesamtwahrheit vom Typ: $\pm r_1 \& \ldots \& \pm r_n$ (eine Vollkonjunktion), die unsere Welt korrekt beschreibt, der wir uns möglichst annähern sollten. Diese Vollkonjunktionen sind so gemeint, dass wir hier alle Konjunktionen der n Aussagen bzw. ihrer Negationen betrachten wie z.B. $r_1 \& \neg r_2 \& \ldots \& \neg r_n$. Um die ganze Wahrheit über unsere Welt zu wissen, genügt es nicht, einzelne der Konjunkte zu kennen, sondern unser Ziel ist es, möglichst für alle Aussagen r_i zu entscheiden, ob sie wahr oder falsch sind.

Wird unsere Modellwelt etwa vollständig korrekt beschrieben durch die Aussage $r_1 \& r_2 \& \ldots \& r_{10}$ und wir haben zwei Theorien: T1 $\equiv r_1$ und T2 $\equiv r_1 \& \ldots \& r_9 \& \neg r_{10}$, dann scheint es schon plausible, dass T1 zwar wahr und T2 falsch ist, aber T2 wegen seines viel größeren Wahrheitsgehalts trotzdem *wahrheitsnäher* ist als T1.

Vereinfachen wir nun unsere Modellwelt noch weiter und nehmen an, unsere Welt (oder die für uns im Moment relevanten Teilbereiche davon) ließe sich durch nur drei logisch unabhängige Aussagen p, q und r vollständig beschreiben. Dann würden wir gerne für alle drei Aussagen wissen, ob jeweils sie oder ihre Negation wahr ist. Wir erhalten somit 8 unterschiedliche, vollständige Welt- oder Zustandsbeschreibungen: $\pm p \& \pm q \& \pm r$. Nennen wir die 8 resultierenden Vollkonjunktionen unsere *möglichen Welten*. Für unsere aktuale Welt gelte die vollständige Beschreibung: p & q & r. Ziel der Wissenschaft ist es nun also, sich dieser *vollständigen Wahrheit* so weit wie möglich anzunähern. Akzeptieren wir etwa p und sind unentschieden bzgl. der anderen beiden Aussagen q und r, so haben wir zwar bereits einen Zipfel der Wahrheit erhascht, aber eben noch nicht die ganze Wahrheit. Unser erstes Ziel mag dabei schon erreicht sein, aber das zweite ist noch nicht vollständig umgesetzt. Das können wir auch gleich konkret ausrechnen.

Schon Popper hatte in *Conjectures and Refutations* (1963) die Idee, dass wir uns der Wahrheit langsam annähern, und dass diese Annäherung eng verknüpft ist mit dem Gehalt unserer Theorien. Fortschritt in der Wissenschaft besteht dann u.a. darin, dass wir *gehaltvolle Theorien* entwickeln, um uns der *vollständigen* Wahrheit anzunähern. (Zusätzlich zu diesem Ziel der Wahrheitsannäherung wird allerdings für wissenschaftlichen Fortschritt dazugehören, dass wir nicht nur zufälligerweise wahre Überzeugungen aufweisen, sondern dass diese vor allem gut begründet sind.) Popper hatte bereits erste Ideen für eine weitergehende Explikation von *Wahrheitsnähe* entwickelt, die erwiesen sich aber leider als unbrauchbar (s. etwa Oddie 2013 und 2014). Moderne verbesserte Versionen dieser Ideen finden sich z.B. bei Schurz und Weingartner (2010).

Einen verwandten und sehr intuitiven Ansatz bieten die Vorschläge, von *Wahrheitsähnlichkeit* zu sprechen und sich dabei auf den Abstand unserer möglichen Welten zu beziehen. Für unser kleines Beispiel liegt ein einfaches Maß für den Abstand einer Welt w* (= p&¬q&¬r) von unserer aktualen Welt w (= p&q&r) nahe: Zähle einfach die Anzahl der voneinander abweichenden Aussagen zwischen den Vollkonjunktionen. Also hier d(w,w*) = 2. Mit diesem sogenannten »city-block« Abstandsmaß haben wir intuitives Maß für den Abstand einer möglichen Welt von der vollständigen Wahrheit erhalten.

Die Probleme beginnen aber, sobald wir dieses Maß auf andere Aussagen unserer kleinen Welt ausdehnen möchten. Schon Popper war etwa der Ansicht, dass schwächere wahre Aussagen wie p einen größeren Abstand von der Wahrheit haben als stärkere wahre Aussagen wie p&q. Den *Gehalt* von Aussagen versucht man nun dadurch zu berücksichtigen, dass wir vom *Bereich* (»range«) einer Aussage sprechen, als der Menge der möglichen Welten, in denen sie wahr ist. Dabei wird eine Aussage wie z.B. p&¬q durch eine Menge möglicher Welten, in denen sie wahr ist (hier: {p&¬q&r, p&¬q&¬r}), repräsentiert. Einige Ansätze wie Spohn (2012) starten direkt damit, dass Aussagen als Mengen solcher möglichen Welten oder möglichen Situationen aufgefasst werden.

Jedenfalls stellt sich die Frage, wie groß der Abstand einer solchen *Menge* von möglichen Welten von der aktualen Welt ist. Dazu wurden unterschiedliche Vorschläge diskutiert, die alle gewisse Vorteile, aber

auch Schwächen aufweisen (vgl. Oddie 2013, 2014). So dachte man u.a. daran, für den Abstand einer Aussage A und unserer aktualen Welt w, sowohl die nächstgelegene Welt in range(A) zu w und die entfernteste Welt in range(A) zu w heranzuziehen und daraus den Abstand zu mitteln. Dieser komplexen Debatte möchte ich hier nicht lange nachgehen, aber sie hat zu bestimmten Vorstellungen für Wahrheitsnähe geführt, von denen einige besonders grundlegende Intuitionen von Schurz (2014, 339) genannt werden.

(1) Unter den wahren Theorien nimmt die Wahrheitsähnlichkeit mit der logischen Stärke der Theorien (bzw. ihrem Gehalt) zu (<):
p ∨ q < p < p & q < p & q & r
(2) Für (teilweise) falsche Theorien gilt: ¬p & ¬q < ¬p < p & ¬q

Ein allgemeines Maß für die Wahrheitsähnlichkeit (W), für das Oddie (2013) ausführlich anhand plausibler allgemeiner Prinzipien argumentiert, ist das *Durchschnittsmaß*. Demnach ist der Abstand d(A,w) unserer aktualen Welt w von einer Aussage A gleich dem Durchschnitt der Abstände der Welten im Bereich von A (range(A)) zu unserer Welt: Wahrheitsähnlichkeit(A) = W(A) = $1/n \sum_{w* \in range(A)} d(w,w*)$, wobei n die Anzahl der möglichen Welten in range(A) darstellt. Für unser kleines Beispiel ergibt sich damit (s. Oddie 2014):

A	Wahrheit	W(A)
p&q&r	wahr	0
p&q	wahr	0.5
p&q&¬r	falsch	1.0
p	wahr	1.3
p&¬q	falsch	1.5
¬p	falsch	1.7
¬p&¬q&r	falsch	2.0
¬p&¬q	falsch	2.5
¬p&¬q&¬r	falsch	3.0

Tabelle 2.1: Der Wahrheitsabstand W(A) unterschiedlicher Aussagen

Dabei zeigt sich – was wir bereits angenommen haben–, dass manchmal sogar falsche Theorien näher an der Wahrheit liegen können als wahre,

wenn sie sehr gehaltvoll sind. Das belegt noch einmal, dass wir eine Verrechnung unserer beiden Ziele vornehmen müssen. Eine informativere falsche Theorie kann sogar wahrheitsnäher und damit fortschrittlicher sein als eine weniger informative wahre Theorie. Daher kann es im Sinne des Ziels größerer Wahrheitsnähe geboten sein, der riskanteren bzw. weniger gut begründeten, aber gehaltvolleren Theorie den Vorzug vor einer besser begründeten, aber weniger gehaltvollen Theorie zu geben. Popper betonte sogar oft das Ziel des größeren Gehalts der Theorien aufgrund seines speziellen Ansatzes über Gebühr. Das werden wir später noch sehen.

Insbesondere dürfen wir uns nicht nur an der Wahrscheinlichkeit der Theorien oder unserer Überzeugungen orientieren. Laut Wahrscheinlichkeitsrechnung gilt: $P(p) \geq P(p\&q) \geq P(p\&q\&r)$. Solange wir die Wahrheit der drei Aussagen noch nicht sicher kennen und wir es mit unabhängigen empirischen Aussagen zu tun haben, gilt für ein reguläres Wahrscheinlichkeitsmaß sogar: $P(p) > P(p\&q) > P(p\&q\&r)$. Das belegt wiederum, wie Wahrscheinlichkeit und Wahrheitsnähe in Konflikt geraten können. Wir suchen in diesem Fall nach der ganzen Wahrheit, auch wenn das gerade die unwahrscheinlichste der drei Theorien ist.

Die Idee, dass falsche Theorien sogar wahrheitsnäher als wahre Theorien sein können, wird vielleicht noch einleuchtender, wenn wir an eine Modellwelt denken, die durch die 10 wahren Aussagen $a_1,\ldots,a_{10}$ korrekt beschrieben wird. Vergleichen wir nun die wahre Theorie $T_1=a_1$ mit der falschen Theorie $T_2=a_1\&\ldots\&a_9\&\neg a_{10}$. Da die zweite Theorie einen wesentlich größeren Wahrheitsgehalt als die erste aufweist, spielt der kleine Fehler in der zweiten Theorie keine so große Rolle mehr. Es ist daher durchaus naheliegend, die zweite Theorie als wahrheitsnäher anzusehen als die erste.

Das Problem, wie die Verrechnung der zwei Ziele genau vorgenommen werden soll, ist natürlich nicht endgültig gelöst. Unsere Vorstellungen von Fortschritt sind an dieser Stelle nicht so klar vorgegeben, dass sich eine unbestreitbare Verrechnungsmethode anbieten würde. Jedenfalls spielt für die Erreichung des übergeordneten epistemischen Ziels der Ermittlung der *vollständigen Wahrheit* der Gehalt unserer Theorien eine wesentliche Rolle, was oft vergessen wird, wenn wir nur darauf

abzielen, dass unsere Theorien gut begründet bzw. sehr plausibel oder sehr wahrscheinlich sein sollen.

Man kann nun noch einen Schritt weitergehen und fragen, was wir neben einer genaueren Klärung der Ziele aus der Innenperspektive des epistemischen Subjekts für unser erkenntnistheoretisches Projekt gewinnen. Da wir nicht wissen, welches die wahre Welt (zu einem bestimmten Thema) ist, können wir natürlich den Abstand von unseren Theorien A zur Wahrheit nicht direkt angeben. Um das Konzept der Wahrheitsnähe dennoch aktiv zum Einsatz bringen zu können, benötigen wir zumindest noch eine Wahrscheinlichkeitsverteilung P(w) auf den möglichen Welten, wie sie etwa der Bayesianismus vorgibt. Dann lässt sich die zu *erwartende Wahrheitsnähe* für unsere Theorie A bestimmen als Durchschnittsabstand von A zu w über alle möglichen Welten w gemittelt, wobei d(A,w) den Abstand von A und irgendwelchen Welten w einbringen soll, etwa im Sinne unseres obigen Durchschnittsmaßes von Wahrheitsnähe:

$$EW(A) = \sum_w P(w) \cdot d(A,w)$$

Dabei handelt es sich eigentlich um den zu *erwartenden Wahrheitsabstand*, denn mit größer werdenden Zahlen entfernen wir uns von der Wahrheit. Unser erkenntnistheoretisches Ziel könnten wir dann so formulieren, dass wir die Theorie wählen sollten, für die EW(A) möglichst minimal wird. Dabei wird wiederum deutlich, dass als Entscheidungsmerkmal nicht nur eine möglichst hohe Wahrscheinlichkeit von A zählt, sondern ebenfalls der Gehalt der Theorien zu berücksichtigen ist. So ist der Abstand einer Tautologie von allen Welten recht groß, und sie stellt im Normalfall daher keine Lösung unseres Minimierungsproblems dar, obwohl ihre Wahrscheinlichkeit immer 1 ist.

In einer ähnlichen Richtung wurde auch eine epistemische Entscheidungstheorie entwickelt worden, die mit *epistemischen Nutzenwerten* arbeitet. Der Nutzen setzt sich darin zusammen aus der Wahrscheinlichkeit der Hypothesen, aber auch ihrem Gehalt, um den bisherigen Überlegungen Rechnung zu tragen. Zu wählen ist dann die Theorie mit dem höchsten epistemischen Nutzen und nicht die wahrscheinlichste Theorie und das muss natürlich keinesfalls übereinstimmen (vgl. Oddie

2014). Viele Ansätze, die wir im Folgenden diskutieren werden, berücksichtigen diesen zweiten Aspekt jedoch nicht.

Meine These ist, dass der am besten im Rahmen des *Schlusses auf die beste Erklärung* zum Tragen kommt, den wir in Kapitel 4 genauer untersuchen werden. Bei diesem Schlussverfahren suchen wir gezielt nach Theorien, die grundlegende Kausalzusammenhänge und Naturgesetze aufdecken und somit viele Phänomene erklären können. Die erklärten Beobachtungen sind aus der Theorie ableitbar bzw. werden von der Theorie vorhergesagt und stützen daher die Theorie induktiv, die andererseits so gestaltet sein muss, dass sie möglichst gehaltvolle Erklärungen dieser Phänomene liefert. Damit werden beide Aspekte berücksichtigt. Diese Idee wird uns im Folgenden immer wieder begegnen.

Die genannte Gefahr einer zu Empirie nahen Forschung ist auch durchaus real. In den empiristisch gesinnten Bereichen der Sozialwissenschaften findet man immer wieder die Tendenz, nur relativ beobachtungsnahe Hypothesen aufzustellen und diese dann durch einfache Signifikanztests direkt abzusichern, und das Ganze als große Fortschritte in der Forschung zu verkaufen. So möchte man allzu Spekulatives von vornherein vermeiden. Dabei wird vergessen, dass es gerade in der physikalischen Forschung immer auch eine *theoretische Physik* mit vielen recht spekulativen Theorien gab, die erst zu den wirklichen großen Fortschritten in der Physik geführt hat. Wenn wir uns die Physik also zum Vorbild für erfolgreiche empirische Forschung nehmen, dann sollten wir den strikt empiristischen Weg überdenken.

Außerdem gibt es auch keinen einfachen Königsweg von Daten zu dadurch gestützten Theorien. Der Signifikanztest scheint geradezu dafür gemacht zu sein, simple empirische Hypothesen zu stützen. Wissenschaftstheoretiker und Wissenschaftler aus den entsprechenden Disziplinen sind aber zu Recht skeptisch gegenüber diesem Verfahren. Damit lassen sich zwar leicht – und tatsächlich zu leicht (vgl. Kapitel 6) – einfache Hypothesen als »signifikant« erweisen, doch es bleibt unklar, was das genau bedeutet. Wir können nämlich leider überhaupt nichts darüber sagen, wie stark signifikant gestützte Theorien nun bestätigt sind.

Tatsächlich gibt es keinen einfachen Algorithmus, der für uns darüber entscheidet, welche Theorien besonders gut durch die Daten gestützt werden. Hier sind vielmehr komplexere Abwägungen erforderlich, für die wir im Folgenden aber durchaus hilfreiche Anleitungen finden werden. Wenn es keinen einfachen Weg gibt, bedeutet das natürlich nicht, dass wir in der feyerabendschen Anarchie eines »anything goes« landen.

Nur dem Wunsch der Wissenschaftler nach einem quasi algorithmischen Verfahren, das wir möglichst dem Computer überlassen können, zur Entscheidung, wie gut unsere Theorien durch unsere Daten gestützt werden, dem können wir leider nicht nachkommen. Beim Schluss auf die beste Erklärung wird sich das schon zeigen, denn es sind immer wieder Abwägungen erforderlich, wie gut unsere beteiligten Hilfshypothesen sind und wie gut die resultierenden Erklärungen im Vergleich aussehen. Dabei tritt wiederum das genannte Problem auf, dass gerade die gehaltvolleren und damit weniger wahrscheinlichen Theorien die besseren Erklärungen liefern können. Das verlangt nach einer Abwägung, welche epistemischen Risiken wir für bestimmte Erklärungsgewinne eingehen sollen. Bayesianer versuchen zumindest die Abwägung des epistemischen Risikos in Zahlen zu gießen – die epistemischen Wahrscheinlichkeiten –, was ihnen aber von klassischen Statistikern eher als unseriöse subjektive Elemente ihres Verfahrens ausgelegt wird. Wir werden jedoch sehen, dass sich die klassischen Statistiker etwas zu früh freuen.

Der Drang nach gehaltvollen, tiefen Theorien ruft natürlich auch die andere Gefahr auf den Plan, *zu spekulativ* vorzugehen. In der Psychologie erschien vermutlich die z.T. recht phantasievolle Hypothesenbildung im Rahmen der freudschen Psychoanalyse vielen Psychologen als zu spekulativ und hat in Form einer Gegenreaktion zu einer stärker empiristisch orientierten Disziplin beigetragen. In der Physik bemängeln die Kritiker, dass die Ausarbeitung der so beliebten *Stringtheorien* zu sehr im theoretischen Raum stattfindet und keine oder nur wenige Anbindungen an die Empirie aufweist. Auch die klassischen Ökonomen werden gern dafür kritisiert, dass sie immer komplexere und subtilere Anwendungen der *Spieltheorie* entwickeln, aber nicht so viel Wert darauf legen, herauszufinden, ob Menschen sich auf den Märkten tatsächlich als rational im Sinne der Spieltheorie verhalten. Um dieser Gefahr zu

begegnen, hat sich dann die sogenannte »bevavioral decision theory« gebildet, die gerade versucht, die Spieltheorie zu verbessern im Sinne einer empirisch angebundenen Hypothesenbildung. Die Wissenschaftler dieser Disziplinen müssen den geeigneten Mittelweg jeweils für ihre Fächer finden. Einen einfachen Ratschlag dazu kann ich nicht geben.

Nichtsdestotrotz müssen wir nun erst einmal genauer bestimmen, wie gut unsere Begründungen – also der erste Aspekt der Theorienwahl – denn sein sollen, damit wir von *wissenschaftlichem Wissen* sprechen können. Das soll im nächsten Abschnitt weiter verfolgt werden.

2.3 Wissen

2.3.1 Platons Wissenskonzeption

Die Frage, welche Kriterien Wissen auszeichnen, stellen sich Philosophen seit der Antike. Lange Zeit galt eine Antwort Platons als völlig ausreichend, wonach wir unter Wissen eine Überzeugung verstehen, die sowohl wahr wie auch gut begründet ist. Demnach sind nur drei Anforderungen zu stellen, wenn wir sagen wollen, dass eine Person S über das Wissen verfügt, dass p der Fall ist, wobei »p« eine beliebige Aussage vertritt.

(PW) ***S weiß, dass p*** gdw. (genau dann, wenn gilt:)
(1) S ist von p überzeugt.
(2) p ist wahr.
(3) S verfügt über gute Gründe für p.

Man kann sich leicht davon überzeugen, dass diese drei Bedingungen tatsächlich notwendig für Wissen sind. Ist eine von ihnen verletzt, sprechen wir normalerweise nicht mehr davon, dass S über das Wissen verfügt, dass p. Lange Zeit glaubte man, sie seien auch ausreichend für Wissen, doch dann entwickelte Edmund Gettier (1963) erste Gegenbeispiele, die belegten, dass wir unsere Wissenskonzeption verstärken müssen. Ich möchte hier nicht sehr tief in die von Gettier ausgelöste Debatte einsteigen, aber doch einige für unser Projekt wichtige Aspekte davon darstellen.

Gettiers Gegenbeispiele folgen alle einem gewissen Schema: Zwar sind in seinen Beispielen die drei Bedingungen der platonischen Wissenskonzeption erfüllt, aber es ist genau genommen nur ein *glücklicher Zufall*, dass die vorliegenden Gründe für p auch mit einem wahren p zusammentreffen. Die Gründe haben eigentlich mit dem Vorliegen von p in unserem konkreten Fall nicht viel zu tun, sondern wir haben es eher mit einem Fall von *epistemischem Glück* zu tun. In vielen ähnlich gelagerten Fällen würden wir sogar erwarten, dass p falsch wäre, obwohl entsprechende Gründe für p vorlägen.

So könnte etwa Folgendes der Fall sein: Einer meiner Studenten X fährt regelmäßig mit einem Audi A3 vor und erzählt mir, dass er sich den von seinem Ersparten gekauft hätte. Da ich keinen Grund habe, ihm zu misstrauen, glaube ich seinen Ausführungen, und die sollten daher durchaus einen guten Grund für meine Überzeugung abgeben, dass gilt: p ≡ *Dem Student X gehört ein Audi A3.* Damit sind schon zwei Bedingungen von (PW) erfüllt.

Allerdings sei X nun doch ein Aufschneider, obwohl ich keine Anhaltspunkte dafür hatte, und der Audi gehört ihm überhaupt nicht. Er hat ihn sich immer nur von seiner Mutter ausgeliehen. Damit wäre die dritte Bedingung von (PW) nicht erfüllt und es wäre auch kein Fall von Wissen. Man könnte sagen, dass das dann der Normalfall ist. Wenn ich auf einen Aufschneider hereinfalle und ihm glaube, so gelange ich normalerweise zu falschen Überzeugungen.

Aber da dem Vater (und natürlich ebenso der Mutter) von X das Spiel mit dem ständigen Ausleihen nun doch zu viel wurde, hat er inzwischen einen neuen Audi A3 auf den Namen seines Sohnes gekauft, ohne dem das schon mitzuteilen. Damit wäre durch diesen *glücklichen Zufall* nun doch die Wissensbedingung (2) aus (PW) erfüllt, denn p ist jetzt also tatsächlich wahr. Trotzdem würde man mir sicher nicht zubilligen, ich hätte es bereits gewusst, denn ich hatte schließlich von den Handlungen des Vaters keine Ahnung. Meine Rechtfertigung für p beruhte nur auf *irreführenden Informationen* und *Lügen* und hatte leider keinen Bezug zu der Wahrheit von p. Insbesondere hätte ich X genauso geglaubt und wäre damit dem Irrtum p aufgesessen, wenn sein Vater nicht so großzügig gewesen wäre.

Ähnlich sind viele Beispiele vom Gettiertyp aufgebaut, die alle zeigen, dass die Bedingungen (PW) nicht genügen, damit Wissen vorliegt. Bekannt ist u.a. das Beispiel der *Scheunenattrappen*. Nehmen wir an, in der Schweiz wären inzwischen alle Scheunen nur noch Attrappen für den Tourismus. Es stehe nur noch die Vorderfront. Franz fährt durch die Schweiz und staunt an jeder dieser Attrappen: »Was für eine romantische Scheune«. Einmal hat er jedoch Glück und sieht per Zufall tatsächlich noch eine letzte verbliebene echte Scheune. Hat er in diesem Fall das Wissen, dass eine Scheune vor ihm steht? Wohl kaum, denn es handelt sich nur um einen *Glückstreffer* unter lauter Nieten, die sich für ihn genauso darstellen. Er hätte genauso bei einer der vielen Scheunenattrappen die falsche Überzeugung entwickelt, dass es sich um eine Scheune handelt. Normalerweise führt ihn sein Verfahren der Überzeugungsbildung in den geschilderten Situationen nur zu falschen Annahmen. Nur seine eine Überzeugung war durch einen besonderen Zufall wahr.

Aber nichtsdestoweniger sind alle drei PW-Bedingungen erfüllt. Er glaubt daran, sich einer Scheune gegenüberzusehen, und hat mit dem Sehen der Scheune schließlich einen guten Grund dafür. Tatsächlich steht er in diesem einen Fall ausnahmsweise und zufälligerweise einer echten Scheune gegenüber; seine Überzeugung ist also wahr. Für Wissen ist das trotzdem zu wenig, denn die Wahrheit beruht nur auf Glück. Daher verlangen moderne Wissenskonzeptionen meist mehr von einer Begründung, die zu Wissen führen soll. Sie muss nicht nur im Allgemeinen gut sein, sondern auch im Einzelfall nicht nur zufällig richtig liegen.

2.3.2 Modale Anforderungen

Die Frage in der Wissensdebatte ist nun, welche weiteren Anforderungen wir an echtes Wissen stellen sollten. In der letzten Zeit sind einige modale Anforderungen genannt worden, die ich zunächst kurz vorstellen möchte, da sie einige unserer Intuitionen zum Wissensbegriff gut einfangen. Nozick (1981) und andere setzen darauf, dass unsere Überzeugungen *sensitiv* gegenüber der Wahrheit bzw. Falschheit von p sind bzw. auf eine dafür *sensitive Weise* gebildet wurden, d.h. auf einer *sensitiven Methode*

beruhen. Unsere Überzeugungen *verfolgen somit die Wahrheit* oder *sind der Wahrheit auf der Spur* (»track the truth«) und sind nicht nur rein zufällig wahr. Dabei wird etwa als vierte Wissensbedingung verlangt:

(4) **Sensitivität**: Wenn p falsch gewesen wäre, hätte S p nicht geglaubt.

Wäre die Welt also ein wenig anders gewesen und p eben nicht wahr gewesen, so hätte S sogleich entsprechend darauf reagiert und in dieser Situation p nicht mehr geglaubt. Das wird oft so formuliert, dass in allen zu unserer Welt am *nächsten gelegenen möglichen Welten* (bzw. in den nächstgelegenen Situationen), in denen p nicht mehr gilt, S auch p nicht mehr glaubt. Unsere Überzeugungsbildung reagiert hier sensibel auf die Falschheit der Überzeugung. Unter normalen Bedingungen (mit normalen Scheunen) erfüllt das Scheunensehen die Sensitivitätsanforderung. Die nächsten Welten, in denen keine Scheune vor uns steht, sind solche, in denen keine Scheune oder Scheunenattrappe zu sehen ist. Also glaubt S in diesen Welten auch nicht, dass dort eine Scheune vor ihm steht. Doch diese Bedingung wird in dem Land mit lauter Scheunenattrappen eben gerade nicht erfüllt, denn dort ist unter den nächsten Welten, in denen keine Scheune vor S steht, eine Welt, in der eine Scheunenattrappe vor S steht, und S daher trotzdem glaubt, dass eine Scheune vor ihm steht. Also ist hier die Bedingung (4) nicht erfüllt und gestattet es daher, p als kein Wissen auszuweisen.

Allerdings hat die Bedingung (4) gewisse bekannte Defizite. So bleibt die Sensitivität nicht erhalten unter deduktiven Schlussfolgerungen, und Wissen wäre dann nicht mehr deduktiv abgeschlossen. Doch die deduktive Abgeschlossenheit wirkt wie eine wesentliche Anforderung an eine wissenschaftliche Konzeption von Wissen. Schließlich möchten wir aus unserem Wissen weitere Schlussfolgerungen ziehen dürfen und so zu weiterem Wissen gelangen.

Andererseits haben wir aber noch weitere Probleme mit der Abgeschlossenheitsforderung. Es sei mit $W_S(p)$ ausgedrückt, dass S weiß, dass p, und mit G wird der entsprechende Begriff des Glaubens dargestellt, dann besagt die Forderung:

(DAW) Deduktive Abgeschlossenheit von Wissen
$W_S(p)$ & $W_S(p \Rightarrow q)$ & $G_S(q)$, dann gilt: $W_S(q)$

Die platonischen Wissensbedingungen werden hierbei von q offensichtlich erfüllt, denn q muss wahr sein, wird von S geglaubt und natürlich haben wir mit der Folgerungsbeziehung auch eine Begründung für q, wenn wir über Begründungen für die Aussagen p und p⇒q verfügen. Aber Probleme bereitet uns an dieser Stelle wieder einmal der Skeptiker. Er nutzt (DAW) sofort für seine Zwecke. Wenn p ein ganz normales Wissen über meine Außenwelt darstellt, wie dass (p) *ich gerade auf einem Stuhl sitze*, wird er sagen, dass mit (DAW) dann auch folgt, dass (q) *ich kein Gehirn im Topf* im Sinne von Putnam bin.

Ein solches Gehirn im Topf (GiT) würde – ähnlich wie im Film *Matrix* – direkt durch einen Computer stimuliert, der ihm eine virtuelle Welt vorspielt. In Putnams Beispiel wird das Gehirn allerdings noch dem Körper entnommen und schwimmt in einer Nährlösung. Keinesfalls würde ich als GiT also auf einem Stuhl sitzen. Könnte ich den Skeptiker demgemäß mit ganz normalem Alltagswissen und (DAW) aus dem Felde schlagen? Ich weiß, dass ich auf einem Stuhl sitze, also weiß ich, dass ich kein Gehirn im Topf bin.

Das lässt sich der Skeptiker so nicht gefallen und kann immerhin darauf verweisen, dass wir gegen die skeptischen Hypothesen überhaupt keine speziellen Einwände vorbringen können, und unsere Überzeugungen gegenüber diesen Möglichkeiten demnach nicht wirklich *sensitiv* sind. Wären wir nämlich Gehirne im Topf, würden wir trotzdem weiterhin glauben, auf einem Stuhl zu sitzen. Die Forderung (DAW) gerät also nicht nur in Konflikt mit der Forderung der Sensitivität für Wissen, sondern ist daneben noch aus anderen Gründen problematisch. Trotzdem bleibt hier der Verdacht zurück, dass die Sensitivitätsbedingung schon zu »viel« leistet und uns zu schnell Wissen verschaffen könnte, obwohl das dann nicht stabil unter deduktiven Schlussfolgerungen ist.

Es gibt allerdings ebenfalls recht alltägliche Situationen, in denen die Sensitivitätsbedingung versagt, weil sie *zu sensibel* reagiert und damit zu viel fordert. Sosa (1999) schildert das folgende Beispiel: Wenn wir unseren Müll durch einen Müllschlucker in den Keller schicken, können wir kurz danach sehr wohl das Wissen haben, (p) *dass unser Müll nun im Keller angekommen ist.* Doch diese Überzeugung reagiert nicht wirklich sensibel auf mögliche Störungen im Müllschacht, die zu einem Feststecken des Mülls im Schacht führen würden. Wir hätten dieselbe

Überzeugung über unseren Müll im Keller auch gebildet, wenn er sich im Schacht verkeilt hätte. Nehmen wir die Sensitivitätsbedingung also ernst, hätten wir sogar im Normalfall des Mülls im Keller kein Wissen mehr, dass p. Damit würde dann aber unser ganz normales Wissen entwertet.

Deshalb plädieren Sosa (1999) und etwa auch Pritchard (2005, 2007, 2008) anstatt der *Sensitivität* für die Forderung der *Sicherheit*, um damit epistemisches Glück als Quelle von Wissen auszuschließen. Unsere Überzeugung p ist nicht nur glücklicherweise wahr, wenn p auch in den *meisten nächsten möglichen Welten*, in denen S weiterhin p glaubt (Sosa 1999), bzw. p auf diese Weise gebildet wird (Pritchard 2007), auch tatsächlich wahr ist. Das ist eine Forderung nach einer Art von *praktischer Sicherheit* unserer Überzeugungen, um Wissen zu erhalten.

(4*) **Sicherheit**: In den *meisten* nächstgelegenen Welten, in denen S p glaubt, ist p wahr.

Im Prinzip geht es darum, dass zumindest in den meisten Situationen, die aus Sicht von S relevante Alternativen darstellen, in denen wir an p festhalten würden, p tatsächlich wahr ist. Damit wird unser Müllbeispiel korrekt erklärt, denn in den meisten benachbarten Welten, in denen wir glauben, dass der Müll unten angekommen ist, ist er das auch, also handelt es sich im Normalfall damit um Wissen. Jedenfalls dann, wenn der Müllschlucker in den meisten Fällen korrekt funktioniert und nicht dauernd verstopft ist. In dem Fall hätten wir natürlich tatsächlich im Normalfall nicht mehr das Wissen, dass der Müll im Keller gelandet ist.

Allerdings gibt es wiederum Problemfälle. Wenn ich ein Los von einer Million Losen kaufe, von denen nur eines gewinnt, so spricht man mir normalerweise trotzdem kein Wissen zu, dass mein Los verliert. Andererseits scheint es sich um eine praktisch *sichere* Überzeugung zu handeln, denn in den meisten nahe gelegenen Welten (bzw. in den nächstgelegenen relevanten Situationen) hätte ich ebenfalls eine Niete gezogen. Damit wäre die Sicherheitsbedingung erfüllt, es läge aber trotzdem kein Wissen vor. Freilich müssen wir in diesem Beispiel noch einmal unsere Intuition befragen, ob wir hier nicht doch lieber von Wissen sprechen sollten und unsere neue Wissenskonzeption nicht doch richtig liegt. Allerdings weiß Ram Neta (2009) andere noch phantasievollere Gegenbeispiele zu nennen, die ohne Lotterien auskommen.

Pritchard (2007) verstärkte daher seine Konzeption von Sicherheit noch, indem für besonders eng benachbarte Welten gefordert wird, dass dort p in jedem Fall wahr ist:

> **Verstärkte Sicherheit**: S's belief is safe iff in most near-by possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world, and in all very close near-by possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world, the belief continues to be true.

Die modalen Zusatzforderungen scheinen uns also intuitiv schon den richtigen Weg zu weisen, sind aber nur schwer nachprüfbar und nicht immer leicht verständlich. Sie beruhen vor allem auf genauen Intuitionen darüber, wie eng benachbart bestimmte mögliche Welten sind. So plausibel der Ansatz also auch startet, so schwierig wird es, wenn wir versuchen, ihn weiter zu präzisieren. Deshalb scheint es mir für die Suche nach wissenschaftlichem Wissen hilfreich zu sein, einen alten Ansatz wiederzubeleben und nach expliziten nicht-modalen Bedingungen zu suchen, die uns für die Wissenschaften helfen, besser zu verstehen, wonach wir eigentlich suchen, wenn wir Wissen erzeugen möchten. Wissen sollte vor allem über stabile und letztlich unanfechtbare Begründungen verfügen. Das hat insbesondere damit zu tun, wie stabil Wissen unter Hinzunahme bestimmter Wahrheiten bleibt. Diese Idee findet sich schon in den Arbeiten von Keith Lehrer (1974).

2.3.3 Unanfechtbares Wissen

Wenn q ein guter Grund für p ist, so schreiben wir das (nur) in diesem Kapitel kurz: »q⇒p«. (Sonst meine ich damit, dass p aus q deduktiv folgt.) Hier sind mit dem Pfeil alle Arten von induktiven und deduktiven Rechtfertigungen gemeint, die wir intuitiv als gute Begründungen betrachten. Außerdem gehen wir davon aus, dass es sich bei q und p um Aussagen handelt, während »⇒« die (meist induktive) stützende Verbindung zwischen den beiden Aussagen bezeichnen soll. Dabei ist das so gemeint, dass q nicht unbedingt die kompletten Prämissen unserer Begründung angeben muss, denn es handelt sich um eine bloß induktive Begründung

und wir denken uns diese Beziehung eigentlich als dreistellig, wonach weiteres Hintergrundwissen in die Stützungsbeziehung eingehen darf, nennen explizit aber nur die hervorgehobene Aussage q, die allerdings bereits eine Konjunktion aus mehreren Aussagen darstellen kann.

So ähnlich sprechen wir schon umgangssprachlich oft über epistemische Rechtfertigungen. Johns verkniffener Gesichtsausdruck, als er zu James schaute, ist unser Grund dafür anzunehmen, dass er ärgerlich auf James ist. Dann steckt in dem »⇒« u.a. das *Hintergrundwissen*, dass zum Ärger über einen Menschen häufig ein bestimmter Gesichtsausdruck gehört. Typischerweise werden bei induktiven Schlüssen nicht alle Annahmen explizit aufgeführt. Deshalb sprechen wir hier normalerweise davon, dass einzelne zentrale Aussagen q (wesentliche Prämissen des induktiven Schlusses) den Grund darstellen, an etwas Bestimmtes zu glauben. In unserem Beispiel müsste q gerade die Aussage sein, *dass John einen verkniffenen Gesichtsausdruck annahm, als er zu James schaute*. Nicht das oben genannte Ereignis selbst ist also rechtfertigend, sondern die Beschreibung des Ereignisses durch eine entsprechende Aussage q. Um in manchen Formulierungen aber nicht zu umständlich zu werden, werde ich nicht immer sprachlich ganz präzise sein, weil eigentlich unmissverständlich ist, was gemeint ist.

Manchmal wird jedoch schon dieser Ansatz bestritten. Es wird etwa eingewandt, dass Begründungen sich ebenso in anderen als doxastischen Zuständen also etwa in bestimmten nicht-propositionalen mentalen Zuständen finden ließen. Doch es gibt auch gute Gründe dafür, an Aussagen als den Trägern der Stützung von anderen Aussagen festzuhalten, wofür u.a. Williamson (2000) argumentiert. In dem hier behandelten Kontext wird das noch deutlicher. Erstens geht es mir nicht in erster Linie darum, zu explizieren, wann eine Person S in ihren Überzeugungen gerechtfertigt ist, sondern zunächst mehr darum, wann eine Aussage eine andere im Lichte weiterer Aussagen (etwa des sogenannten Hintergrundwissens) epistemisch stützt. Das können wir in eine einzelne Person verlegen, müssen das aber nicht unbedingt tun. Man kann z.B. stattdessen ebenso Zusammenhänge betrachten, die von einer ganzen Wissenschaftlergemeinschaft akzeptiert werden.

Zweitens wird es mir insbesondere um *wissenschaftliches Wissen* gehen und dafür gelten allemal besondere Spielregeln. Alle relevanten

Begründungen müssen in diesem Fall *explizit* auf den Tisch gelegt werden und müssen damit in die Form von Aussagen gebracht werden. Nicht-verbalisierte Sinneseindrücke werden erst dann gezählt, wenn jemand sie in propositionale Form bringt und so *schildert*, was er genau gesehen oder auf andere Weise wahrgenommen hat. Deshalb gehe ich davon aus, dass die Gründe für eine Überzeugung immer in Form von Aussagen oder im Falle einer bestimmten Person in Form ihrer anderen Überzeugungen vorliegen (vgl. a. Bartelborth 2015). Außerdem ist die Position *internalistisch* gemeint, wonach unsere epistemischen Subjekte vollen kognitiven Zugriff auf die in Frage stehenden Aussagen haben, gleichgültig, ob es sich um tatsächliche Personen oder eine Wissenschaftlergemeinschaft handelt. Das ist wiederum für wissenschaftliche Begründungen offensichtlich unabdingbar.

Genauer diskutieren möchte ich statt der genannten Grundlagenfragen die komplizierte Dialektik von Gründen und Gegengründen bzw. Einwänden in der Generierung von Wissen. Um die besser beschreiben zu können, unterscheiden wir zunächst (in Anlehnung an John Pollock 1974) zwischen *Unterminierern* (»undercutting defeater«) und *Gegengründen* (»rebutting defeater«). Beide können sich als Hindernisse auf dem Weg zum Wissen, dass p der Fall ist, herausstellen (vgl. a. Grundmann 2008).

Ein Unterminierer findet sich in Aussagen u, die eine Begründung q für eine Aussage p untergraben. Zunächst gilt also q⇒p. Doch dann, wenn ich zu q den Unterminierer u hinzunehme, stellt u & q keine Begründung für p mehr dar: ¬((u & q) ⇒ p). (Hierbei bezeichnet das »¬«-Zeichen die Negation der Aussage.) Das spricht natürlich noch nicht dafür, dass p falsch ist. Nur *eine* Begründung für p ist weggebrochen. Dabei kann sich der Unterminierer u direkt gegen q wenden und so q entwerten, oder er richtet sich eher gegen den Pfeil, wonach q nicht mehr speziell für p spricht, sondern etwa als mehrdeutig ausgewiesen wird und damit seine Bestätigungswirkung für p verliert. Im ersten Fall stellt u schlicht einen Gegengrund gegen q dar, für die wir gleich Beispiele sehen werden, d.h., es gilt hier: u⇒ ¬q. Für den zweiten Fall werden wir danach Beispiele aus unserer Wahrnehmung kennenlernen. In beiden Fällen können wir einen Unterminierer dadurch kennzeichnen, dass er gegen q⇒p spricht, d.h.: u⇒ ¬(q⇒p).

(U) u ist ein ***Unterminierer*** von q⇒p gdw gilt: q⇒p und ¬((u & q)⇒p) (und damit gilt: u⇒¬(q⇒p))

(WU) u ist ein ***wahrer Unterminierer*** von q bzw. von q⇒p, wenn u ein Unterminierer von q ist und u außerdem wahr ist.

Unterminierer untergraben also zunächst eine Begründungsbeziehung, wobei wir erst einmal offenlassen, ob sie tatsächlich wahr sind, oder wir sie nur als wahr akzeptieren.

Einen anderen Einwand gegen ein mögliches Wissen, dass p, finden wir neben den Unterminierern noch in den *Gegengründen* g zu p. Das sind einfach Gründe, die nun tatsächlich für non-p sprechen und sie können von q ganz unabhängig sein: g⇒¬p. Wir werden sehen, dass wir von Wissen typischerweise verlangen, dass es keine (auch keine unbekannten) *relevanten Unterminierer* und nur bestimmte *Gegengründe* geben darf. Doch zunächst möchte ich das an einfachen Beispielen erläutern. In unserem Audi-Beispiel wäre etwa ein Unterminierer das Wissen darum, dass X ein großer Aufschneider ist. Dieses Wissen würde zumindest meinen einen Grund dafür zerstören, dass X tatsächlich der Eigentümer eines A3 ist. Aber selbst ein Aufschneider kann natürlich einen A3 besitzen, das wird durch die neuen Erkenntnisse nicht direkt bestritten. Erst wenn wir die weitere Aussage der Mutter von X hätten, dass X kein Auto sein Eigen nennt und keinesfalls das Geld dafür hätte, wäre das ein Gegengrund gegen p im Audi-Beispiel.

Unterminierer kennen wir bereits aus der Wahrnehmung. Wenn wir etwa schließen, ein bestimmter Gegenstand sei rot (p), weil wir ihn als rot wahrnehmen (q), dann wäre ein Unterminierer (u) die Information, dass wir ihn nur in rotem Licht gesehen haben (u). Hier spricht der Unterminierer nicht dagegen, dass uns der Gegenstand rot erscheint (also nicht gegen q und auch nicht gegen p), aber er besagt, dass sich diese Wahrnehmung unterschiedlich deuten lässt. Neben der Möglichkeit, dass der Gegenstand rot ist, gibt es noch andere Möglichkeiten wie die, dass er eigentlich weiß ist und nur das rote Licht reflektiert hat, die in unserer Situation unsere Wahrnehmung genauso gut erklären würden. Wir erhalten so alternative Hypothesen für unsere Wahrnehmung, die genauso gute Erklärungen für sie darstellen wie die, die wir gerne begründen möchten.

Damit wurde p durch u als mehrdeutig erwiesen und spricht nicht mehr eindeutig für p. Wenn die Unterminierer oder die Gegengründe jeweils stark genug sind, dann sprechen wir davon, dass sie eine bestimmte positive Begründung q für p *neutralisieren*, also ¬(u & q⇒p). Wir werden später auf die Frage nach der Stärke von Argumenten noch zurückkommen, müssen damit aber zunächst einmal intuitiv umgehen. Die Stärke ließe sich eventuell mit Hilfe von Wahrscheinlichkeiten darstellen, aber die wären hier letztlich nur ein Buchhaltungssystem für unsere intuitiven Einschätzungen. Leider wird die Idee der Unanfechtbarkeit und der Neutralisierung sogleich komplexer, sobald wir kompliziertere Beispiele betrachten.

Relevante Unterminierer. Wir erleben in manchen Beispielen das komplexe Wechselspiel von Gründen, Unterminierern und Gegengründen sowie weiteren Gegengründen oder Unterminierern, die sich gegen die Unterminierer und Gegengründe selbst richten können etc. Dabei erweist sich *Wissen* als das angestrebte relativ *stabile Ziel*. Dazu müssen wir allerdings für die Begründungen von p fordern, dass sie stabil gegenüber Unterminierern und Gegengründen sind. Als vierte Wissensbedingung werden wir daher zunächst verlangen, dass es zu unserer Begründung von p keine *relevanten Unterminierer* (auch unbekannterweise) gibt, denn gäbe es sie, wären unsere Begründungen nicht viel wert und die Wahrheit von p wiese mit unseren Begründungen dafür keinen genuinen Zusammenhang mehr auf, wäre also wiederum nur ein glücklicher Zufall, so dass sich erneut Gettier-Beispiele entwickeln ließen. Allerdings hatte ich schon ganz bewusst von »relevanten« Unterminierern gesprochen.

Das bekannte *Tom-Grabbit-Beispiel* zeigt, dass wir uns nicht vor allen Unterminierern fürchten müssen. Nehmen wir an, wir hätten jemanden in der Bibliothek beobachtet, der dort ein Buch gestohlen hat und wir glauben, ihn als Tom Grabbit erkannt zu haben (q). Da Tom Grabbit tatsächlich der Dieb war, ist p ≡ »Tom Grabbit hat in der Bibliothek ein Buch gestohlen« Wissen für uns, das durch q begründet wird: q⇒p. Doch nehmen wir nun weiter an, dass es einen Unterminierer u dazu gäbe: Die Mutter von Tom Grabbit erzählt jemandem, dass Tom einen kleptomanischen Zwillingsbruder hat, der im Moment in der Stadt weilt, während Tom selbst sich auf den Malediven befindet (u). Bereits diese Erzählung (u besteht hier in den Aussagen der Mutter und nicht

in den darin behaupteten Fakten) untergräbt unsere Belege für die Täterschaft Toms. Würden wir sie unseren Gründen für p hinzufügen, käme zusammengenommen keine Begründung mehr für Toms Diebstahl heraus: $\neg((u \,\&\, q) \Rightarrow p)$.

Allerdings könnte dieser Unterminierer selbst wieder durch andere Aussagen unterminiert werden. So erzählt vielleicht der Psychiater von Toms Mutter, dass Toms Mutter in einer Phantasiewelt lebe, in der alles Schlechte, was Tom verbricht, auf einen imaginären Zwillingsbruder von Tom geschoben würde, den es aber nicht gibt, denn Tom ist ein Einzelkind (u*). Dann wären q und u und u* zusammen wieder eine Begründung für p, weil unser Unterminierer u in seiner unterminierenden Wirkung selbst untergraben wurde bzw. durch Gegengründe *neutralisiert* wurde. Damit wäre u kein relevanter Unterminierer mehr, denn es gilt: $(u^* \,\&\, u \,\&\, q) \Rightarrow p$. Oder man könnte es auch so ausdrücken, dass u* einen guten Grund dafür darstellt, dass u unsere ursprüngliche Begründung nicht mehr untergräbt, d.h., danach gilt: $u^* \Rightarrow \neg(u \Rightarrow \neg(q \Rightarrow p))$. Man sieht also hier schon, dass die Iterationen übersichtlicher sind, wenn wir die einfachere erste Schreibweise wählen, was ich daher im Folgenden überwiegend tun werde, obwohl diese vereinfachte Schreibweise nicht die komplexen inneren Zusammenhänge deutlich macht.

Leider lässt sich das Spiel fortsetzen. Auch u* könnte wiederum neutralisiert werden, etwa durch den Beichtvater des Psychiaters, dem er anvertraut hat, dass er aus bloßem Hass auf Tom Toms Mutter der Lüge bzw. der Verrücktheit bezichtigt (u**) und wir erhielten wieder:

$\neg((u^{**} \,\&\, u^* \,\&\, u \,\&\, q) \Rightarrow p)$
was wiederum bedeuten würde: $u^{**} \Rightarrow \neg(u^* \Rightarrow \neg(u \Rightarrow \neg(q \Rightarrow p)))$

Also können Unterminierer, die neutralisiert werden, doch wieder relevant werden, wenn ihr Neutralisierer ebenfalls wieder neutralisiert wird usw. Die Relevanz von Unterminierern muss also letztlich eine *holistische Stabilitätsbedingung* erfüllen. Danach ist ein Unterminierer *relevant*, wenn es für ihn keine stabile Neutralisierung gibt, d.h., wenn es für ihn *keine* Neutralisierung gibt, die sich auch bei Hinzunahme weiterer wahrer Aussagen letztlich als stabile Neutralisierung darstellt. Wir gehen hier natürlich davon aus, dass es nicht zu einem unendlichen Spiel

von Neutralisierungen und Gegenneutralisierungen kommt, sondern wir nach endlich vielen Schritten einen stabilen Zustand erreichen, der unser lokales Argumentgefüge nicht mehr ins Wanken bringt. Bleibt dann kein stabiler Neutralisierer für u mehr übrig, ist u ein relevanter Unterminierer. Intuitiv dürfte damit hoffentlich geklärt sein, was wir unter einem *relevanten Unterminierer* zu verstehen haben.

In Pollock (1994) finden wir eine formale Explikation solcher Beziehungen in Form einer speziellen nicht-monotonen Logik. Er bringt die möglichen Gründe und Gegengründe sowie die Unterminierer in einen Inferenzgraphen und betrachtet dann die Bewertungen der Knoten als neutralisiert oder nicht neutralisiert. An die Bewertungen werden rekursive Anforderungen gestellt. Die Ursprungsknoten werden als prima facie plausibel und damit nicht neutralisiert angesehen. Sie stellen unsere akzeptierten Daten dar. Welche Schlussfolgerungen dann daraus zu ziehen sind, zeigen uns die Bewertungen. Nur wenn bei allen zulässigen Bewertungen unsere Gründe für p nicht neutralisiert werden, gilt p als stabil begründet. In der Praxis werden solche Kaskaden von Unterminieren sicher nur selten vorkommen. Daher werde ich ihnen hier nicht sehr viel Platz einräumen. Es dürfte auch so gut verständlich sein, wonach wir suchen. Etwas einfacher lässt sich das so darstellen:

(UK) $u = u_0, u_1, \ldots, u_n$ ist eine ***Unterminiererkette*** zu $q \Rightarrow p$ gdw gilt:
$u \Rightarrow \neg(q \Rightarrow p)$
$u_1 \Rightarrow \neg(u \Rightarrow \neg(q \Rightarrow p))$
...
$u_n \Rightarrow \neg(u_{n-1} \Rightarrow \neg(u_{n-2} \Rightarrow \ldots \neg(u \Rightarrow \neg(q \Rightarrow p)\ldots)$

Für einen wahren und relevanten Unterminierer u gilt dann, dass jede solche maximale Kette $u=u_0, u_1, \ldots, u_n$, für die es also keine weiteren Unterminierer gibt, nach einer geraden Anzahl von Schritten endet (n=2k), weil damit die ursprüngliche Behauptung $q \Rightarrow p$ weiterhin unterminiert bleibt, und es gilt: $\neg(u_n \,\&\, \ldots \,\&\, u_1 \,\&\, u \,\&\, q \Rightarrow p)$.

(RWU) u ist ein ***relevanter wahrer Unterminierer*** von q bzw. von $q \Rightarrow p$ gdw gilt:

(1) u ist wahr und

(2) Alle maximalen Unterminiererketten $u = u_0, u_1, \ldots, u_n$ zu u und $q \Rightarrow p$ mit ausschließlich wahren Aussagen u_i enden mit einem geraden $n = 2k$.

Damit wird ein wahrer Unterminierer dann relevant, wenn sich jeder wahre Neutralisierer, den wir für ihn finden, selbst wieder neutralisieren bzw. entwerten lässt, so dass er schließlich doch seine unterminierende Wirkung entfalten kann. Im Folgenden spreche ich meist etwas kürzer von Unterminieren, meine aber meistens *wahre* Unterminierer. Auf die möglicherweise falschen Unterminierer kommen wir gleich und besonders im Rahmen der Debatte unserer epistemischen Pflichten zurück.

Tatsächlich können wir derartig komplexe Fälle für Zeugenaussagen in einem Kriminalfall finden – ganz ähnlich wie wir das im Tom-Grabbit-Fall rekonstruiert haben. Bleiben dann relevante Unterminierer übrig, die also nicht wieder stabil neutralisiert werden können, so gefährden sie unsere Wissensansprüche und führen in der Regel zu neuen Gettier-Beispielen.

Es kann solche Unterminierer nun sowohl für die Gründe für p wie für die Gegengründe gegen p geben. Man findet solche Ketten in Ansätzen typischerweise bei Zeugenaussagen. Auf der ersten Stufe gibt es Gründe q_{1i} sowie Gegengründe g_{1j} für die Tatsache p ≡ »M war der Täter«. Das sind jeweils *direkte Zeugenaussagen* von verschiedenen Zeugen der Tat oder des Tatumfeldes. Die ersteren sprechen sich für p aus und die zweiteren dagegen, dass p stattgefunden hat. Alle Aussagen seien prima facie plausibel. Doch dann tauchen auf der zweiten Ebene dazu Unterminierer u_{2ij} auf, die jeweils die Gründe und die Gegengründe in Zweifel ziehen, indem diese Zeugen Aussagen über die direkten Zeugen machen und diese etwa als unglaubwürdig ausweisen. Zu diesen Aussagen kann es schließlich wieder Unterminierer u_{3ij} geben, die die Leumundszeugen der zweiten Stufe untergraben usf. Das sind schon recht spezielle Fälle, aber wir sollten im Prinzip wissen, wie wir damit umzugehen haben. Das können die Beispiele belegen. Zunächst werden uns die Gegengründe noch nicht weiter interessieren und wir betrachten nur die Gründe und ihre Unterminierer.

Als *relevante Unterminierer* zu den Begründungen von p werden jedenfalls nur die Unterminierer betrachtet, die in diesem Spiel letztlich nicht selbst unterminiert werden. Auch in der Wissenschaft können solche Debatten über einige Stufen laufen, bis sich eine gewisse Stabilität einstellt. So kann es etwa alternative Erklärungen bestimmter Daten geben (wie wir das etwa in der Klimadebatte immer wieder gesehen haben s.u.), die die Bestätigungswirkung dieser Daten in Frage stellen. Dann können die alternativen Erklärungen wiederum an anderen Daten scheitern und damit werden bestimmte Konkurrenzhypothesen eliminiert, so dass sie die ursprünglichen Belege nicht mehr unterminieren etc.

Unterminierer in der Wissenschaft. Unterminierer müssen nicht zwingend wahr sein, um unsere Wissensansprüche untergraben zu können. Wenn uns die Unterminierer bekannt sind und wir sie nicht definitiv als falsch ausschließen können, können sie ebenfalls unsere Begründungen zunichte machen. Dafür finden sich gute Beispiele in der Wissenschaft. Tatsächlich werden wir feststellen, dass eine enorm wichtige Gruppe von Unterminierern in der Wissenschaft von vielen Ansätzen zum induktiven Schließen nicht richtig berücksichtigt wird. Viele Ansätze lassen zunächst nämlich nicht erkennen, wieso das Auftreten von *alternativen Erklärungshypothesen* für die Daten unsere bisherige Begründung einer Theorie unterminieren kann.

Kuhn hatte auf dieses Phänomen in seinen Fallstudien zur Wissenschaftsgeschichte hingewiesen, aber in den meisten Begründungsansätzen ist dafür eigentlich kein Platz vorgesehen. Ist es der Fall, dass ein Datum E eine Theorie T bestätigt (E⇒T), so wird der Pfeil darin in Frage gestellt durch eine neue Theorie T', die E ebenso gut erklären kann (bzw. ebenso gut zu E passt) wie T, aber inkompatibel zu T ist. E wird dadurch zumindest als mehrdeutig ausgewiesen. Es weist nicht mehr nur auf eine bestimmte Theorie hin, sondern zugleich auf verschiedene untereinander inkompatible Theorien, was selbstverständlich die Bestätigungskraft für die einzelnen Theorien entsprechend abschwächt. Das ist ähnlich, wie wenn die Verteidigung in einem Mordprozess alternative Verdächtige benennen kann, die ebenso gut zu den Indizien passen wie der Angeklagte selbst. Erst wenn es dem Ankläger gelingt, diese Alternativen als nicht wirklich verdächtig auszuschalten, richten sich

seine Beweise wieder gegen den Angeklagten. Erst dann hat er die Unterminierer selbst wieder unterminiert, wodurch sie ihre Relevanz verlieren.

Das ist für wissenschaftliche Hypothesen ganz ähnlich. Erst wenn es uns gelingt, die Konkurrenzhypothesen zu eliminieren (wenn wir sie etwa anhand der Daten falsifizieren), dann sprechen unsere Daten wieder für unsere ursprüngliche Hypothese. Das vermutlich anspruchsvollste und wichtigste Geschäft der Wissenschaftler ist daher die *Elimination von Konkurrenzhypothesen*, da sie sonst Unterminierer für unsere Wissensansprüche in der Wissenschaft darstellen. Wir sind deshalb gezwungen, diese alternativen Erklärungen bewusst zu suchen und zu eliminieren, da relevante Unterminierer unsere Wissensansprüche selbst dann untergraben, wenn sie uns nicht bekannt sind.

Relevante Unterminierer sind aber nur die Konkurrenzhypothesen, die uns bekannt werden und die nicht anderweitig eliminiert werden können. Daher werden unsere Wissensansprüche nicht schon dadurch untergraben, dass wir irgendwelche Konkurrenzhypothesen nicht beachtet haben. Erst wenn sie nicht selbst zu unterminieren sind, sind sie tatsächlich bedrohlich für uns. Das ist wichtig für einen Einwand gegen das abduktive Schließen (das wir in Kapitel 4 behandeln werden), wonach es auch durch »ungeborene« Hypothesen gefährdet würde. Zumindest gefährden nur die *relevanten* ungeborenen Hypothesen unsere Wissensansprüche. Außerdem dürfen wir bei diesem Einwand nicht vergessen, dass ein induktives Schließen niemals eine Wahrheitsgarantie liefert, weshalb man zwar sagen kann, dass diese neuen Hypothesen unsere Schlüsse zwar im Einzelfall unbrauchbar machen können, aber nicht generell gegen das abduktive Schließen sprechen.

Da wir allerdings oft nicht so leicht entscheiden können, ob sich solche Konkurrenzhypothesen eliminieren lassen, sind wir als sorgfältige Wissenschaftler gezwungen, sie so systematisch zu suchen wie möglich und dann einzeln zu widerlegen. Nur Konkurrenzhypothesen, die wir von vornherein für abwegig halten, müssen wir nicht unbedingt einbeziehen, da wir davon ausgehen dürfen, dass sie sich bei genauerer Betrachtung als *nicht relevant* entpuppen würden, weil sie im Prinzip selbst wieder unterminierbar sind. Diese Idee der Elimination von Unterminieren werden wir als Nächstes im Rahmen der eliminativen Induktion wie-

derfinden, aber sie gehört letztlich in alle Induktionsverfahren. Fehlt sie also an irgendeiner Stelle, deutet das nur auf die Unvollständigkeit dieser Verfahren hin. Daher wird es ein wichtiges Anliegen des Buches sein, diese *komparative Form der Theorienbestätigung* immer wieder aufzudecken und in die Verfahren zu integrieren.

Direkte Gegengründe. Leider gilt es, neben den Unterminierern ebenfalls noch die möglichen *direkten Gegengründe* gegen p zu berücksichtigen. Sind die sehr stark, können sie genauso Wissen verhindern wie die Unterminierer und neutralisieren so indirekt ebenfalls bestimmte unserer Begründungen für p. Überhaupt bleibt leider das Grundproblem offen, wie wir mehrere Gründe und Gegengründe miteinander verrechnen sollen. Dazu gibt es zunächst kein einfaches Verfahren, das belegen schon die Beispiele der Zeugenaussagen. Bei einander widersprechenden Zeugenaussagen ist es sehr schwierig, sich ein Gesamturteil zu bilden, und wir sind vielmehr auf intuitive Urteile oder weitere Informationen angewiesen, als dass es dafür eine Art von Algorithmus gäbe. So werden wir normalerweise mehrere Zeugen, die übereinstimmend M als den beobachteten Täter identifizieren, als eine stärkere Begründung für p betrachten, als wenn wir weniger oder nur einen Zeugen hätten. Sollten allerdings die Zeugenaussagen einander wieder zu ähnlich sein und sich sogar im Wortlaut gleichen, kann das ins Gegenteil umschlagen. Das kann auf gegenseitige Absprachen oder Beeinflussungen hinweisen – bis hin zu einer Verschwörung –, die sogar dafür sprechen würde, dass M nicht der Täter ist. Wir müssen uns jeweils fragen: *Was ist insgesamt die beste Erklärung für diese Übereinstimmung?* Besteht die beste Erklärung darin, dass die Zeugen eben dasselbe gesehen haben, oder gibt es andere Erklärungen wie die, dass sie sich abgesprochen haben oder noch ganz andere Erklärungen.

Ein kleines Experiment, das ich kürzlich im Fernsehen sah, belegt, wie brisant diese Frage tatsächlich ist. Darin wurden mehrere Zeugen, die einen vermeintlichen Täter kurz, aber deutlich gesehen hatten, direkt nach der Tat 5 Verdächtigen gegenübergestellt. Fast alle von ihnen identifizierten den Verdächtigen Nummer 5 als den Täter, obwohl sich der wahre Täter überhaupt nicht unter den Verdächtigen befand. Trotzdem waren sich viele der Zeugen sehr sicher in ihrer Einschätzung, und Kriminologen bestätigen, dass das leider kein Einzelfall ist.

Im geschilderten Experiment war es leicht zu erklären, worauf dieses angebliche Wiedererkennen beruhte. Der tatsächliche Täter hatte Locken und der einzige Verdächtige mit Locken war Nummer 5, der dem Täter allerdings im Übrigen nicht besonders ähnlich sah. Die Übereinstimmung der Zeugen war hier also anders zu erklären als dadurch, dass die Zeugen alle den Täter erkannt haben. Das ist ein weiteres Beispiel dafür, warum selbst in diesen einfachen Fällen keine Akkumulation der Begründungsstärke vorliegt und wir wieder nach alternativen Erklärungen für die Übereinstimmung Ausschau halten müssen. Außerdem sollte klar sein, dass die subjektive Sicherheit der Zeugen ebenso wenig als guter Hinweis auf die Wahrheit ihrer Aussagen dienen kann.

Insbesondere die »Verrechnung« der positiven Gründe mit Gegengründen ist also kein leichtes Geschäft – noch nicht einmal für Zeugenaussagen. Dafür können wir jedenfalls keine einfachen Regeln formulieren und es helfen auch keine probabilistischen Tricks. Wir hätten z.B. zu beurteilen, wie groß $P(p|q_1 \& \ldots \& q_n \& g_1 \& \ldots \& g_k)$ mit positiven Gründen q_i für p und Gegengründen g_i ist. Dazu gibt es aber keine allgemeinen Rechenregeln, die das auf einfachere Ausdrücke zurückführen würden, und das letzte Beispiel der Zeugenaussagen deutet schon darauf hin, dass das auch besser so ist, denn je nach speziellem Kontext sollten völlig andere Ergebnisse dabei herauskommen. (In Kapitel 5.9.8 werden wir noch erfahren, wie der Bayesianismus das Problem behandelt.)

Haben wir es mit starken Einwänden zu tun, spricht das in wissenschaftlichen Kontexten zunächst einmal dafür, p nicht als hinreichend begründet anzusehen. Doch wir wissen schon aus der Wissenschaftsgeschichte, dass manchmal Theorien akzeptiert werden, obwohl es starke Gegengründe dagegen gibt. Hier kommen wieder spezielle Regeln für das Akzeptieren wissenschaftlicher Theorien ins Spiel, da diese keine einfachen Konjunktionen von Aussagen sind, sondern eine komplexere Struktur haben, die es ihnen ermöglicht, eine differenzierte Reaktion auf bestimmte Anwendungsprobleme zu zeigen (vgl. etwa Bartelborth 1996 Kap. VII).

Ein bekanntes Beispiel ist Bohrs Atommodell, nach dem die Elektronen nur auf ganz bestimmten stabilen Bahnen um den Atomkern kreisen konnten. Obwohl das im Widerspruch mit der im Wesentlichen akzeptier-

ten klassischen Elektrodynamik stand, wurde die bohrsche Theorie nicht gleich in Bausch und Bogen verworfen. Für dieses Verhalten lässt sich eine Erklärung finden, denn hier sind bestimmte Kohärenzüberlegungen am Werk. Bohr nahm als Hilfshypothese an, dass die *gebundenen Elektronen* im Atom nicht der Maxwellschen Elektrodynamik unterliegen. Er verkleinerte also einfach den *intendierten Anwendungsbereich* der klassischen Elektrodynamik, so dass keine direkten Widersprüche zu seinem Atommodell mehr auftraten. Solange man aber die Maxwellsche Elektrodynamik nicht wirklich in Zweifel zog, lieferte sie natürlich weiterhin starke prima facie Gegengründe gegen das Bohrsche Atommodell, denn diese Ausnahmeregelung wirkte recht ad hoc. Trotzdem konnte das Atommodell mit einer gewissen Berechtigung beibehalten werden (vgl. Bartelborth 1989).

Jedenfalls steht uns für entwickelte wissenschaftliche Theorien dieser Ausweg immer offen, dass wir ihren *intendierten Anwendungsbereich* geeignet einschränken, wodurch sonst auftretende Inkonsistenzen vermieden werden können. Das gilt vor allem für den Fall, in dem eine Theorie direkt mit bestimmten Daten in Konflikt gerät. Das ist zwar prima facie ein Gegengrund für das Akzeptieren der Theorie, es gibt aber hier denselben Ausweg wie oben, der sogar wieder zu Wissen führen kann, denn *jede* Theorie ist immer auf einen ganz bestimmten *intendierten Anwendungsbereich* bezogen. Jede Theorie erklärt bestimmte Phänomene und andere eben nicht. Die Gravitationstheorie ist zwar zunächst auf alle Objekte anwendbar, ist aber keineswegs dazu gedacht, die Bewegungen und Handlungen von Menschen erklären zu können. Psychologische Theorien, die das leisten sollen, treten also keinesfalls in Konkurrenz zur Gravitationstheorie, die die Anziehungskräfte unter den Menschen anders bestimmt.

Newton hoffte zwar, seine Partikelmechanik sogar auf Lichtphänomene (als Theorie der Lichtpartikel) anwenden zu können, jedoch eine ganze Reihe solcher Phänomene wie die Interferenz oder die Beugung ließen sich nicht unter seine Theorie subsumieren. Deshalb ist Newtons Theorie noch nicht gleich in Gänze zu verwerfen, sondern eine naheliegende Antwort auf diese Herausforderung ist, nur diese speziellen Phänomene aus dem Anwendungsbereich seiner Theorie zu streichen. Für andere Phänomene scheint die Theorie hingegen erfolgreich anwendbar zu

sein, und wir können sie noch heute für bestimmte Bereiche zumindest als approximativ gültig ansehen. Wir müssen jeweils genau hinsehen, welche Hypothese mit welchem Anwendungsbereich zur Debatte steht, wenn wir die Frage nach wissenschaftlichem Wissen entscheiden wollen. Gegen die newtonsche Partikelmechanik mit verkleinertem Anwendungsbereich stellen die Lichtphänomene eben keine Gegengründe mehr dar.

Und selbst Gegengründe, die wir auf diese Weise nicht auflösen können, müssen nicht immer sogleich dazu führen, dass kein Wissen mehr vorliegt. Zunächst einmal ist klar, dass Gegengründe, die uns bekannt sind, in jedem Fall zählen und zu berücksichtigen sind. Wiegen sie schwer genug, dass sie unsere bisherigen Gründe für p aufwiegen, stellt p kein Wissen dar. Da hilft es auch nicht, wenn es für die Gegengründe Unterminierer gibt, solange diese uns nicht bekannt sind. Sind sie uns bekannt, können sie die Gegengründe natürlich neutralisieren und somit entkräften, so dass wir wieder berechtigt sind, an p zu glauben.

Doch wie sieht das mit *unbekannten Gegengründen* aus? Das ist schon wesentlich schwieriger zu beantworten. Zunächst einmal zählen wiederum nur *bestimmte* unbekannte Gegengründe (vgl. auch Steup 2006). Betrachten wir dazu ein Beispiel: Nehmen wir etwa an, Kommissar K hätte gute Gründe (q), daran zu glauben, dass M der gesuchte Mörder von Fritz ist (p), und er wäre es tatsächlich. Nun gibt es allerdings einen Menschen in Sibirien, der behauptet (ohne dass Kommissar K davon wüsste), er hätte M zum Tatzeitpunkt dort gesehen (g), wodurch dieser ein Alibi hätte. Damit wäre die Aussage des sibirischen Zeugen unser Gegengrund zu der Aussage p ≡ »M hat Fritz ermordet.« Jedoch, wie das eben so passieren kann, hat er sich dabei geirrt (u). Er hat jemand anderen für M gehalten. Die Tatsache des Irrtums wäre ein Unterminierer für g, und ein solcher muss letztlich vorliegen, da wir angenommen haben, dass p wahr ist. Wenn p aber wahr ist, so kann das Wissen von Kommissar K, dass p, doch nicht durch solche unbekannten Zeugenaussagen gefährdet werden. Die Aussage des Alibizeugen scheint für eine epistemische Bewertung von Ks Überzeugungen offensichtlich bedeutungslos, denn er hat letztlich nichts Relevantes zu p zu sagen. Sobald der Unterminierer u bekannt würde, ist klar, dass der Alibizeuge nur eine irrelevante Anmerkung zu bieten hatte. Außerdem befindet sich der Alibizeuge weit außerhalb der Ermittlungen von Kommissar

K. Sollten solche Gegengründe bereits gegen Wissen ausreichen, käme es wohl nur selten zu Wissen, gerade auch in der Wissenschaft. Gegengründe, die wir nicht kennen, scheinen daher keine Gefährdung von Wissen darzustellen. Ganz so überzogen sind unsere Anforderungen an eine Wissenskonzeption dann doch nicht. Das gilt selbst dann, wenn wir keine Unterminierer angeben könnten, denn die Gegengründe sind zumindest de facto nicht sehr stark, weil p schließlich wahr ist.

Epistemische Pflichten. Unsere *epistemischen Pflichten* scheinen allerdings anders auszusehen, wenn wir einen derartigen Gegengrund *kennenlernen*, vor allem ohne einen Unterminierer dagegen zu kennen. Dann sollten wir genau genommen nicht mehr fest an p *glauben*, und damit wäre schon ein erstes Abweichen von den klassischen Wissensbedingungen gegeben. Auch verfügen wir – alles zusammengenommen – nicht mehr über eine Begründung für p, denn es gilt schließlich: ¬(q & g⇒p). Diese Schwächung der persönlichen epistemischen Position des Subjekts wird dazu führen, dass wir noch nicht einmal mehr sagen dürfen, wir könnten p noch rationalerweise glauben, und resultiert damit in einem Verlust des Wissensstatus für p. Das ist typisch für den Wissensbegriff, in dem hier zwei Aspekte zusammenkommen: Erstens müssen die subjektiven (internalistischen) Bedingungen für einen rationalen Glauben erfüllt sein, was in unserem Beispiel fehlt; zweitens müssen bestimmte objektive (externalistische) Anforderungen eingehalten werden, was in unserem Beispiel zwar erfüllt ist, denn der Gegengrund wird womöglich objektiv gesehen durch einen Unterminierer neutralisiert, aber das kann die Mängel des ersten Aspekts nicht wettmachen. Kommissar K sollte also in dem Fall, dass der Alibizeuge ihm gegenüber seine Behauptung macht, kein Wissen mehr zugeschrieben werden, dass M der Täter ist.

Umgekehrt sieht es selbstverständlich genauso aus. Selbst wenn wir festen Glaubens an p sind und über allerbeste Gründe für p aus unserer Sicht verfügen, sollte p nicht wahr sein oder sollten diese Gründe gettieranfällig sein, liegt trotzdem kein Wissen vor. Besonders die Defizite der objektiven Wissensaspekte sind nicht durch die subjektiven zu heilen. Wurden die DNS-Spuren, die auf M als Täter hinweisen, tatsächlich falsch ausgewertet oder sogar verwechselt, aber M sei trotzdem der Täter, haben wir wieder kein Wissen mehr auf Seiten von Kommissar K, sondern nur

ein neues Gettierbeispiel konstruiert. Trotz irreführender Gründe liegen wir zufälligerweise richtig mit unseren Vermutungen. Das vorherige Beispiel belegt aber insbesondere die spannendere andere Richtung und zeigt die hohe Sensibilität des Wissensbegriffs schon gegenüber kleinen subjektiven epistemischen Defiziten.

Das Problematische an dieser Stelle ist nun insbesondere, dass die Wissensfrage bereits daran hängen kann, in welchen Kontexten der Alibizeuge auftritt. Gehört der Alibizeuge zum Umfeld des Tatgeschehens (ist er z.B. ein Nachbar des Tatverdächtigen) und M behauptet, er wäre zum Tatzeitpunkt in seinem Garten gewesen, dann hätte es zu den Aufgaben des Kommissars gehört, diesen Nachbarn zu befragen. Hat er das schlicht vergessen oder möchte seine bisherigen Ermittlungsergebnisse einfach nicht durch mögliche Alibis von M gefährden, so würden wir diese *epistemischen Pflichtversäumnisse* Kommissar K anlasten und nicht mehr von Wissen sprechen, selbst wenn sich der Nachbar geirrt hat. Kommissar K hätte bei sorgfältiger Arbeit die Zeugenaussage kennen müssen. Wir werden ihn erkenntnistheoretisch nicht belohnen, in dem wir ihm Wissen aufgrund seiner Nachlässigkeiten zuschreiben. Es wäre schon seltsam, wenn wir einfach dadurch unser Wissen aufrechterhalten könnten, dass wir naheliegende Gegengründe nicht zur Kenntnis nehmen. Auch Kommissar K selbst sollte ein ungutes Gefühl bei seiner Präsentation der Ermittlungsergebnisse haben. Fragt man ihn, ob er auch sorgfältig in beide Richtungen ermittelt hat, sollte ihm klar sein, dass er nicht wirklich nach Entlastungszeugen gesucht hat. Dann liegt wieder ein Defizit auf der subjektiven Seite vor, das m.E. bereits groß genug ist, um Wissen zu verhindern. Es können in diesem speziellen Fall also auch unbekannte und möglicherweise sogar selbst unterminierte (und damit irrelevante) Gegengründe dem Wissen entgegenstehen. Das ist immer dann der Fall, wenn es sich um Gegengründe handelt, die das epistemische Subjekt hätte zur Kenntnis nehmen müssen, wenn es seinen epistemischen Pflichten genügt hätte.

Das findet sich in ähnlicher Form für *wissenschaftliche Beispiele*. Hat ein Wissenschaftler S gute Gründe für die wahre Hypothese h und ist fest von h überzeugt, so schadet es dem Wissensstatus von h nicht, wenn ein besonders renommierter Wissenschaftler N (womöglich ein Nobelpreisträger) behauptet, er habe zwingende Daten, die gegen h sprächen, wenn

N das nur in engstem Familienkreise tut. Sollte S mit seiner Hypothese h Recht behalten und die Aussagen des Nobelpreisträgers sind für S nicht zugänglich, so kann S durchaus über das Wissen verfügen, dass h gilt. Hat der Nobelpreisträger seine Resultate aber in einer geeigneten Fachzeitschrift veröffentlicht, die S eigentlich hätte zur Kenntnis nehmen müssen, untergräbt das bereits wieder die Wissensansprüche von S, selbst wenn er es nicht gelesen hat. Man wird in dem Fall nämlich gegen ihn einwenden, dass es zu seinen *epistemischen Pflichten* für das wissenschaftliche Arbeiten gehört, solche Arbeiten zur Kenntnis zu nehmen, wenn sie für ihn im Prinzip zugänglich sind. Hier verläuft ein schmaler Grat zwischen Wissen und Nichtwissen, der auch eine Grauzone aufzeigt, in der unsere Intuitionen nicht ganz klar sind, ob es sich noch um Wissen handelt oder ob die Verstöße gegen unsere epistemischen Pflichten schon zu groß sind.

Für die Wissenschaft hat die Rede von *epistemischen Pflichten* noch einen relativ klaren Inhalt, aber in anderen Kontexten scheint das schon viel weniger klar zu sein. Welche derartigen Pflichten hat etwa ein Redakteur des Kulturteils einer Zeitung oder jemand am Stammtisch im Freundeskreis, wenn er dort seine Überzeugungen vertritt? Die Vagheiten und Unklarheiten, die dadurch noch in der Wissensdefinition enthalten sind, finden sich aber bereits im umgangssprachlichen Wissensbegriff selbst. Jedenfalls soll das intuitive Ergebnis dieser kurzen Beispieldebatten sein, dass wir zumindest solche Gegengründe zur Kenntnis nehmen müssen, die uns entweder bekannt sind oder die uns im Prinzip zugänglich sind, und die wir gemäß unseren epistemischen Pflichten zur Kenntnis nehmen müssten, sogar wenn die Gegengründe eigentlich irrelevant sind, da es dazu Unterminierer gibt.

Weitere Verstärkungen unserer Anforderungen für das Vorliegen von Wissen sind zwar denkbar und werden manchmal erwogen, ergeben jedoch keine sinnvolle Wissenskonzeption. Selbst noch so gute Gründe für eine Aussage p stellen normalerweise *keine Wahrheitsgarantie* für p dar. Eine solche absolute Irrtumssicherheit erwarten wir auch nicht von guten Gründen. Gründe sind *Hinweise* auf die Wahrheit einer Behauptung p, aber sind nicht unfehlbar (sonst hätten wir es wohl mit deduktiven Begründungen zu tun). Eine solche Forderung der Wahrheitsgarantie würde selbst wissenschaftliche Begründungen überfordern und

wäre sogar ein Rückfall in dogmatische Denkstrukturen. Nach einer wahrheitsgarantierenden Begründung hätten wir keinen Grund mehr, anderen in dieser Sache noch einmal zuzuhören. Sie könnten keine relevanten Gegengründe oder Zusatzinformationen mehr liefern, denn unsere Sache wäre ein für alle Mal entschieden. Diese dogmatische Denkweise entspricht aber überhaupt nicht der wissenschaftlichen Einstellung, die vorsieht, dass wir immer offen für neue Daten sein müssen und unsere bisherigen Theorien alle als *fallibel* betrachten sollten. Daher wird man normalerweise keine solche Verstärkung unserer Begründungen verlangen, obwohl das ebenfalls ein Weg zu einer Wissensdefinition im Sinne einer Vermeidung der Gettierbeispiele sein könnte.

Allerdings dürften wir dann kaum noch echte Instanzen von Wissen erwarten, weil diese Anforderungen praktisch unerfüllbar wären. Wir stoßen jedoch immer wieder auf die Behauptung, Wissen wäre auf solche zwingenden Begründungen angewiesen. Linda Zagzebski (1994) behauptet sogar nachweisen zu können, dass nur wahrheitsgarantierende Gründe uns vor Gettierbeispielen schützen können. Da ich ihr hierin nicht folgen kann, werde ich diesen Punkt aber nicht weiter diskutieren.

Denken wir noch einmal an unser Beispiel von Kommissar K. Zusammenfassend gilt: Wenn sich der Alibizeuge nicht bei uns in Deutschland meldet und unser ermittelnder Kommissar nicht den geringsten Hinweis darauf hat, dass er existiert, hat er dann kein Wissen, dass p (M ist der Mörder), weil es irgendwo einen solchen Gegengrund gibt? So starke Anforderungen stellen wir nicht. Hat der Kommissar sorgfältig ermittelt, sprechen wir ihm nun Wissen zu. So weit entfernte Gegengründe können das nicht verhindern. Anders sähe es nur aus, wenn der Kommissar sich sofort auf M festgelegt hätte und nicht einmal mehr die Zeugen aus dem Umfeld von M und Fritz vernommen hätte. Dann ist er in diesem Fall seinen *epistemischen (und polizeilichen) Sorgfaltspflichten* nicht genügend nachgekommen. Gibt es also relevante Gegengründe, auf die jemand gestoßen wäre, wenn er seinen Sorgfaltspflichten in der entsprechenden epistemischen Situation genügt hätte, dann können diese Gegengründe unserem Wissen entgegenstehen. Sie müssen zumindest gezählt werden. Sind sie dann etwa genauso so stark wie unsere ursprünglichen Gründe für p liegt kein Wissen mehr vor. Hier könnte uns höchstens noch eine komplexe Abwägung retten, die sagt, dass unsere ursprünglichen Gründe

für p deutlich stärker als alle relevanten Gegengründe wären, um doch noch Wissen zu erreichen.

Wissenschaftliches Wissen. Diese Wissenskonzeption lässt sich besonders gut auf den Fall von *wissenschaftlichem Wissen* übertragen. Gerade Wissenschaftler haben bestimmte Sorgfaltspflichten zu erfüllen, bevor sie etwas als Ergebnis vorstellen dürfen. Dem werden wir im Weiteren noch nachgehen, aber hier können wir schon nennen, dass sie u.a. gezielt nach Daten Ausschau halten müssen, die ihren Hypothesen entgegenstehen und dass sie außerdem nach alternativen Erklärungen für ihre Daten suchen müssen. Dazu müssen sie u.a. die einschlägige Literatur durchforsten, auf der Suche nach Hinweisen auf solche Gegengründe durch Kollegen. Verletzen sie ihre epistemischen Pflichten, kann das der Zuschreibung von Wissen entgegenstehen, selbst wenn sie zufälligerweise trotzdem richtig liegen. Damit kommen wir schließlich zu der folgenden Wissensdefinition:

(WW) Wissenschaftliches Wissen: S weiß, dass p bei einem Überzeugungssystem X von begründet akzeptierten Aussagen gdw. gilt:
(1) S ist von p überzeugt.
(2) p ist wahr.
(3) S verfügt in X über gute Gründe G für p.
(4) Es existieren keine wahren relevanten Unterminierer U zu G und keine bekannten, noch nicht widerlegten Unterminierer U zu G.
(5) Alle möglichen (starken) Gegengründe zu G, zu denen es in X keine relevanten Unterminierer gibt, liegen außerhalb des Bereichs der epistemischen Pflichten von S in der hier relevanten Situation und sind S nicht bekannt.

Die fünfte Bedingung weist sowohl eine größere Vagheit als die anderen Bedingungen auf als auch eine gewisse *Kontextabhängigkeit*, die in den anderen Bedingungen so nicht vorliegt. Wir müssen damit zugestehen, dass der Wissensbegriff vermutlich nicht völlig kontextfrei verstanden werden kann.

Außerdem wird deutlich, dass die Unterminierer nur zählen, wenn sie wahr und stabil sind oder wenn sie uns bekannt und noch nicht widerlegt sind. Dafür müssen sie im ersten Teil also noch nicht einmal dem epistemischen Subjekt bekannt sein (hier haben wir es also mit einer typischen externalistischen Bedingung zu tun). Demgegenüber zählen die Gegengründe erst, sobald sie uns bekannt werden oder sobald sie uns bekannt sein sollten, wenn wir unsere epistemischen Pflichten ernst nähmen. Allerdings werden sie durch bekannte Unterminierer neutralisiert, die selbst nicht unbedingt wahr sind, sondern nur begründet akzeptiert sein müssen. Damit handelt es sich bei der fünften Bedingung überwiegend um eine internalistische Bedingung.

Außerdem zeigen sich schon weitere Merkmale wissenschaftlichen Wissens, die uns im Folgenden weiter beschäftigen werden. Um von Theorien als Wissen sprechen zu können, benötigen wir eine holistische Theorienbewertung, die komplexe Abwägungsprozesse beinhaltet und sich nicht auf einfache Zusammenhänge zwischen Theorien und wenigen Daten beschränkt. Wir werden später sehen, dass diese am besten als ein Schluss auf die beste Erklärung im größeren Rahmen einer Kohärenztheorie der epistemischen Rechtfertigung beschrieben werden können.

Damit haben wir eine erste Explikation von Wissen vorliegen, die mir vor allem für das wissenschaftliche Wissen eine gute Grundlage zu bieten scheint. Das möchte ich im nächsten Kapitel noch weiter verfolgen und in den folgenden Kapiteln als Zielvorstellung vor Augen haben. Aus vergangenen leidvollen Erfahrungen sollten wir allerdings gelernt haben, dass kaum ein Wissensbegriff gegen alle phantasievollen Gettier-Beispiele gefeit ist. Das wird auch hier vermutlich der Fall sein. Darin zeigt sich die Komplexität unserer Wissensvorstellung und schließlich finden sich immer kontextuelle Elemente im Wissensbegriff. Wir könnten uns nun also noch extremere Beispiele ausdenken, die nach immer komplexeren Wissensdefinitionen verlangen. Doch das scheint mir nicht der richtige Weg zu sein. Die Definitionen werden so schließlich unhandlich und zeigen nicht mehr klar die wesentlichen Aspekte des (wissenschaftlichen) Wissens auf, die m.E. in (WW) sehr gut eingefangen sind.

Mir genügt es also, wenn (WW) die Kernbedeutung von Wissen wiedergibt und zumindest die meisten normalen Instanzen von Wissen insbesondere in der Wissenschaft korrekt als solche einstuft. Von Wissen erwarten wir vor allem, dass es eine *stabile Basis* für weitere Schlüsse oder Bewertungen anderer Überzeugungen bereitstellt und das verlangt, dass die Begründungen für unser Wissen nicht sogleich umfallen, sobald neue Erkenntnisse hinzukommen, die relevant sind; was heißen soll, dass sie nicht auf tönernen Füßen stehen dürfen. Wissen sollte also zumindest im direkten Umfeld unanfechtbar sein, bzw. es sollte so gut begründet sein, dass es nicht sofort in Frage gestellt wird, sobald wir uns nur sorgfältig umsehen. Das ist intuitiv verwandt mit der modalen Forderung nach praktischer Sicherheit von Wissen, die wir bereits kennengelernt haben.

Außerdem erwarten wir, dass unser Wissen nicht auf einen simplen Trick zurückzuführen ist, nämlich einfach widerspenstige Daten nicht zur Kenntnis zu nehmen. Gegengründe sind jedenfalls dann zu berücksichtigen, wenn sie für uns leicht zugänglich sind. Gerade in der Wissenschaft gehört diese Offenheit gegenüber Einwänden zu den grundlegenden Spielregeln, auch wenn dagegen de facto manchmal verstoßen wird. So wissen wir leider (Fröhlich 2003), dass selbst die Reviewer wissenschaftlicher Zeitschriften viel eher die Artikel »durchlassen«, die ihrer eigenen Meinung entsprechen, als solche, die davon abweichende Meinungen zeigen. Wirkliche Offenheit gegenüber anderen Ansichten sieht anders aus.

Auch im Falle von »Climategate« (so wurde ein Hackerzwischenfall 2009 an einem Klimaforschungszentrum genannt, in dem Absprachen zwischen Klimaforschern in gehackten E-Mails aufgedeckt wurden) war das Ziel, anderslautende Meinungen von vornherein zu unterdrücken. Das beweist jedoch nicht, dass die meist ungeschriebenen Regeln des wissenschaftlichen Arbeitens nicht gelten. Vielmehr offenbart besonders die dabei an den Tag gelegte Heimlichtuerei geradezu, dass auch den Regelverletzern diese Regeln bekannt sind und sie keineswegs offen dagegen opponieren würden.

Insbesondere müssen wir uns darüber im Klaren sein, dass der Wissensbegriff immer eine Reihe von *externen Bestandteilen* bzw. objektiven Anforderungen aufweist. Das sind Bedingungen, die für Wissen

erfüllt sein müssen, die den Wissenden aber selbst nicht direkt und irrtumssicher kognitiv zugänglich sind und uns als Zuschreibern von Wissen oft genauso wenig bekannt sind. Wenn ich eine Aussage über meine eigenen Wünsche formuliere, dann ist mir ihre Wahrheit vielleicht noch direkt zugänglich, wenn ich sie aber z.B. über die Wünsche eines Anderen treffe, so ist das schon nicht mehr der Fall. Das heißt, wenn ich oder ein Dritter nun mir das entsprechende Wissen über die Wünsche des Anderen zuschreiben möchte, ist er dabei immer auf Vermutungen darüber angewiesen, ob die Bedingung (2) von (WW) auch tatsächlich erfüllt ist. Nur dann sprechen wir aber von Wissen. Solche Wissensbehauptungen sind also selbst immer nur Vermutungen oder man könnte stattdessen sagen, Wissen ist das *Ziel* der Wissenschaft, aber wir können uns (in den meisten Fällen) nicht ganz sicher sein, das Ziel erreicht zu haben, sondern haben dafür nur bestimmte Indizien, denn damit tatsächlich Wissen vorliegt, müssen jeweils bestimmte externalistische Bedingungen gegeben sein.

Ähnliches lässt sich auch für die Bedingung (4) sagen und selbst für (5) gilt, dass sie zumindest z.T. noch externalistisch ist (vgl. dazu Bartelborth 1996 Kap. III.A). Wir müssen also bei allen Wissenszuschreibungen einschränkend sagen, dass S *vermutlich* das Wissen hat, dass p, denn wir haben dann Gründe für die Annahme, dass es keine relevanten Unterminierer gibt und S wohl alle seine epistemischen Pflichten erfüllt hat. Sollte dann trotzdem ein Unterminierer auftauchen, müssen wir eben zugeben, dass wir uns mit dieser Wissenszuschreibung geirrt haben. Von diesen Problemen der externalistischen Elemente in (WW) unabhängig, bleibt aber Wissen das erklärte Ziel der Wissenschaft, und es ging mir hier vor allem darum, dieses Ziel zu präzisieren. Das können wir recht intuitiv noch einmal so beschreiben:

(Wissenschaftliches) Unanfechtbares Wissen
besteht aus einer wahren begründeten Meinung, wobei
(1) (*Stabilität gegenüber Unterminierern*) die Begründung so stabil ist, dass sie durch relevante Wahrheiten nicht neutralisiert werden kann und

(2) (*Stabilität gegenüber Gegengründen*) die Begründung so stark ist, dass sie nicht durch Kenntnis von Gegengründen aus dem Bereich unserer epistemischen Pflichten neutralisiert werden kann.

Das zweite Ziel. Das zweite Ziel der Wissenschaft, nach möglichst *gehaltvollen* Hypothesen zu suchen, wird wieder nicht explizit in (WW) erwähnt. In (WW) geht es nur darum, dass die wissenschaftlichen Erkenntnisse gut abgesichert wurden, so dass sie relativ unanfechtbar sind. Wir sollten nun hinzufügen, dass wir vor allem nach *wertvollem* oder *relevantem* oder *informativem* Wissen streben. Logisch gültige Aussagen sind sicher unanfechtbar, aber kaum das, was wir in den empirischen Wissenschaften anstreben. Aussagen wie »Ein Pferd hat entweder vier Beine oder es hat eine andere Anzahl an Beinen« sind bestechend unanfechtbar, aber ebenso sicher keine wissenschaftliche Erkenntnis. Man muss noch nicht einmal ein Pferdeexperte sein, um eine derartige Aussage treffen zu können.

Wir suchen hingegen nach informativen und empirisch gehaltvollen Einsichten. Das hat wiederum einen pragmatischen Aspekt, aber daneben einen objektiv informationstheoretischen. Unanfechtbarkeit können wir also leichter für relativ triviale Aussagen erzielen, die keine gewagten empirischen Behauptungen über unsere Welt aufstellen. Doch das ist gerade nicht das Ziel der empirischen Wissenschaften. Sie streben nach tiefen Einsichten in das Funktionieren unserer Welt und unseres sozialen Zusammenlebens. Es gehört regelrecht zum Wesen der Wissenschaft, vor allem gewagte Behauptungen aufzustellen und anschließend zu untersuchen bzw. empirisch zu testen. Popper (1984) hat das immer wieder betont und stellt deshalb den *Falsifizierbarkeitsgrad* einer Theorie ganz in den Vordergrund. Seiner Meinung nach ist unter den nicht-falsifizierten Theorien immer die gewagteste Hypothese mit dem höchsten Falsifizierbarkeitsgrad zu wählen. Das ist die Theorie, die die meisten potentiellen Falsifikatoren aufweist – also die meisten Beobachtungsaussagen, die aus der Theorie folgen und an denen die Theorie scheitern könnte. Einmal, weil wir nach möglichst gehaltvollen

Theorien suchen, und zum anderen, weil wir es für solche Theorien am schnellsten aufdecken können, wenn sie denn tatsächlich falsch sind.

Für die *Tiefe der Theorien* denkt er schon an deren *Erklärungskraft*, die, wie ich schon ausgeführt habe, ganz im Zentrum des wissenschaftlichen Erkenntnisstrebens steht. Das wird in der Erkenntnistheorie gern übersehen und man spricht etwa nur davon, wie wahrscheinlich eine Theorie ist (im Empirismus oder im Bayesianismus), ohne sich darüber klar zu werden, dass die hohe Wahrscheinlichkeit für sich allein ein recht langweiliges Ziel darstellt, wenn wir es nicht mit der Forderung nach hohem Gehalt kombinieren. Wir werden später sehen, wie gerade dieser Aspekt in einigen Ansätzen zum induktiven Schließen ganz ausgeblendet wird, weshalb sie nicht besonders gut geeignet sind, die wissenschaftliche Theorienwahl nachzuzeichnen.

Um die Besonderheiten wissenschaftlichen Wissens noch weiter zu klären, möchte ich im Weiteren auf einige Beispiele und einige typische Spielregeln der Wissenschaften eingehen, da der Begriff wissenschaftlichen Wissens einen Ausgangspunkt der weiteren Arbeit darstellt. Durch die Festlegung, in welchem Kontext wir von Wissen sprechen möchten, gewinnt der Wissensbegriff zunächst selbst an Profil. Es wird immer wieder (vermutlich zu Recht) gegen ihn eingewandt, dass er zu viele Facetten aufweist und wir in unterschiedlichen Kontexten u.a. unterschiedlich strenge Anforderungen an das Wissen stellen und dieses deshalb kaum allgemein definierbar sei. Dem begegnen wir hier durch die Konzentration auf den Kontext wissenschaftlichen Wissens, das schärfere Intuitionen dazu bereitstellt, was wir uns von einem solchem Wissen erwarten.

Also werde ich weiter für eine Form von Unanfechtbarkeitskonzeption plädieren, wie sie in (WW) vorgeschlagen wurde. Wissenschaftliches Wissen sollte demnach vor allem gegen bestimmte Anfechtungen abgesichert sein. Die Gründe für unser Wissen sollten von ganz besonderer Qualität sein und vor allem einer kritischen wissenschaftlichen Debatte ein gutes Stück weit standhalten können. Typischerweise ist dabei zu beachten, dass wir bereits bei den *Daten* verschiedene Irrtumsrisiken auszuschließen haben. Idealerweise erheben wir unsere Daten in Experimenten und dabei sind bestimmte Regeln des Experimentierens zu beachten, deren Grundlage wir im Kapitel 7 im Rahmen des kau-

salen Schließens untersuchen werden. Aber selbst in anderen Fällen von Datenerhebungen sind wir z.B. auf Vergleiche zweier Gruppen angewiesen, die möglichst homogen zu sein haben, um den Effekt eines bestimmten Unterschieds besser aufzeigen zu können. Speziell das noch zu schildernde Beispiel von Snows Entdeckung der Cholera bietet ein gutes Vorbild für eine umsichtige Datenerhebung. Auf der Ebene der *Theorien* wird es vor allem darum gehen, möglichst alle plausiblen alternativen Erklärungen zu unserer Hypothese zu finden und als eindeutig minderwertig zurückweisen zu können. Das wird in diesem Kapitel besonders im Rahmen des abduktiven Schließens zu erläutern sein. Die Entdeckung des Kindbettfiebers durch Semmelweiss kann schließlich aufzeigen, wie kreativ man bei der Entwicklung von Alternativhypothesen sein sollte.

2.3.4 Zuverlässige Methoden

Alvin Goldman (z.B. Goldman 1986) hat in mehreren Aufsätzen und Büchern eine *reliabilistische* Konzeption von Wissen entworfen. Danach gehört zum Wissen dazu, dass es durch einen zuverlässigen Prozess oder wir würden besser sagen durch eine *zuverlässige Methode* entstanden ist. Das passt zunächst gut zu einer Analyse wissenschaftlichen Wissens, denn Viele würden das so formulieren, dass das wissenschaftliche Wissen gerade dadurch gekennzeichnet ist, dass es gemäß einer wissenschaftlichen Methode gewonnen wurde.

Wir wissen zumindest ungefähr, was damit gemeint ist, obwohl unsere gleich folgende Debatte über die Pseudowissenschaften und unsere Suche nach einem Abgrenzungskriterium für Wissenschaftlichkeit keine einfachen Ergebnisse liefert. Und auch die Debatte etwa der statistischen Verfahren in Kapitel 6 wird keine einfachen Vorschläge für eine wissenschaftliche Methode liefern. Doch zunächst ist der Vorschlag intuitiv sehr plausibel und knüpft an die modalen Vorschläge an, die genauer angaben, was unter Zuverlässigkeit verstanden werden könnte. Goldmans Vorschlag geht etwas allgemeiner an die Sache heran. Eine *zuverlässige Methode* ist einfach eine Methode, die (auf unserer Erde) eher zu wahren als zu falschen Behauptungen führt. Man könnte

geradezu sagen, dass die Erfolgsquote der Methode ihren Grad an Verlässlichkeit darstellt.

Je verlässlicher die Methode, mit der wir eine Überzeugung p gewonnen haben, umso eher handelt es sich dann bei p um Wissen. Das klingt erst einmal sehr überzeugend. Eine erste Frage ist natürlich, wie zuverlässig die Methode genau sein muss, damit Wissen vorliegt. Doch wir wollen uns nicht in kleinlichen Details verlieren. Eine zweite Frage ist, wie man die Reliabilitätsforderung genau in die Wissensbedingung einbringen kann. Sie kann rein externalistisch gemeint sein und einfach die Begründungsbedingung ersetzen oder es kann sich um eine Mischform handeln, die auch weitere internalistische Anforderungen stellt und etwa verlangt, dass wir eine zuverlässige Fähigkeit eingesetzt haben, über deren Zuverlässigkeit wir auch noch reflektieren können, die dann zu p geführt hat, wie das z.B. in der Tugendepistemologie bei Ernest Sosa (2007) verlangt wird.

Am schwierigsten ist aber eine dritte Frage zu beantworten, die wir anhand unseres Scheunenattrappenbeispiels einführen können. Das Problem in diesem Fall (in dem wir eine letzte echte Scheune in einem Land mit lauter Scheunenattrappen sehen) ist, dass das Sehen einer Scheune zwar normalerweise zuverlässig zu einer entsprechenden Beobachtungsüberzeugung führt, dass eine Scheune vor mir steht (Methode M), aber in unserem speziellen Fall diese Zuverlässigkeit gerade nicht zum Tragen kommt, weil sie in dieser speziellen Umgebung nicht gegeben ist. M ist in normalen Welten und deren Ländern zwar zuverlässig, aber in diesem speziellen Land eben nicht mehr. Dass die Methode i.A. zuverlässig ist, sagt uns nicht genug über den Einzelfall.

Man könnte nun natürlich sagen, wir müssen zur Methode M* übergehen, wonach wir einen Gegenstand mittlerer Größe bei guten Lichtverhältnissen sehen und außerdem eine normale Welt vorliegt, in der es kaum gute Attrappen von solchen Gegenständen gibt. Aber auch diese Methode muss vielleicht verfeinert werden: Wir dürfen nicht betrunken oder sehschwach sein, sonst lassen sich wieder Problemfälle konstruieren. Jede mögliche Fehlerquelle muss im Prinzip ausgeschlossen werden, sonst können wir wieder neue Problemfälle konstruieren. Erst wenn also die Erfolgsgarantie in die Methode eingebaut wird, können wir damit

für Wissen zufrieden sein. Das wäre aber viel zu anspruchsvoll. Unser Wissensbegriff wäre dann kaum noch interessant für uns.

Man spricht auch vom *Allgemeinheitsproblem* (oder vom Referenzklassenproblem), denn der konkrete Prozess, der zu unserer Überzeugung p geführt hat, kann eine Instanz recht unterschiedlicher Typen von Methoden sein. Welchen Typ davon sollen wir heranziehen, um zu beurteilen, ob p in diesem konkreten Fall Wissen ist? Die schöne Idee auf eine allgemeine Konzeption von Methode auszuweichen und sich nicht in die Untiefen der konkreten Begründung von p zu begeben (wie wir es in der Uanfechtbarkeitskonzeption machen), verliert durch diese Vagheit gleich wieder an Biss. Vermutlich finden wir zu jedem spannenden Beispiel von Überzeugungsbildung Methoden, unter die der Prozess fällt, die zuverlässig sind und solche, die es nicht sind. Der Weg zur am engsten beschriebenen Methode – wie oben – droht in einer Methode zu münden, die nur eine Anwendungsinstanz hat und für Wissen die Wahrheit garantieren müsste, denn sein Zuverlässigkeitsgrad wäre damit eins.

So intuitiv die Rede von der zuverlässigen Methode auch ist, sie scheint uns also nicht gleich weiterzuhelfen. Eine derartige Methode ist übrigens, nach möglichst *unanfechtbaren Begründungen* für unsere Behauptungen zu suchen und nur die entsprechend begründeten Aussagen dann zu akzeptieren. Für Wissen geht es uns um den konkreten Einzelfall und daher nicht nur um die eingesetzte Methode und deren allgemeine Zuverlässigkeit (in anderen Fällen). Nichtsdestotrotz werden wir im Folgenden immer wieder bestimmte Methoden unter die Lupe nehmen und analysieren, ob wir Gründe für die Annahme haben, dass es sich um zuverlässige Methoden handelt, die dann zumindest i.A. zu guten Begründungen unserer Überzeugungen führen sollten. Die Frage, ob es sich dabei um Wissen handelt, ist damit aber noch nicht entschieden.

2.4 Wissenschaft und Pseudowissenschaft

Eine Aufgabe, die immer wieder an die Wissenschaftstheorie herangetragen wird, ist die, eine möglichst einfache und klare Demarkationslinie zwischen Wissenschaft und anderen Aktivitäten, etwa den sogenannten

Pseudowissenschaften, festzulegen. Man wünscht sich (verständlicherweise) ein einfaches Kriterium, das es einem gestattet, zügig bestimmte Dinge wie die Astrologie, die Homöopathie, den Kreationismus u.a. als pseudowissenschaftlich zu erweisen, so dass man sie abhaken kann, ohne sich mit ihnen auseinandersetzen zu müssen.

Es wäre natürlich schön und könnte uns eine Menge harter Arbeit ersparen, wenn unser Erkenntnisstreben so einfach wäre, und wir über ein derartiges Hilfsmittel verfügten. Doch andererseits wäre es sehr überraschend, wenn wir ein einfaches Kriterium hätten, das den möglichen Erfolg oder Misserfolg bestimmter Theorien schon früh und in letztlich *hellseherischer* Weise vorhersagen könnte, denn viele (wissenschaftliche) Theorien wurden zwischenzeitlich als unwissenschaftlich gebrandmarkt und dann auch wieder ernsthafter betrachtet. So wurde in neuerer Zeit die kalte Fusion zunächst ernsthaft untersucht, dann als recht abwegig begraben und in *Nature* für tot erklärt und inzwischen wird sie wieder etwas ernster genommen. Vermutlich wird sie sich als Fehlentwicklung herausstellen, aber wir haben keine einfachen Entscheidungshilfen, die den Erfolg einer solchen Theorie vor allem Experimentieren und anderem Räsonieren vorhersagen können. Wir müssen uns wohl damit abfinden, kein einfaches Kriterium ausmachen zu können, das uns schon zu Beginn einer Debatte sicher erkennen lässt, was aus einer Theorie einmal werden wird und welche Hypothese sich wohl nie zu einer sinnvollen wissenschaftlichen Theorie entwickeln wird. Das gilt zumindest für die etwas spannenderen Fälle.

Natürlich wird niemand die Theorien über den Klabautermann als ernsthafte wissenschaftliche Theorien untersuchen, aber schon im Falle der Homöopathie oder Astrologie muss man sich wenigstens die Zeit nehmen, inhaltlich darauf einzugehen, was gegen ihre Annahmen spricht. Die Astrologie behauptet etwa, dass ein Einfluss der Stellung der Gestirne bei der Geburt eines Menschen auf seinen Charakter vorliegt, den sie auszunutzen hofft und der zu korrekten Vorhersagen führen soll. Hier wird man nachfragen, wie dieser Einfluss vonstattengehen soll, welches also die Mechanismen sind, durch den die Gestirne diesen Einfluss haben. Weiterhin kann man die Vorhersagen empirisch überprüfen und zusätzlich alternative Erklärungen für den Charakter eines Menschen geben – etwa über seine genetische Ausstattung und Umwelteinflüsse auf

seinem Lebensweg. Das ist in Einzelfällen auch alles schon geschehen und belegt, dass die Astrologie schlecht abschneidet (vgl. Thagard 1978, Ivo Ponocny & Elisabeth Ponocny-Seliger 2009). Natürlich hat sie viele Tricks zur Verfügung, um sich vor einfachen Falsifikationen zu schützen, wie übrigens in ähnlicher Weise echte wissenschaftliche Theorien, aber damit wird sie in der Regel zugleich langweilig und das können wir ebenso aufdecken. Sie kann etwa in ihren Vorhersagen wie den Horoskopen so vage bleiben, dass genau genommen nichts Konkretes mehr daraus folgt, oder sich in Horoskopen zu so tollen Tipps versteigen wie: »Sie sollten sich heute nicht überanstrengen.« Na klar, wann sollte man das schon.

Dass es nach einer solchen negativ verlaufenen Debatte immer noch Anhänger der Astrologie geben wird, ist klar, aber ein Kriterium für Wissenschaftlichkeit könnte das auch nicht verhindern. Die Astrologen könnten kontern: »Natürlich ist das ein schönes Kriterium für Wissenschaftlichkeit (wenn sie es denn überhaupt anerkennen würden), aber die wirklich tiefen Einsichten werden dann eben außerhalb der Wissenschaft gewonnen. Für viele praktische Fragen wird man sich gerade den Erkenntnissen der Astrologie zuwenden müssen.« Es bleibt also leider nur der ständige Kleinkrieg gegen die Unvernunft und kein Kriterium könnte so überzeugend sein, dass der mit einem Mal beendet würde. Außerdem ist es vielleicht sogar noch nicht einmal verkehrt, dass man über bestimmte (einfache) Zusammenhänge noch einmal nachdenkt, um sich darüber klar zu werden, was tatsächlich gegen die Homöopathie oder die Astrologie spricht und was die Wissenschaften demgegenüber anzubieten haben. Das ist nicht immer leicht zu beantworten, doch die Wissenschaft sollte sich solchen externen Herausforderungen stellen, um die eigenen Qualitätsbehauptungen zu überprüfen.

Allerdings müssen wir dazu unser Bild von der Wissenschaft etwas zurechtrücken. Erstens ist sie kein ganz so homogenes Unternehmen und zweites gibt es keine scharfe Trennungslinie zwischen wissenschaftlichem und Alltagswissen oder auch pseudowissenschaftlichen Hypothesen. Wir haben es eher mit einem kontinuierlichen Übergang zwischen gut begründeten Aussagen und weniger gut begründeten Aussagen zu tun, bis hin zu Aussagen, für die wir momentan mehr Gegengründe als

Gründe haben. Ab wann man hier von »unwissenschaftlich« sprechen kann, ist nicht genau bestimmt.

Selbst in den Methoden finden wir keine so scharfen Abgrenzungen. Wissenschaftstheoretiker denken dazu, dass auch Wissenschaftler Menschen sind (s.o.). Das soll heißen, dass sie als Wissenschaftler ihre Erkenntnisse nicht gleich auf völlig andere Art und Weise begründen als der »normale« Mensch. Die Verfahren sind im Kern dieselben und werden nur noch präzisiert etwa durch geeignete statistische Methoden. Allerdings können sich die wissenschaftlichen Modelle der Welt mit Hilfe der Mathematik ein gutes Stück von anschaulichen Alltagsmodellen entfernen. Trotzdem gibt es hier auch innerhalb der empirischen Wissenschaften ein Kontinuum von Methoden von der Physik über die Medizin, Ökonomie, Soziologie bis hin zu den Geschichtswissenschaften, die z.B. weit weniger mathematisierte Methoden einsetzen, wo wir aber nichtsdestoweniger zwischen besser und schlechter begründeten Hypothesen unterscheiden können.

Was können wir also tun? Das Beste, was uns zu tun übrig bleibt, ist nun, *Indikatoren* für wissenschaftliches Arbeiten anzugeben bzw. bestimmte *Spielregeln der Wissenschaft* explizit zu machen, um zumindest Hinweise zu finden, wann wir auf eine mehr oder weniger wissenschaftliche Tätigkeit stoßen. Startpunkt sind die beiden Ziele, die wir schon für die Wissenschaften ausgemacht haben, nämlich das Streben nach Wissen also insbesondere Wahrheit, aber auch das Streben nach informativen und insbesondere *erklärungsstarken Theorien*, dem sich etwa Astrologen mit vagen Aussagen zu entziehen suchen.

Unsere Debatte zum Wissensbegriff hat deutlich gemacht, dass wir auch nach möglichst *unanfechtbaren Begründungen* für unsere Theorien suchen, um das erste Ziel der Wissenschaften zu erreichen. Zugleich müssen wir uns für empirische Theorien immer wieder fragen, welchen Gehalt sie tatsächlich aufweisen, wann sie also echte Erklärungen liefern und wann vielleicht nur Pseudoerklärungen. Was echte Erklärungen auszeichnet, wird später noch erläutert. Ausführlicher findet sich das in Bartelborth (2007). Zunächst mag aber unser intuitives Verständnis dafür ausreichen, das Ziel zu verstehen.

Es war vor allem Popper (1984), der immer darauf hingewiesen hat, dass die *Falisfizierbarkeit* einer Theorie das wichtigste Kriterium für

ihre Wissenschaftlichkeit sei. Je höher der Falsifizierbarkeitsgrad bzw. je höher der empirische Gehalt und die Vorhersagekraft einer Theorie umso besser für sie. Damit bildete er einen wichtigen Gegenpol zu den logischen Empiristen, die praktisch nur auf die Begründung der Theorien abzielten und zeitweise sogar hofften, die wissenschaftliche Methodologie auf eine *induktive Logik* zu reduzieren, die den Grad der Begründung (»confirmation«) c(H,E) einer Hypothese H durch die Daten E (für »evidence«) durch bestimmte logische Wahrscheinlichkeiten angibt. Auf die Probleme dieses Verfahrens werde ich im 5. Kapitel zu sprechen kommen, aber zunächst einmal ist klar, dass der Aspekt des Gehalts einer Theorie damit verlorengeht. Je höher der Gehalt einer Theorie, umso unwahrscheinlicher ist sie im Normalfall, insbesondere gilt: Wenn H aus der stärkeren Theorie H' logisch folgt (H'⇒H), so ist P(H') ≤ P(H). In gewisser Weise ergeben sich die weiteren Spielregeln aus diesen beiden oberen Zielen, wobei hier oft das erste Ziel im Vordergrund der Überlegungen stehen wird.

Als Rahmen, innerhalb dessen wir diese Ziele verfolgen, nennt Schurz (2006, 26 ff.) bestimmte *minimale erkenntnistheoretische Annahmen*, die allen empirischen Wissenschaftsdisziplinen gemeinsam seien, und nennt als Erstes die Forderung des *minimalen Realismus*. Danach gehen wir davon aus, dass eine Realität existiert, die von den Annahmen des Erkenntnissubjekts unabhängige Eigenschaften aufweist, die es zu erforschen gilt. Das schließt für diese Eigenschaften auch einen *korrespondenztheoretischen Wahrheitsbegriff* mit ein.

Diese Annahme ist erforderlich, da wir sonst nicht wirklich von Erklärungen sprechen können. Möchte ich z.B. das Verhalten eines Menschen erklären und beschreibe es so: »Er hat sich verhalten, *als ob* er von einem Dämon beherrscht worden wäre«, dann hat das keine Erklärungskraft. Jedenfalls dann nicht, wenn ich davon ausgehe, dass es diesen Dämon nicht wirklich gab. Mit solchen »Klabautermann-Erklärungen« können wir etwa unser Erstaunen zum Ausdruck bringen, und sagen, dass dieses Verhalten für den Betreffenden nicht typisch ist und Ähnliches mehr, aber *warum* er sich so seltsam verhalten hat, wird für uns nicht aufgeklärt (vgl. Kap. 4.7). Das deutet schon das »als-ob« in unserem Satz an. Nehmen wir entscheidende Bestandteile unserer Erklärung nicht ontologisch ernst, sondern betrachten sie nur als fiktive

Größen, die uns eine anschauliche Beschreibung erlauben, so liegt noch keine Erklärung vor.

Für eine Erklärung muss ich auf tatsächliche gesetzesartige Zusammenhänge in der Welt verweisen, die sich etwa aus den Eigenschaften bestimmter Dispositionen ergeben, während die »Als-ob«-Antwort die Frage offen lässt, was die wirklichen Ursachen und Motive seines Verhaltens waren. Auf einen gewissen Realismus sind Wissenschaftler daher festgelegt, wollen sie ihr zweites Ziel überhaupt einlösen können. Ein Astrologe müsste also darauf bestehen, dass es einen kausalen Einfluss der Gestirne auf unsere genetische Ausstattung bei unserer Geburt tatsächlich gibt.

Leider wird selbst diese naheliegende Annahme oder Spielregel für die Wissenschaft inzwischen von einigen Wissenschaftlern bestritten. Die sogenannten *sozialen Konstruktivisten* bestreiten z.T. die Existenz einer von uns unabhängigen Welt und halten stattdessen alles für sozial konstruiert. Hier ist allerdings nicht der Ort, um das weiter zu diskutieren. Es zeigt aber zumindest, dass die Forderung des minimalen Realismus nicht völlig trivial ist.

Die *Erklärungsstärke* wissenschaftlicher Theorien war auch ein wesentlicher Punkt in dem Gerichtsverfahren, das 1983 in Arkansas um die Frage geführt wurde, ob der *Kreationismus* eine Wissenschaft sei, die dann gleichberechtigt neben der biologischen Evolutionslehre an der Schule zu unterrichten sei (vgl. Bird 2002, 3 ff.). In den USA gab es in den 90er Jahren in einigen Staaten weitere Gesetze, die das vorsahen, und eine Mehrheit der Bevölkerung sprach und spricht sich ebenfalls dafür aus. Der Richter von 1983 William R. Overtone sah das anders und stützte sich in seiner Urteilsbegründung auf die folgenden Grundsätze, die eine Theorie erfüllen müsse (wobei er sich auf die Wissenschaftstheorie berufen konnte), um als wissenschaftlich gelten zu können:

Grundsätze für Wissenschaftlichkeit bei Gericht 1983

(a) It is guided by natural law.
(b) It has to be explanatory by reference to natural law.
(c) It is testable against the empirical world.
(d) Its conclusions are tentative, i.e. are not necessarily the final word.
(e) It is falsifiable.

Die Bedingungen (a) und (b) zielen auf die Erklärungskraft der wissenschaftlichen Theorien ab. In (c) und (e) geht es darum, dass wissenschaftliche Theorien einen empirischen Gehalt haben müssen und in (d) wird gefordert, dass sie undogmatisch vertreten werden. In diesem Rahmen liegen immer wieder die Anforderungen an Wissenschaftlichkeit, die nur von verschiedenen Autoren etwas unterschiedlich formuliert werden. Wir können das in drei größere Komplexe einordnen:

Indikatoren für Wissenschaftlichkeit
1. Gute Begründungen: Es sind spezielle (u.a. intersubjektiv überprüfbare) Methoden bei der Erstellung und Begründung wissenschaftlicher Theorien einzusetzen.
2. Gehaltvolle empirische Theorien: Wissenschaftliche Theorien müssen eine bestimmte Leistungsfähigkeit und Erklärungskraft aufweisen, indem sie nomische Muster in unserer Welt aufdecken.
3. Verhalten der Wissenschaftler: Wissenschaftler müssen eine undogmatische Einstellung gegenüber ihren Theorien (z.B. Verhalten gegenüber Einwänden) einnehmen.

Für diese drei Bereiche können die Indikatoren nun jeweils weiter spezifiziert werden. Den ersten Aspekt werden wir in diesem Buch weiter verfolgen, aber wir können schon bestimmte allgemeine Anforderungen an die Begründungen formulieren. Zum Spiel Wissenschaft gehört es dazu, dass die Begründungen so gestaltet sind, dass sie im Prinzip von anderen überprüft und bewertet werden können. Wenn also etwa jemand persönliche Erleuchtungserlebnisse hat, in denen ihm Gott erschienen ist und ihm besondere Einsichten E vermittelt hat, so mag das für ihn vielleicht einen besonders guten Grund bieten, an E zu glauben, wissenschaftlich ist das aber nur dann eine zulässige Begründung, wenn diese Einsichten auch *intersubjektiv* zugänglich sind. Wenn sie also von anderen auf eine bestimmte Weise reproduziert werden könnten. Ist das nicht der Fall, zählen sie nicht im Spiel Wissenschaft, selbst wenn sie uns persönlich als gute Informationsquellen erscheinen.

Sind sie *reproduzierbar*, bleibt allerdings immer noch die Frage, welches Gewicht ihnen dann zukäme, und das ist von weiteren Faktoren abhängig. Die Begründungsverfahren müssen selbst einer kritischen Bewertung unterzogen und begründet werden, z.B. durch bestimmte

erkenntnistheoretische Überlegungen und durch ihre Erfolgsquote in ihren Anwendungsfällen. So werden wir auch statistische Verfahren auf ihre erkenntnistheoretische Plausibilität und im Hinblick auf die Frage überprüfen müssen, ob ihre Ergebnisse in anderen Anwendungen sich tatsächlich als intuitiv brauchbar erweisen. Die Begründungsverfahren müssen also selbst intersubjektiv zugänglich begründet werden. Wir müssen Gründe dafür haben, dass sie die Ziele der Wissenschaft (also z.B. Wissen zu generieren) tatsächlich befördern und schließlich hin zu unanfechtbarem Wissen führen. Die im Folgenden rekonstruierten Überlegungen stellen einen Teil der langen und oft heftig geführten Debatte darüber dar, welches denn die richtigen Begründungsverfahren sind. Die Uneinigkeit belegt schon, dass hier weiterer Diskussions- und Begründungsbedarf besteht.

Eine gewisse Einigkeit besteht in vielen Kreisen von Praktikern, dass etwa *Signifikanztests* das geeignete Mittel der Auswahl von Theorien sind. Doch unter den Methodologen, die sich mit diesem Verfahren selbst auseinandersetzen, besteht ebensolche Einigkeit, dass das Verfahren nicht das leisten kann, was man sich oft von ihm verspricht. Darauf werden wir in Kapitel 6 genauer eingehen. Jedenfalls kann selbst Einigkeit eine Begründung nicht ersetzen. Wir müssen uns immer fragen, was die Verfahren genau beweisen und wo ihre Grenzen sind. Leider gibt es nicht ein perfektes Induktionsverfahren für alle Anwendungsbereiche, aber zumindest eine *Rahmentheorie*, nämlich den *Schluss auf die beste Erklärung*, in den wir die anderen Verfahren einbetten und mit deren Hilfe wir sie bewerten können.

Den zweiten Aspekt der Wissenschaftlichkeit haben wir schon ein Stück weit beschrieben. Es geht u.a. um die Erklärungskraft und Vorhersagekapazität unserer Theorien. Manche pseudowissenschaftliche Theorie versucht durch die Vagheit ihrer AUssagen einer genauen Überprüfung oder Widerlegung zu entkommen. Im größeren Rahmen ist aber auch die Frage wesentlich, wann eine Disziplin *genuine Fortschritte* macht, und das ist nicht so leicht zu explizieren und ebenfalls nicht leicht zu ermitteln. Wir haben bereits den Vorschlag Poppers untersucht, wonach Fortschritt darin besteht, dass wir uns der Wahrheit nähern sollten. Doch zunächst ist es schon nicht ganz leicht, dieser recht intuitiven Vorstellung eine präzise Form zu geben. Poppers eigener Vorschlag, sich

jeweils an den Inklusionsbeziehungen der Mengen der wahren und der falschen Aussagen der Theorien zu orientieren, war leider unbrauchbar (vgl. Oddie 2014). Allerdings gibt es bessere Vorschläge, von denen wir bereits einen kurz diskutiert haben. Das nächste Problem ist jedoch, dass wir mit diesem Fortschrittskriterium keine für uns *einfach anwendbaren* Kriterien finden, ob tatsächlich ein Fortschritt vorliegt.

Da ist es um die *Vorhersagekraft* von Theorien in puncto Zugänglichkeit schon besser bestellt. Für die können wir empirische Belege sammeln. Eine andere Idee sieht den Fortschritt in einer *zunehmenden Vereinheitlichung* bzw. Systematisierung oder der Kohärenz unseres Wissens. Auch die scheint aber vor allem für die Erklärungskraft von Bedeutung zu sein. Dazu gehört u.a. (und das betrifft ebenso den ersten Aspekt), dass es wenige Anomalien in unserem Paradigma geben sollte.

Außerdem sollten wir eine gewisse *dynamische Stabilität* in unserem Wissen verzeichnen können. Würden wir kurzfristig immer wieder zwischen konkurrierenden Theorien hin und her schwanken (also z.B. zwischen einer Wellentheorie des Lichts und einer Partikeltheorie), weil die Daten mal für die eine und mal für die andere Theorie sprechen, so könnten wir kaum behaupten, entscheidende Fortschritte gemacht zu haben.

Der dritte Aspekt der Wissenschaftlichkeit betrifft noch einen ganz anderen eher sozialen Bereich des Unternehmens Wissenschaft. Es geht nicht so sehr um die Theorien selbst, ihre Leistungsfähigkeit und unsere Begründungen dafür, sondern es geht uns um das *Verhalten der Wissenschaftler*. Welchen epistemischen Status geben sie ihren Theorien und wie verfahren sie mit ihren Theorien? Hier geht es insbesondere um die Frage, ob wir unsere Theorien auf *dogmatische Weise* verteidigen. Jeder Wissenschaftler ist ein Stück weit immer der Anwalt seiner eigenen Theorien und versucht sie im besten Licht erscheinen zu lassen. Schließlich möchte er dafür die Anerkennung seiner Kollegen und Forschungsgelder bestimmter Institutionen erhalten. Aber das darf eine bestimmte Grenze nicht überschreiten. Er darf weder Daten fälschen noch sie weglassen, wenn sie nicht zu seiner Theorie passen, und ebenfalls nicht dogmatisch auf Einwände reagieren, d.h., diese schlicht nicht mehr zur Kenntnis nehmen und völlig unbeirrt an

der eigenen Theorie festhalten, ganz gleich, wie gut die Einwände sein mögen.

Das betrifft ebenso die Methoden selbst, die er reflektiert und undogmatisch anzuwenden hat. So gehören zur Wissenschaft Kriterien für die Methoden und die darauf aufbauende Qualität ihrer Produkte, aber ebenfalls Verhaltensnormen für die Wissenschaftler. Sie müssen ihre Ergebnisse veröffentlichen und damit einer öffentlichen Debatte aussetzen. Das findet wiederum Eingang in die üblichen Systeme der Begutachtung und der wissenschaftlichen Debatte.

Allerdings können solche Systeme nachgebildet werden. So gibt es auch für den Kreationismus wissenschaftliche(?) Zeitschriften mit einer Begutachtung durch Reviewer. In den Zeitschriften werden dann z.B. Theorien zu der Frage diskutiert, wann Gott die Insekten oder die Viren geschaffen habe, insofern sie durch die Bibel noch nicht explizit beantwortet werden.

Außerdem sollte man sich von den Begutachtungsverfahren auch nicht zu viel versprechen. Die empirischen Forschungen zum Begutachtungswesen kommen zu vielen ernüchternden Einsichten. Arbeiten, die vor drei Jahren noch akzeptiert wurden, werden heute von denselben Zeitschriften abgelehnt; Reviewer akzeptieren vor allem die Artikel, die ihrer eigenen Meinung entsprechen u.v.m. (vgl. Fröhlich 2003). So werden schwere Fehler in Artikeln eher durch die Leser als durch die Gutachter entdeckt, die ihrerseits verständlicherweise meist Besseres zu tun haben, als ihre ganze Aufmerksamkeit den möglichen Artikeln anderer Leute zu widmen.

Damit haben wir zwar keine scharfen Abgrenzungskriterien für die Wissenschaft gegenüber den Pseudowissenschaften erhalten, aber zumindest noch eine Reihe von *Indikatoren* für Wissenschaftlichkeit. Je mehr von ihnen erfüllt sind und umso besser sie erfüllt werden, umso eher haben wir es mit einer Wissenschaft zu tun. Letztlich haben wir es mit einer Abstufung von schlecht begründeten und uniformativen Behauptungen bis hin zu gut begründeten Naturgesetzen zu tun, und was davon als wissenschaftlich anzuerkennen ist, hängt auch von den Standards und Möglichkeiten in einzelnen Wissenschaftsdisziplinen ab. Der Wunsch nach einem »hellseherischen« Kriterium, das gleich zu Beginn der Untersuchung einer Behauptung bereits angibt, ob sich

daraus einmal eine gut begründete Behauptung ergeben kann oder eben nicht, kann natürlich nicht erfüllt werden.

Ein anderer Ansatz kommt von der Gegenseite her und versucht, spezielle Merkmale der *Pseudowissenschaften* für eine Abgrenzung anzugeben. Natürlich sind da zunächst einfach die Negationen der obigen Kriterien zu nennen, wir können das aber vielleicht zusätzlich durch genauere *Spezifika pseudowissenschaftlichen Vorgehens* beschreiben wie die Überbetonung von *Analogien* und weiteren *Assoziationen*. Jedenfalls schreibt etwa Paul Thagard (1978, 227f.):

> We can now propose the following principle of demarcation: A theory or discipline which purports to be scientific is pseudoscientific if and only if: it has been less progressive than alternative theories over a long period of time, and faces many unsolved problems; but the community of practitioners makes little attempt to develop the theory towards solutions of the problems, shows no concern for attempts to evaluate the theory in relation to others, and is selective in considering confirmations and non confirmations.

Man erkennt hier vor allem die Verletzungen der obigen positiven Regeln wieder, die Paul Thagard im Bereich der Astrologie vermutet.

Es ist gut ersichtlich, wie die Spielregeln den genannten Zielen der Wissenschaft dienen. Für die Indikatoren der Bereiche (1) und (2) ist das offensichtlich, aber auch der Verhaltenskodex für Wissenschaftler soll letztlich diesen Zielen dienen. Doch wie effizient werden diese Regeln tatsächlich eingehalten? Das ist ein Thema für empirische Studien zu diesem Thema. In Zankl (2003) werden jedenfalls etliche Beispiele vorgestellt, in denen die Wissenschaftler ganz bewusst gegen die Regeln verstoßen haben.

2.5 Fälschungen als Unterminierer

Die in der Regel ungeschriebenen Gesetze der Wissenschaft oder *Spielregeln*, wie ich sie nenne, werden in der Praxis natürlich nicht immer befolgt; genauso wenig wie die sozialen Spielregeln für den Umgang der

Menschen untereinander. Das bedeutet jedoch nicht, dass sie nicht wirklich gelten oder anerkannt werden, sondern oft wissen die Regelverletzer sehr genau, dass sie unseriös agieren. Das erkennt man u.a. daran, dass sie versuchen, ihre Regelverstöße zu kaschieren oder sogar mit einem gewissen Aufwand ganz geheim zu halten.

Einen großen Bereich von Problemen für Wissensansprüche finden wir jedenfalls in den *Fälschungen* in der Wissenschaft. Die sind in der tatsächlichen Wissenschaft weiterverbreitet, als wir es uns normalerweise eingestehen wollen (vgl. Zankl 2003 dazu). Dabei treten fast immer Unterminierer der bisherigen Begründungen auf. Betrachten wir das kurz an einem fiktiven Fall: Nehmen wir an, Prof. Sorgfalt möchte die Theorie H überprüfen, dass *Grüntee bei Sportlern die Abwehrkräfte stärkt*, wenn er wenigstens über ein Jahr lang regelmäßig getrunken wird. Bei der Durchführung der Untersuchungen wird er unterstützt durch seinen Assistenten Dr. Sorglos. Außerdem sei es tatsächlich so, dass die Theorie H stimmt. Was kann nun alles schiefgehen, so dass Prof. Sorgfalt nicht zu H-Wissen gelangt? (»H-Wissen« sei hier die Kurzform, für den Ausdruck: »Wissen, dass H« bzw. »Wissen, dass H der Fall ist.«)

Nehmen wir an, sein Assistent Dr. Sorglos möchte gerne in einem möglichst erfolgreichen Projekt teilnehmen und dem Professor ebenso gerne zugleich einen Gefallen tun. Der ist schließlich ungenießbar, wenn seine Studien nicht so laufen, wie er sich das wünscht. Also fälscht Sorglos bestimmte Resultate, indem er etwa die Angaben über die Anzahl der T-Zellen im Blut der Sportler im Sinne der Hypothese H »schönt«. Da er fest von H überzeugt ist, scheint ihm das auch eine lässliche Sünde zu sein, verhilft sie doch nur dem gesunden Grüntee zum Durchbruch. Die Studie enthält eine Versuchsgruppe E der Grüntee-Trinker und eine Kontrollgruppe K der Grüntee-Abstinenzler (die einen Placebo-Grüntee erhalten?). Leider lassen sich nach einem Jahr keine signifikanten Unterschiede in den Blutwerten der beiden Gruppen festmachen. Um die positiven Wirkungen des Grüntees doch besser sichtbar zu machen, »korrigiert« Dr. Sorglos die Werte der Blutproben aus E etwas nach oben, so dass sich das gewünschte Resultat einstellt.

Nehmen wir an, Prof. Sorgfalt kommt nicht auf die Idee, dass so etwas passiert sein könnte – er hält jedenfalls bisher Dr. Sorglos für absolut vertrauenswürdig – und glaubt aufgrund dieser Resultate nun fest an

H. Dazu hat er sicher gute Gründe, sagen doch die Auswertungen der umfangreichen Studie seiner Meinung nach, dass der Abwehrstatus der Grüntee-Trinker deutlich besser ist, als der aus der Kontrollgruppe. Und nehmen wir außerdem an, H sei wahr. Trotzdem können wir Prof. Sorgfalt kein Wissen zubilligen, denn die Fälschungen seines Assistenten *unterminieren* die Gründe des Professors, selbst wenn er sie nicht kennt. Die Aussage über die Fälschungen von Dr. Sorglos bieten zusammen mit den Ergebnissen der Studie keine Begründung mehr für die Annahme H und gerade intuitiv scheint der Fall ganz einfach zu sein: Wir können Prof. Sorgfalt kein H-Wissen zubilligen. Jedenfalls, wenn er keine anderen überzeugenden Gründe für H hat, als die, die ihm sein Assistent geliefert hat, wovon wir hier einmal ausgehen. Sollten sich also Fehler in der Datenerhebung finden lassen, unterminieren sie unsere Wissensansprüche selbst dann, wenn wir guten Glaubens auf die Daten vertrauen und unsere Überzeugungen wahr sind.

Was kann sonst noch zu einer Unterminierung führen? In der wissenschaftlichen Praxis sind viele mögliche Unterminierer bekannt und ihre Kenntnis und Vermeidung gehört zur wissenschaftlichen Methodologie, über die es ganze Bücher gibt (z.B. Bortz & Döring 2006). In unserem Beispiel lassen sich eine ganze Menge typischer Unterminierer aufzeigen. Einige Beispiele für solche Unterminierer für verschiedene Ergebnisse des Experiments sind etwa die folgenden: Die Teilnehmer einer Studie könnten eine *mangelnde* »*compliance*« an den Tag gelegt haben. Vielleicht sind die Ergebnisse der Grünteetrinker nur deshalb noch genauso schlecht wie die der Nichttrinker, weil die Teetrinker zwar weiterhin behauptet haben, sie hätten sich immer an die Vorgaben gehalten (5 Tassen Tee pro Tag), aber sie hatten dann doch bald keine Lust mehr dazu und haben das Teetrinken deutlich reduziert. Oder: Bei der Auswahl der Gruppe E haben sich insbesondere die Versuchspersonen gemeldet und sind nach vorne gedrängt, die schon gesundheitlich angeschlagen waren und daher von vornherein das schwächere Immunsystem aufwiesen. Diese Personen waren daher auch besonders daran interessiert, innerhalb der Versuchsgruppe an dem Experiment teilzunehmen.

Gegen das letzte Problem versuchen die Versuchsleiter in der Regel die *Randomisierung* einzusetzen, wonach die Versuchspersonen per

Zufallsauswahl den beiden Gruppen zugeteilt werden. Aber selbst die *Besprechungen* mit den Versuchsteilnehmern können Auswirkungen auf die Ergebnisse haben, sowie die begleitenden Kontakte mit den Teilnehmern im Sinne des Rosenthal-Effekts sowie des Placebo-Effekts. Beim Rosenthal-Effekt hat die eventuell etwas andere (aber möglicherweise unbewusst andere) Behandlung der Personen aus Gruppe E und K eine Auswirkung auf das Ergebnis, während der Placebo-Effekt auf der Kenntnis der Teilnehmer beruht, ob sie zu Gruppe E oder K gehören. Beide Effekte sind empirisch sehr gut belegt und keinesfalls vernachlässigbar. Dem versucht man meist durch eine *doppelte Verblindung* (der durchführenden Personen sowie der Versuchsteilnehmer) zu begegnen, aber schon in diesem einfachen Beispiel zeigt sich, dass das nicht gelingen dürfte. Die Personen in K müssten einen glaubwürdigen Placebo-Tee erhalten, doch der dürfte sich im Normalfall schon durch den Geschmack und das Aussehen von echtem Grüntee unterscheiden. Es gibt viele weitere Probleme der Datenerhebung, die hier denkbar sind, und die typischerweise Unterminierer darstellen.

Was sind mögliche *Gegengründe* gegen H? Andere Experimente zu den Wirkungen des Grüntees könnten zu anderen Ergebnissen gekommen sein. Es ließ sich vielleicht in diesen anderen Experimenten keine Wirkung feststellen. Wie können wir das miteinander verrechnen? Da in der Praxis hier oft Signifikanztests durchgeführt werden, gibt es kein einfaches Verrechnungsverfahren. Man spricht dann meist von einer Metaanalyse, aber die ist keineswegs ein klares Verfahren. Um die Gegengründe zunächst abzuschwächen, könnten wir z.B. auf entsprechende Unterminierer der anderen Experimente hinweisen, sollten die nicht sauber durchgeführt worden sein. Doch wie stark das wiederum zu Buche schlägt, ist auch nicht klar. Besser stehen hier vor allem die Bayesianer da (aber auch die Likelihoodisten), die zumindest im Prinzip eine solche Verrechnung durchführen können. Unsere Wissenskonzeption hilft uns an dieser Stelle zumindest, die aufgestellten Spielregeln als sinnvoll zu erkennen und die typischen Regelverstöße einzuordnen.

Letztlich suchen wir nach Theorien, die gut begründet sind, zu deren Begründungen es keine relevanten Unterminierer und zu denen es keine starken Gegengründe gibt. Außerdem werden wir sehen, dass noch ein speziell *komparatives Element* hinzukommt: Es muss uns gelingen,

alle wichtigen *Alternativhypothesen* zu eliminieren, denn die stellen wiederum wichtige Unterminierer in der wissenschaftlichen Praxis dar. Erklärt nämlich eine Hypothese H unsere Daten E in zufriedenstellender Weise, aber ebenso die Hypothese H', die mit H inkompatibel ist, so wird die Stützung von H durch E durch das Vorliegen einer Konkurrenzerklärung deutlich geschwächt. Das zeigt das Schema des Schlusses auf die beste Erklärung, das wir in Kapitel 4 diskutieren werden. Das Problem ist ebenfalls unter dem Stichwort der *Unterbestimmtheit der Theorien* (gegeben unsere Daten) bekannt und wurde insbesondere von Thomas Kuhn (1976) in seinen wissenschaftshistorischen Fallstudien immer hervorgehoben, um zu zeigen, dass unsere Theorien keineswegs immer so gut begründet sind, wie wir uns das denken.

2.6 Probleme der Datengewinnung

Ein Bereich, den wir auch im Folgenden etwas stiefmütterlich behandeln werden, ist der der Datenerhebung und ihrer Resultate selbst. Hier ist aber zunächst zwischen *Beobachtungsaussagen* bzw. echten *Rohdaten*, wie ich sie nennen möchte, und den *interpretierten Daten* zu unterscheiden, die wir dann mit unseren Theorien konfrontieren können. Um von den Rohdaten zu den relevanten Daten zu gelangen, sind meist schon erste induktive Schlüsse oder Interpretationen erforderlich, die also bereits in das Thema des induktiven Schlussfolgerns fallen. Möchte ich etwa elektromagnetische Theorien überprüfen, werde ich diese normalerweise nicht direkt mit bestimmten Beobachtungen wie etwa einer Zeigerstellung auf einem Messgerät vergleichen, sondern werde erst diese Zeigerstellung auf der Drei deuten als: *Es fließen 3 Milliampere in diesem Stromkreis*. Oder: Möchte ich die Hypothese testen, dass eine (vermeintlich) ungerechte Behandlung zu großem Ärger bei den Betroffenen führt, werde ich u.a. ein bestimmtes Verhalten von Probanden als Ausdruck des Ärgers deuten müssen, etwa dass sie energisch und lautstark protestieren o.ä. Dabei können wir es wieder mit Unterminierern zu tun haben, wie etwa ungeeichten Messgeräten etc. Wir befinden uns damit bereits mitten im Bereich induktiven Schlussfolgerns.

Wir können einige der genannten und neue Probleme ebenfalls an konkreten Beispielen aus der Wissenschaft erläutern. Das bringt uns schließlich zu einigen Spielregeln für die Wissenschaft, die helfen sollen, diese Fehler zu vermeiden und so im Hintergrund des wissenschaftlichen Wissensbegriffs stehen. Oft haben wir bereits auf der Ebene der Datenerhebung schon mit erheblichen erkenntnistheoretischen Schwierigkeiten zu kämpfen, die sich im Einzelfall als Unterminierer herausstellen können. Ein einfaches Beispiel finden wir in *ungeeichten Messgeräten*, aber ich denke hier mehr an wirklich problematische Fälle der Datenerhebung. Ein Beispiel ist besonders durch seine politische Brisanz sowie ein bestimmtes Fehlverhalten der beteiligten Akteure, das bekannt wurde, in letzter Zeit in den Blickpunkt der Öffentlichkeit gerückt.

Gerade der aktuelle Fall von »Climategate« ist ein schönes Beispiel dafür, wie hoch die Anforderungen an wissenschaftliches Wissen sind und wie schnell bestimmte Theorien den Status von Wissen wieder verlieren können, weil bestimmte Unterminierer auftauchen. Daten zu erheben wie die Durchschnittstemperatur der vergangenen Jahre, aber auch die der Gegenwart sind schließlich bereits recht hypothetische Unternehmungen. Sie müssen induktiv aus grundlegenderen Daten erschlossen werden, wie etwa den Einlagerungen in Eiskernen oder Bäumen. Dass die Messung der Durchschnittstemperatur schon für die Gegenwart nicht so einfach ist, kann man sich leicht überlegen: Man muss sich nur vor Augen halten, dass wir es in Deutschland im Laufe eines Jahres etwa mit Temperaturschwankungen von ca. 50° C zu tun haben, und daraus wollen wir Erhöhungen der Durchschnittstemperatur von 0,166° pro Jahrzehnt »ablesen«. Das jedenfalls erhalten wir in etwa, wenn wir dem britischen Klimaforscher Phil Jones glauben, der maßgeblich an der Ermittlung der sogenannten Golfschlägerkurve beteiligt war, die einen deutlichen und einzigartigen Anstieg in der neueren Zeit zeigt und entsprechende Prognosen für die Zukunft stützt.

Inzwischen sind einige potentielle und z.T. wohl auch aktuale Unterminierer zu dieser Kurve aufgetaucht. Dazu einige Hinweise (vgl. Spiegel 13/2010): Für die Klimageschichte sind wir u.a. auf Baumringe angewiesen. Dazu müssen aber möglichst viele alte Bäume ausgewertet werden, doch gerade von den besonders alten Bäumen, die älter als

500 Jahre sind, finden sich nicht mehr viele Exemplare. Dazu gibt es Hinweise, dass das Herausfiltern der Durchschnittstemperaturen selbst fehlerhaft war. Die Qualität der Rohdaten wird noch durch viele andere Phänomene gefährdet oder unterminiert. Die realen Temperaturmessungen leiden z.B. darunter, dass um die Mess-Stationen herum Häuser gebaut wurden, etwa weil sich die Städte ausbreiten, und das selbst wieder zu höheren Temperaturen an diesen Stellen führt. Oder die Art und Weise der Messung der Wassertemperaturen wurde umgestellt, was zu Temperatursprüngen führte.

All diese Fehlerquellen musste Jones aus seinen Daten herausrechnen, was in jedem Fall selbst wieder fehleranfällig ist. Leider hat er die Aufzeichnungen über dieses Herausrechnen gelöscht, so dass sich diese Schlüsse nun nicht mehr überprüfen lassen. Das britische Met Office kommt daher zu dem richtigen Schluss, dass alle Daten nun offengelegt und neu bewertet werden müssen, selbst wenn das drei Jahre dauern kann. Außerdem wurden noch andere Spielregeln der Wissenschaft wie eine gewisse Offenheit gegenüber Einwänden verletzt, so dass eine klare Unterminierung der Ergebnisse von Jones u.a. vorliegt.

Das bedeutet aber andererseits nicht, dass die Ergebnisse falsch seien. Wir haben bereits oben darauf hingewiesen, dass ein Unterminierer nichts über die Wahrheit der in Frage stehenden Aussage p (also hier der behaupteten Golfschlägerkurve bzw. der Klimaerwärmung) aussagt, sondern nur bestimmte Begründungen wegbrechen. Er spricht noch nicht dafür, dass die Ergebnisse falsch sind. Es gibt außerdem noch andere Hinweise für die Klimahypothese, die weiter Bestand haben, aber nichtsdestoweniger erkennt man deutlich, dass hier bestimmte Wissensansprüche neu überprüft werden müssen, da für wissenschaftliches Wissen entsprechend hohe Standards gelten. Das bedeutet zugleich, dass die Prognosen aufgrund der Klimahypothese nicht mehr als gesicherte Erkenntnisse gelten dürfen, sondern ihnen ebenso die Grundlage entzogen wurde, denn für solche Schlüsse sind wir eben auf wissenschaftliches Wissen angewiesen, was vermutlich nicht mehr vorliegt. Noch weniger dürfen wir natürlich schließen, es werde keine Klimaerwärmung geben. Dafür bieten die Unterminierer wie gesagt keinen Anlass.

Als zu Beginn der 1980er Jahre die Hypothese aufkam, dass Magengeschwüre womöglich durch Bakterien verursacht werden, hielt man ihr

Unterminierer und Gegengründe entgegen. Die Daten für die Theorie wurden angezweifelt. Die Bakterien, die in Proben aus den Mägen Verstorbener gefunden wurden, wurden als vermutliche Verunreinigungen der Proben bei der Entnahme gedeutet. Außerdem wies man darauf hin, dass im sauren Milieu des Magens keine Bakterien überleben würden (vgl. Thagard 1998, 1999). Ersteres war ein Unterminierer, Zweiteres ein Gegengrund. Man glaubte so, die neue Bakterientheorie schnell wieder ad acta legen zu können, aber wie wir inzwischen wissen, behielt die Theorie schließlich doch die Oberhand. Jedenfalls sind uns derartige Argumentationen aus der Wissenschaft gut bekannt.

Unterminierer treten genauso dort in Erscheinung, wo wir Daten eine andere Interpretation geben möchten, so dass sie nicht mehr ihren ursprünglichen Zweck erfüllen können. Dazu werden meist bestimmte Hilfshypothesen ins Spiel gebracht. Beispiele dafür wie die Versuche, den *Ätherwind* zu messen, werden wir im Zusammenhang mit Popperschen Falsifikationen kennenlernen. Im Folgenden werden wir jedoch der Gewinnung von interpretierten Daten aus Rohdaten und der Sicherheit der Daten kaum noch Aufmerksamkeit schenken, sondern uns ganz dem Thema widmen, wie sich Theorien u.a. anhand von Daten begründen lassen. Das soll nicht bedeuten, dass es sich dabei nicht um eine wichtige Fragestellung handelt, sondern es ist nur eben nicht mein Thema hier.

Das gilt auch für die Frage, ob die Datenaussagen eine epistemische Basis unserer Erkenntnis im Sinne eines erkenntnistheoretischen Fundamentalismus darstellen. Ich neige eher zu einer moderaten empirischen Kohärenztheorie an dieser Stelle, werde das aber in diesem Buch nicht behandeln. Man siehe dazu Bartelborth (2015).

2.7 Wichtige Ergebnisse des Kapitels

In dem Kapitel werden zunächst die Zielvorstellungen für die empirischen Wissenschaften (und unsere weitere empirische Erkenntnis) genauer bestimmt. Bemerkenswert ist dabei, dass es uns zwar darauf ankommt, möglichst *wahre Einsichten* zu erzielen, dass dabei aber ein Aspekt gern übersehen wird, nämlich, dass wir auch nach ganz bestimmten Informationen über die Welt suchen. Wir suchen vor allem

nach *erklärungsstarken* Theorien, die auch *Vorhersagekraft* haben. Nur sie gestatten es zugleich, gut begründete Annahmen über die Zukunft zu entwickeln, auf die wir uns dann in unseren Entscheidungen stützen können. Man kann das auch so beschreiben, dass wir nicht nur einen kleinen Teil der Wahrheit erreichen möchten (oder schlimmer noch: nur wahre, aber triviale Einsichten), sondern dass uns um die *ganze* Wahrheit geht. Der möchten wir uns annähern. Daher sollen die Theorien, die wir in der Wissenschaft schließlich akzeptieren werden, nicht nur gut begründet sein, sondern wir müssen für das zweite Ziel auch immer ein gewisses (erhöhtes) Irrtumsrisiko eingehen. Die Ziele stehen in einem Spannungsverhältnis, das wir nicht leicht ausgleichen können.

Das erste Ziel lässt sich noch genauer beschreiben. Wir suchen nach *Wissen*. Bei Wissen handelt es sich aber nicht nur um eine zufällig wahre Überzeugung, sondern um eine sehr gut begründete Überzeugung. Dabei sollte die Begründung so gestaltet sein, dass sie etwa tatsächlich *sensitiv* gegenüber der Wahrheit ist. Da wir diese Idee nicht so weit explizieren können, wie da wünschenswert ist, habe ich eine ältere Idee wiederbelebt, die mir für wissenschaftliches Wissen besonders geeignet erscheint. Das Wissen muss de facto unanfechtbar sein bzw. stabil unter weiteren Wahrheiten. Es darf dafür keine (wahren) *Unterminierer* geben. Die gezielte Suche nach möglichen Unterminierern ist daher eine wichtige Aufgabe für die wissenschaftliche Praxis. Es scheint auch wichtig, darauf hinzuweisen, da wir aus vielen Untersuchungen von Psychologen wissen, dass Menschen eher dazu neigen, nach bestätigenden Instanzen für ihre Ansichten zu suchen, als nach möglichen Unterminierern. So gehört zum wissenschaftlichen Wissen dazu, dass wir nicht einfach bei einer Theorie stehen bleiben, auch wenn sie eine gute Erklärung für unsere Daten liefert, sondern ganz gezielt nach alternativen Erklärungen Ausschau halten.

Zu dieser objektiven Anforderung kommt allerdings noch eine etwas vagere und subjektivere, nämlich dass wir für (wissenschaftliches) Wissen auch unsere epistemischen Pflichten nicht grob verletzen dürfen. Diese letzte Bedingung wird auch gerne als Antwort genannt, wenn wir gefragt werden, warum wir den wissenschaftlichen Erkenntnissen so vertrauen. Sie wurden gemäß hohen Standards begründet und daher haben wir für

sie die begründete Hoffnung, dass es sich dabei tatsächlich um Wissen handelt.

3 Qualitative Induktionsverfahren

Die Grundidee der qualitativen Induktionsverfahren ist, dass wir prüfen müssen, ob das, was bestimmte Theorien über die Welt behaupten, auch tatsächlich zutrifft. Passen die Behauptungen unserer Theorien zu unseren Daten über die Welt, so scheint das die Theorien zu bestätigen, passen Theorien und Daten nicht zusammen, gilt die Theorie dagegen als geschwächt oder sogar widerlegt. Diese plausible Idee ist leider nicht ganz einfach zu präzisieren. Dabei stoßen wir auf eine Reihe von Problemen, die noch immer nicht vollständig gelöst sind. Den Anfang dieser Überlegungen soll der Falsifikationismus von Sir Karl Raimund Popper bilden. Popper war einer der prominentesten Wissenschaftstheoretiker und stand dem Projekt, ein Induktionsverfahren anzugeben, außerdem noch sehr skeptisch gegenüber.

3.1 Poppers Falsifikationismus

Popper (1984) hielt Humes skeptische Überlegungen gegenüber induktiven Schlüssen für stichhaltig (vgl. Kap. 1.7) und suchte daher nach einem neuen Weg, mit dem Induktionsproblem umzugehen. Seiner Ansicht nach gibt es keine induktive Bestätigung einer Hypothese, weil sich kein Induktionsverfahren zirkelfrei begründen lässt (vgl. Kap. 1.7). Wir dürfen uns daher nur auf deduktive Schlüsse aus einer Hypothese und darauf aufbauende Falsifikationen der Hypothese stützen. Theorien stellen für ihn zunächst eine (gesetzesartige) Allbehauptung auf vom Typ (G) dar:

(1) »(Für alle Zeitpunkte gilt:) Alle Gegenstände aus Metall dehnen sich aus, wenn sie erhitzt werden« oder
(2) »Alle zwei materiellen Objekte ziehen sich gemäß dem newtonschen Gravitationsgesetz gegenseitig an.«

Da solche Gesetze ebenso für die Zukunft wie für die Vergangenheit gelten sollen, ist es ausgeschlossen, dass wir solche Behauptungen jemals

vollständig überprüfen und *verifizieren* können. Wir haben im Gegenteil immer nur einen sehr kleinen Teil der Instanzen der Aussagen (G) überprüft. Insbesondere liegen die alle in der Vergangenheit. Möglich ist dagegen eine Falsifikation dieser Allaussagen. Treffen sie nur in einem Fall nicht zu, müssen wir die jeweilige Aussage als widerlegt ansehen. Das nennt Popper die Asymmetrie von Verifikation und Falsifikation.

Da die Zahl der Instanzen einer Aussage vom Typ (G) zumindest potentiell unendlich ist, kann es für Popper auch keine *probabilistische Bestätigung* von (1) oder (2) geben. Dazu ist die Anzahl der tatsächlich verifizierten Fälle immer viel zu klein. Das wird für beide Aussagen vom Typ (G) deutlich, wenn wir beachten, dass dabei über alle Zeitpunkte quantifiziert wird. Selbst probabilistische Modelle sprechen nach Popper und Miller (1983) gegen eine induktive Bestätigung solcher Naturgesetze. Dafür haben sie recht spitzfindige Argumente entwickelt, die vor allem etwas mit der Induktionseigenschaft der Induktionsverfahren zu tun haben, auf die ich später noch eingehen werde. So lässt sich jede Hypothese H zu einem Datum E auf folgende Art zerlegen: $H \equiv (H \vee E) \& (H \vee \neg E)$. Dabei wird der erste Teil $(H \vee E)$ durch E bestätigt, der zweite aber nicht. Die Frage ist dann, ob das eine sinnvolle Zerlegung von H darstellt und welcher Zusammenhang mit der Induktionseigenschaft besteht. Ich möchte hier demgegenüber nur eine einfache Abschätzung im Sinne von Popper betrachten.

Gesetze sind für Popper Allaussagen (etwa der Form $K \equiv \forall x(Fx \rightarrow Gx)$) über eine potentiell unendliche Zahl von Objekten. Nehmen wir etwa an, dass wir es bei einem Naturgesetz K zumindest mit einer sehr großen Anzahl N von Instanzen a_i zu tun haben, von denen wir nur wenige (etwa die ersten n) überprüfen konnten. Wir betrachten dabei nur F-Objekte und überprüfen dann, ob sie auf die Eigenschaft G aufweisen. Das sei für die ersten n F-Objekte der Fall, für die weiteren bis zur Zahl N>n sei dagegen noch nicht geklärt, ob sie unter G fallen. Für das Gravitationsgesetz könnten also z.B. Tripel bestehend aus zwei Gegenständen und einem Zeitpunkt als solche Instanzen dienen.

Wie groß ist dann die Wahrscheinlichkeit P(K)? Dazu gehen wir probeweise davon aus, die Instanzen seien probabilistisch unabhängig voneinander und hätten vor ihrem Test jeweils die Wahrscheinlichkeit

0,5 dafür, unter K zu fallen, würden sich beim Test aber alle als Instanzen von K erweisen, dann würde gelten:

$P(K) = \prod_{1 \leq i \leq N} P(a_i) = \prod_{n < i \leq N} P(a_i) = (0{,}5)^{N-n}$, da $P(a_i) = 1$ ist für die ersten n Objekte.

Dabei ist also N–n die Anzahl der noch nicht untersuchten Instanzen (auch der zeitlichen). Ist N gegenüber n sehr groß, wird die Wahrscheinlichkeit P(K) nahe bei 0 liegen und im Grenzfall für wachsende N gegen unendlich sogar null sein. Nehmen wir nur einmal an, dass N–n = 64 wäre, so liegt P(K) bereits im Bereich von 10^{-19}. Das ist eine Eins geteilt durch eine Zahl mit 19 Nullen und damit bereits unvorstellbar klein: praktisch gleich null. Jedenfalls sprechen die Überlegungen über solche Anzahlen nicht gerade dafür, dass es so etwas wie eine induktive Bestätigung von Gesetzen gibt, wenn bereits 64 noch offene Instanzen zu einer Wahrscheinlichkeit der Gesetze nahe null führen würden. Wenn wir statt der 0,5 einen anderen festen Wert <1 gewählt hätten, hätte das auch nicht viel geholfen.

Gegner von Popper wie Vertreter der induktiven Logik oder eines objektiven Bayesianismus (wie Williamson 2007) können hier vor allem an der probabilistischen Unabhängigkeitsannahme ansetzen und behaupten, mit zunehmender Anzahl positiver Instanzen für unser Gesetz K würde auch die (Vorher-) Wahrscheinlichkeit der weiteren Instanzen steigen, dass sie K erfüllen, und damit wäre ihre Wahrscheinlichkeit größer als 0,5 (oder womit wir ansonsten starten würden) und könnte mit jeder neuen Instanz weiter gegen 1 wachsen. Doch, wenn wir zunächst noch überhaupt nicht wissen, ob K gilt, dann haben wir eigentlich keinen Grund, die Instanzen von K als besondere Menge zu betrachten, bei der ein Element im Hinblick auf K eine Information über ein anderes tragen wird.

Das hieße, dass die Aufgabe der Unabhängigkeitsannahme praktisch schon einer Vorentscheidung zugunsten von K als Naturgesetz gleichkäme. Doch wie sollte man die begründen? Gerade Empiristen (wie etwa Carnap als einem der Erfinder der induktiven Logik) sollten diese Frage sehr ernst nehmen. Warum sollten wir von bestimmten Ereignissen oder singulären Tatsachen auf bestimmte andere schließen dürfen, wenn

wir nicht bereits einen naturgesetzlichen Zusammenhang vermuten? Wir müssten uns sonst bereits auf ein bestimmtes Hintergrundwissen stützen, dass diesen Zusammenhang begründet. Es ist daher in dieser Ausgangssituation unsere Annahme einer probabilistischen Unabhängigkeit der einzelnen Instanzen von K durchaus plausibel. Letztlich müssen wir aufgrund weniger untersuchter Instanzen auf noch sehr viele nicht untersuchte Fälle schließen. Poppers Bedenken gegen solche Schlüsse sind also sicherlich ernst zu nehmen.

Poppers eigener Lösungsvorschlag für das Induktionsproblem war schließlich radikal. Seiner Meinung nach gibt es schlichtweg *überhaupt keine induktive Bestätigung* von wissenschaftlichen Theorien, und wir benötigen in der Wissenschaft auch keine induktiven Schlüsse, sondern kommen allein mit der deduktiven Logik aus. Dann muss uns Humes Induktionsproblem nicht mehr beunruhigen. Popper stellt dafür die *Asymmetrie von Verifikation und Falsifikation* ganz in den Vordergrund seiner Überlegungen, denn Gesetze sind Allaussagen über eine potentiell unendliche Zahl von Objekten und sind daher niemals verifizierbar. Andererseits genügt bereits ein einzelnes Objekt bzw. System a, das die Eigenschaft F aufweist, aber nicht G (also: Fa & ¬Ga), um K zu falsifizieren. Für die Falsifikation sind wir nach Popper daher nur auf deduktive Schlüsse angewiesen: Aus Fa und K dürfen wir auf Ga schließen. Beobachten wir dann aber ¬Ga, können wir deduktiv schließen, dass K falsch sein muss.

Wir können z.B. aus der Behauptung, dass alle Metalle sich bei Erwärmung erhitzen, einzelne deduktive Schlussfolgerungen ziehen. Die Theorie sagt damit voraus, dass das nächste Stück Metall, das wir erhitzen, sich ebenfalls ausdehnen wird. Jede dieser Vorhersagen stellt einen empirischen *Test* für die Theorie dar, bei dem sie scheitern könnte; das heißt, wenn er fehlschlägt, könnten wir sie definitiv als falsch nachweisen. Das war Poppers Ansatzpunkt: *Eine wissenschaftliche Theorie ist zwar nicht verifizierbar, aber zumindest falsifizierbar.* Wenn sie allerdings ernsthafte Falsifikationsversuche unbeschadet übersteht, so hat sie sich für Popper *bewährt* und wir können uns in unseren Entscheidungen weiter auf sie stützen.

Für Popper war daher die Falsifizierbarkeit einer Theorie das entscheidende *Abgrenzungskriterium* zwischen Wissenschaft und Pseudowissen-

schaft. So verwarf er den Marxismus – verstanden als historische Theorie – und ebenso die Psychoanalyse als vermeintlich *pseudowissenschaftlich*, weil sie praktisch überall nur Bestätigungen der Theorie erblickten, aber nicht sagten, was die Theorie definitiv ausschließt und was sie somit falsifizieren könnte. Stattdessen sieht der Marxist in jedem Streik eine Bestätigung seiner Idee eines Klassenkampfes und der Psychoanalytiker in jeder Neurose eine Bestätigung der Psychoanalyse, ohne dass er ebenso angibt, wann bestimmte Phänomene gegen seine Theorie sprechen könnten. Selbst kritische Stimmen wurden oft nur psychoanalytisch als Verdrängungen gedeutet. Diese Art von »Immunisierung« gegen Kritik und die Fähigkeit alles zu »erklären« und immer nur auf »Bestätigungen« der Theorie zu stoßen, stellt für Popper die typischen Merkmale einer *Pseudowissenschaft* dar. Positives Vorbild war für Popper dagegen Einstein mit seiner allgemeinen Relativitätstheorie, der gleich bei der Erstellung seiner Theorie angab, welches überraschende Phänomen die Theorie vorhersagte – nämlich die Lichtablenkung im Schwerefeld der Sonne – und dass die Theorie zu verwerfen sei, wenn man das nicht beobachten könnte.

Intuitiv stellt die Falsifizierbarkeit zumindest ein naheliegendes Kriterium für die *Empirizität* einer Behauptung dar. Treffen wir auf einen Wetterbericht, der nur sagt: »Morgen regnet es, oder es regnet morgen nicht«, so stellt er schlicht *keine empirische Behauptung* auf. Seine Vertreter könnten dann zwar in jedem Fall morgen triumphieren und feststellen, dass sie das Wetter doch korrekt vorhergesagt haben, aber tatsächlich haben sie solange nicht wirklich etwas vorhergesagt, wie sie nicht ein bestimmtes Wetter ausgeschlossen haben. Erst wenn sie ein bestimmtes Wetter ausschließen, kann ihre Vorhersage aber auch falsifiziert werden. Popper plädierte daher dafür, gerade in der Wissenschaft immer besonders *kühne Hypothesen* aufzustellen, aus denen möglichst viele *potentielle Falsifikatoren* deduktiv folgen. Erst damit erhalten unsere Theorien einen hohen Informationsgehalt und erfüllen so unseren zweiten Anspruch an wissenschaftliches Wissen.

Potentielle Falsifikatoren sind für Popper die sogenannten *Basissätze*, also *singuläre Existenzaussagen* über ein beobachtbares Ereignis an einem bestimmten Ort. Diese werden allerdings – im Unterschied zum Vorgehen der Empiristen – nicht anhand von Beobachtungen akzeptiert,

sondern erst durch einen *Beschluss* der Wissenschaftlergemeinde. Damit enthält sein Ansatz allerdings schon an wesentlicher Stelle ein *konventionelles* Element, das über die deduktive Logik hinausgeht und dessen Rationalität durchaus bestreitbar ist. Auch hier möchte Popper vermeiden, dass wir etwa anhand von Sinneswahrnehmungen einen solchen Basissatz induktiv begründen müssen, denn dann träten natürlich gleich wieder beide Induktionsprobleme auf, denen Popper zu entgehen hoffte.

Kühne Hypothesen haben vor allem den erkenntnistheoretischen Vorteil, dass sie sich schnell falsifizieren lassen, wenn sie falsch sein sollten. Popper entwickelte dazu Vorstellungen vom *Falsifizierbarkeitsgrad* einer Theorie. Je höher der ist, umso gehaltvoller und erklärungsstärker ist eine Theorie und umso besser ist sie damit nach seiner Meinung. Je mehr Falsifikatoren eine Theorie aufweist, umso höher ist jedenfalls ihr empirischer Gehalt. Eine derartige Theorie haben wir anschließend nach Popper möglichst *strengen* bzw. *ernsthaften* Falsifikationsversuchen zu unterwerfen. Sollte sie falsch sein, zeigt sich das bald, und wir können wieder eine fehlerhafte Theorie ad acta legen. Es bleibt aber zu klären, wann ein Test einer Theorie als streng anzusehen ist.

Strenge Tests. Ganz im Sinne von Popper entwickelte Deborah Mayo (1996) eine Explikation von *strengen* oder *ernsthaften* Tests einer Theorie. Ernsthaft ist ein Test für eine Theorie T demnach gerade dann, wenn die Wahrscheinlichkeit hoch ist, dass T durch den Test falsifiziert wird, sollte T denn falsch sein. Es gilt also für einen *ernsthaften Test*: P(Test fällt negativ aus|T ist falsch) ist sehr hoch. Darauf werden wir im Zusammenhang mit den Signifikanztests der klassischen Statistik wieder zurückkommen, die eine Umsetzung dieser Idee darstellen. Sie stellen eine Art von *probabilistischer Falsifikation* dar. Jedenfalls hoffte Popper sich auf dem Weg über das Ausscheiden der falschen Hypothesen, langsam der Wahrheit anzunähern. Wir werden im Rahmen der eliminativen Induktion sehen, dass dieser Weg durchaus gangbar ist, aber nur wenn wir ihn um wesentliche Bestandteile erweitern, die Popper leider nicht akzeptiert hätte.

Zunächst sollte es aber intuitiv klar sein, was hier mit ernsthaften Tests gemeint ist. Denken wir etwa an die Hypothese: (H) »Fritz ist *kein* Spion des Iran.« Dann folgt aus der Hypothese (zusammen mit anderem Hintergrundwissen), dass er keine größeren Überweisungen

des iranischen Staates erhält. Das können wir überprüfen und daran könnte unsere Hypothese theoretisch natürlich sogar scheitern. Erhielte Fritz regelmäßige Überweisungen des iranischen Staates, würde das jedenfalls gegen unsere Hypothese sprechen. Stellen wir dagegen fest, dass er keine derartigen Überweisungen erhält, haben wir damit aber noch lange nicht nachgewiesen, dass es sich bei Fritz um keinen Spion handeln muss, denn es handelt sich wohl kaum um einen ernsthaften Falsifikationsversuch. Selbst wenn unsere Hypothese falsch wäre und es sich bei Fritz also um einen iranischen Spion handeln würde, würde er vermutlich trotzdem keine direkten Überweisungen des iranischen Staates erhalten.

Erst ein Test, bei dem die Hypothese tatsächlich höchstwahrscheinlich scheitern würde, wenn sie denn falsch wäre, würde einen ernsthaften Test abgeben. Das könnte in unserem Beispiel vielleicht eine gründliche Durchleuchtung der Lebensverhältnisse von Fritz sein, die alle auch indirekten Kontakte zu iranischen Staatsbürgern auffinden würde. Hat er keine solchen Kontakte, hätte er den Test überstanden. Das wäre wohl eher ein ernsthafter Test für unsere Hypothese H.

Es gibt andere Vorschläge für die Frage, wann ein Test als streng anzusehen ist, die sich aber auf subjektivere bzw. zufälligere Zusammenhänge stützen und damit weniger geeignet erscheinen, eine Explikation im Sinne Poppers zu formulieren. So könnte man eine Theorie anhand ihrer *überraschenden* Vorhersagen testen oder anhand von *neuen* Beobachtungen, deren Wahrheitswert zum Zeitpunkt des Aufstellens der Theorie noch nicht bekannt war. Im ersten Fall scheint es jedoch vielmehr darauf hinauszulaufen, unser subjektives Empfinden, was uns jeweils als überraschend erscheint, zum Maßstab für das Theorientesten zu erheben, während im zweiten Fall die Qualität von Theorientests von der zufälligen zeitlichen Reihenfolge abhängen würde, in der eine Theorie aufgestellt wurde und bestimmte Beobachtungen aufgetreten sind. Beides scheint nicht besonders gut zu den Intentionen Poppers zu passen, der auf der Suche nach einem möglichst objektiven Verfahren für Theorientests war. Daher scheint mir der Vorschlag von Mayo den besseren Weg zu einer Explikation der Strenge von Theorientests zu weisen.

Es gibt allerdings spezielle Situationen in denen eine Konzeption von »use novelty« – also ein Testen einer Theorie T vor allem an neuen Beobachtungen, die bei ihrer Aufstellung noch nicht bekannt waren – eine besondere Bedeutung bekommt. Bestimmte Datenerhebungen E stellen dann keine ernsthaften Tests und ihr Auftreten damit keine genuine Bestätigung von T mehr dar, wenn die Theorie Parameter aufweist, die mit Hilfe von E so angepasst wurden, dass die Theorie automatisch zu E passt. T wurde dann unter Einbeziehung von E bereits so formuliert, dass die Theorie nicht mehr an E scheitern kann. Damit stellt ein Test mit Resultat E auch keinen ernsthaften Test für T mehr dar. Schurz (2014a) zeigt genauer, wie das passieren kann, und wann bestimmte Daten daher keine *genuinen Bestätigungen* oder Bewährungen mehr darstellen können. Er hat damit einen alten Explikationsvorschlag aus dem Popper Lager wiederbelebt. Es gibt also möglicherweise verschiedene Ideen zur Explikation von strengen Tests.

Bewährung und Theoriendynamik. Poppers Vorschlag zur Theoriendynamik sieht damit schließlich wie folgt aus: Wir starten mit Problemen, die wir lösen möchten. Damit sind normalerweise Phänomene gemeint, die wir beobachten können und gerne *erklären* würden. Dann suchen wir nach einer möglichst gehaltvollen (erklärungsstarken) Theorie, die die Phänomene erklären (oder zumindest ableiten) kann, und unterwerfen diese im weiteren möglichst strengen Tests. Die Theorie soll also z.B. unsere bisherigen Phänomene erklären können und weitere kühne Vorhersagen aufstellen, die wir überprüfen können. Treffen diese Vorhersagen ein, so hat sich die Theorie *bewährt* und wir können uns vorläufig darauf stützen. Wir akzeptieren die Theorie als Arbeitshypothese weiterhin, bis sie tatsächlich falsifiziert wird. Sollten allerdings falsifizierende Basissätze vorliegen, dürfen wir nach Popper nicht versuchen, die Theorie durch Tricks wie *Ad-hoc-Hypothesen* (s.u.) zu retten, sondern sollten sie definitiv verwerfen und nach einer neuen Theorie suchen.

Unter den dann noch nicht-falsifizierten Theorien sollten wir wiederum die kühnste Theorie (mit dem höchsten *Falsifizierbarkeitsgrad*) als nächste Theorie auswählen und unser Verfahren damit entsprechend fortsetzen. Solange eine Theorie sich bewährt, dürfen wir an ihr festhalten, und es ist nach Popper auch rational, sich dann auf sie in seinen

Entscheidungen zu stützen. So nähert sich die Wissenschaft langsam und in einer Art von evolutionärem Prozess der Wahrheit an, indem die falschen Theorien im Laufe der Zeit ausgeschieden werden. Der wissenschaftliche Fortschritt besteht so nach Popper in einer langsamen Annäherung an die Wahrheit. Er versuchte sogar eine formale Explikation von Wahrheitsnähe anzugeben, die freilich an bestimmten formalen Problemen scheiterte (vgl. Oddie 2014).

Allerdings dürfen wir nach Poppers Ansicht diese *Bewährung* nicht mit irgendeiner Form *induktiver Bestätigung* einer Theorie verwechseln, denn sonst würden wir wieder in das humesche Problem der Induktion zurückfallen. Eine bewährte Theorie ist also eben *nicht* »wahrscheinlich wahr« oder »plausibler« als andere nichtfalsifizierte Theorien oder etwas Ähnliches.

Kritiker haben zu Recht bemängelt, dass damit unklar bleibt, was die Bewährung über eine Theorie überhaupt aussagen kann, wenn sie keine positive Auszeichnung darstellt, die wir als Indiz für die Wahrheit der Theorie auffassen dürfen. Die *Bewährung einer Theorie* sagt uns nach Popper nur etwas über die Vergangenheit (nämlich über bestandene ernsthafte Falsifikationsversuche) und nicht über die Zukunft. Die Bewährung einer Theorie macht es insbesondere nicht wahrscheinlicher, dass ihre Vorhersagen eintreffen.

Genau genommen können wir noch nicht einmal für bereits falsifizierte Theorien sagen, dass ihre Vorhersagen für die Zukunft unwahrscheinlicher sind als die der bewährten Theorien, denn natürlich können auch falsche Theorien wahre Vorhersagen liefern und über die Wahrscheinlichkeit von wahren Vorhersagen unserer bewährten Theorien wissen wir nach Popper überhaupt nichts. Warum sollten wir uns dann eher auf bewährte als auf falsifizierte Theorien stützen? Selbst diese scheinbare Selbstverständlichkeit lässt sich innerhalb von Poppers Ansatz nicht mehr begründen. Er schließt entsprechende Begründungsmöglichkeiten aus, weil die alle einen induktiven Bestätigungscharakter hätten, den wir seiner Meinung nach aber zurückweisen müssen.

Außerdem gilt: All unsere Falsifikationen lassen *unendlich* viele Theorien als nicht-falsifiziert zurück. Warum sollten wir eine von denen hervorheben? Und warum gerade die kühnste? Sie hat zwar den Vorteil, vermutlich schneller falsifiziert zu werden, wenn sie falsch sein sollte,

aber das zeichnet sie nicht gerade als Arbeitshypothese aus, auf die wir uns in unseren Entscheidungen stützen sollten. Popper und seine Anhänger konnten darauf keine überzeugenden Antworten finden. Dann hat er aber leider auch *keine Lösung des Induktionsproblems* anzubieten. Das lässt sich noch einmal veranschaulichen, indem wir uns ein Beispiel dazu anschauen.

Nehmen wir an, wir leiden unter *Magengeschwüren* und suchen nach einer Therapie dagegen. Es gibt jedoch zwei Theorien über die Auslöser der Magengeschwüre: Die erste Theorie besagt, dass es sich um Stresssymptome handelt und die beste Therapie daher in einer Stressreduktion besteht. Die zweite Theorie macht das Bakterium Helicobacter Pylori für die Magengeschwüre verantwortlich und empfiehlt daher eine Antibiotikatherapie (vgl. Thagard 1998). Wenn wir keine Daten gegen eine der Theorien haben, warum sollten wir dann die kühnere wählen und wonach richtet sich das? Und selbst wenn die erste Theorie falsifiziert wurde, hätten wir nach Popper keinen guten Grund, der anderen Theorie die wahrscheinlicheren Prognosen zuzusprechen. Das gilt erst recht, wenn beide Theorien noch nicht falsifiziert wurden und nur die zweite hätte sich schon mehrfach bewährt. Popper könnte nur behaupten, dass es tatsächlich *rational* sei, die eine der beiden Theorien vorzuziehen, wenn er behaupten dürfte, dass sie *wahrscheinlicher wahr* ist als die andere, aber genau diese Art von positiver Auszeichnung kann es in seinem Ansatz nicht geben.

Die *Bewährung* einer Theorie sagt nach Popper insbesondere nichts über die Zukunft und kann uns deshalb nicht wirklich für unsere Entscheidungen weiterhelfen. Also hat Popper das Induktionsproblem keineswegs gelöst, obwohl das sein Ziel war. Genau den Fall, der uns interessiert, kann seine Konzeption nicht behandeln. Er betont zwar immer, dass es rational sei, einer bewährten Theorie zu vertrauen, aber das ist solange nicht überzeugend, wie er unter Bewährung nicht doch eine Art von positiver Bestätigung versteht. Genau genommen hat Popper noch nicht einmal *komparative* Aussagen für die Zukunft anzubieten, nach denen wir einer nicht-falsifizierten Theorie wenigstens *eher* vertrauen sollten als einer falsifizierten. Das hilft uns also in Wahrheit nicht weiter in unserem Projekt, Entscheidungshilfen bereitzustellen. Wir werden die poppersche Konzeption um Nicht-poppersche Komponenten ergänzen

müssen, um damit in unserem Projekt voranzukommen. Die sogenannte *eliminative Induktion* stellt eine solche Ergänzung bereit.

Das Duhem-Quine-Problem. Da uns die Falsifikation in einem geeigneten Rahmen weiterhin begleiten und interessieren wird, sind noch einige andere Probleme der popperschen Konzeption zu untersuchen. Das bekannteste ist wohl das *Duhem-Quine-Problem*. Bereits Pierre Duhem (1978) hatte darauf hingewiesen, dass wir für die Ableitung von Beobachtungsaussagen aus einer Theorie in der Regel auf weitere Annahmen angewiesen sind. Das sind etwa Annahmen über die Zuverlässigkeit der Messgeräte für die eingesetzten Größen oder über Randbedingungen, die vorliegen müssen, damit unsere Theorie konkrete Vorhersagen abgeben kann oder anderes Hintergrundwissen, das erforderlich ist, um aus der Theorie bestimmte Beobachtungen schlussfolgern zu können.

So können wir zwar mit Hilfe der newtonschen Gravitationstheorie T die nächste Sonnenfinsternis vorhersagen, aber dazu benötigen wir eine ganze Reihe weiterer Hilfsannahmen A. Wir müssen z.B. wissen, wie groß die Massen der Planeten und der Sonne sind, wie groß ihre Impulse zu einem bestimmten Zeitpunkt sind, welche Planeten es überhaupt gibt und dass wir bestimmten Messinstrumenten jeweils vertrauen dürfen. Erst dann kann die Gravitationstheorie angewandt werden und eine entsprechende Vorhersage E – etwa eine Sonnenfinsternis – ableiten. Sollte sich E aber als falsch herausstellen, dann folgt nicht mehr gleich ¬T, sondern vielmehr nur noch ¬(T&A) ≡ ¬T∨¬A. Somit können wir T nur noch als falsifiziert betrachten, wenn wir die Hilfsannahmen A als bereits sehr gut begründet ansehen und in diesem Theorientest nicht mehr in Frage stellen.

Doch das ist in der Praxis natürlich nicht immer der Fall. So wurde bei Anomalien im Planetensystem aus Sicht von Newtons Theorie sogar mehrfach vermutet, dass es einen neuen Planeten geben müsste, der nur noch nicht bekannt sei, aber für die Bahnabweichungen verantwortlich gemacht werden könne. Somit können wir bei Konflikten mit der Erfahrung keineswegs immer einen Schuldigen ausmachen und die Vertreter einer Theorie tendieren dann verständlicherweise dazu, die Hilfsannahmen (etwa darüber, welche Planeten im Spiel sind) in Frage zu stellen. Der französische Astronom Le Verrier erklärte 1845 auf diese Weise Bahnabweichungen des Uranus durch einen weiteren äußeren

Planeten, den Neptun. Der wurde daraufhin bald an entsprechender Stelle entdeckt.

Später nahm Le Verrier an, dass auch die Abweichung der *Periheldrehung des Merkurs* von den Vorhersagen der newtonschen Gravitationstheorie durch einen weiter innen liegenden Planeten namens Vulkan verursacht würde; doch das erwies sich nach längerer Suche als falsch und erst Einstein konnte mit seiner allgemeinrelativistischen Gravitationstheorie diese Anomalie erklären und überwinden. Popper hätte sicher in beiden Fällen die Annahme eines weiteren Planeten als *Ad-hoc-Hypothese* verdammt. Doch niemand konnte vorhersehen, in welchen Fällen sie sich letztlich bestätigen würde und in welchen anderen Fällen sie sich schließlich als Blindgänger herausstellen würde. Wenn wir nun berücksichtigen, dass meistens Hilfsannahmen für eine Falsifikation erforderlich sind, erhalten wir die folgende Konzeption von Falsifikation:

Einfaches Schema für die Falsifikation einer Theorie

(1) Aus einer Theorie T und bestimmten begründeten Hilfsannahmen $A_1,\ldots, A_n$ folgt deduktiv ein Beobachtungssatz/Basissatz B.
(2) Wir beobachten, dass B falsch ist.
(3) Schlussfolgerung: Also wird T als falsifiziert verworfen, da die Hilfsannahmen gut begründet sind und vermutlich nicht zu dem Versagen der Theorie geführt haben.

Tatsächlich können wir normalerweise jeweils neue Hilfsannahmen erfinden, die es gestatten, die Theorie wieder mit den widerspenstigen Daten zu versöhnen. Der Physiker Lorentz nahm an, dass sich das Licht als Wellenphänomen im Medium Äther ausbreitet und wir daher Unterschiede in der Lichtgeschwindigkeit in unterschiedliche Richtungen feststellen würden, da sich die Erde schließlich schnell durch den Äther bewegen müsste. Das kann man sich ähnlich vorstellen wie bei Wasserwellen von einem Boot aus betrachtet. Als Lorentz dann versuchte mit dem von ihm entwickelten Interferometer den Ätherwind zu bestimmen und dazu die Lichtgeschwindigkeit in einer Richtung

mit der dazu senkrechten Richtung verglich, erhielt er bekanntermaßen immer wieder nur ein Nullergebnis. Doch statt daraus zu schließen, dass kein Ätherwind feststellbar ist und daher auch kein Äther existieren kann, nahm er an, dass das Interferometer sich beim Durchgang durch den Äther in Längsrichtung um gerade den Betrag *verkürzt*, dass der Effekt dadurch nicht mehr beobachtbar ist. So ließ sich die Äthertheorie trotz der Nullergebnisse aufrechterhalten.

Ad-hoc-Hypothesen. Popper wollte solchen Vorgehensweisen einen Riegel vorschieben und verbot deshalb die Annahme von sogenannten Ad-hoc-Hypothesen – und als solche betrachtete er die Verkürzungsannahme. Typischerweise werden die Hypothesen als Ad-hoc-Hypothesen betrachtet, die nur dazu dienen, eine Theorie zu retten, aber ansonsten keine eigene Stützung durch andere unserer Annahmen oder Daten aufweisen. Allerdings lassen sich Ad-hoc-Hypothesen nicht immer leicht als solche identifizieren. Lorentz hatte für seine Verkürzungshypothese durchaus eigenständige Gründe anzubieten, die sich aus dem Zusammenspiel der Atome mit dem Äther ergaben. Außerdem haben sich manche Ad-hoc-Hypothesen später als wahr und als fruchtbare Weiterentwicklungen der Theorie herausgestellt wie in unserem Beispiel des Uranus.

Popper will hier zu viel. Wir können nicht immer sofort und von vornherein sagen, ob eine bestimmte *Theoriemodifikation* sich schließlich als fortschrittliche Entwicklung oder degenerative Entwicklung herausstellen wird. Popper hoffte sogar, so ein einfaches Kriterium für Wissenschaftlichkeit entwickeln zu können; ein Ansinnen, was ich oben bereits zurückgewiesen habe. Es existiert leider kein einfaches Kriterium (keine einfache Regel), die uns gleich zu Beginn einer Theorieentwicklung sagen könnte, ob diese in eine Sackgasse führt oder nicht. Leverriers Annahmen weiterer Planeten haben sich einmal als hilfreich und im anderen Fall als Sackgasse erwiesen. Man müsste schon hellseherische Fähigkeiten haben, wenn wir das zum damaligen Zeitpunkt hätten vorhersagen wollen. Wir sollten daher nicht einfach behaupten, Ad-hoc-Annahmen seien immer zurückzuweisen, sondern wir sind gezwungen, die weiteren Entwicklungen abzuwarten.

Weitere Probleme des Falsifikationismus. Selbst die angesprochene logische Asymmetrie zwischen Verifikation und Falsifikation ist nicht so eindeutig, wie Popper sich das vorstellte. Damit wird Poppers einfache Falsifikationskonzeption problematisch. Es gibt auch immer wieder *Existenzannahmen* in wissenschaftlichen Theorien (bzw. eine gemischte Quantorenstruktur in den Theorien), so dass Theorien nicht immer sogleich durch ein Gegenbeispiel widerlegt werden können. Im newtonschen Gravitationsgesetz wird etwa behauptet, dass es *eine* Gravitationskonstante γ gibt, die für alle Paare von Objekten die Gravitationskraft mitbestimmt. Ist ein Versuch mit einer bestimmten Gravitationskonstante gescheitert, heißt das noch nicht zwingend, dass wir nicht eine andere Konstante und eine andere Zuordnung von Massen zu den Objekten finden können, so dass das Gravitationsgesetz doch wieder erfüllt wird. Doch das möchte ich mehr unter die technischen Probleme rechnen und Popper zugestehen, dass wissenschaftliche Theorien meistens die von ihm angenommene Allaussagenstruktur aufweisen.

Viel problematischer ist allerdings eine andere Schwierigkeit für den Falsifikationismus, die Popper selbst schon gesehen hat. Etliche Theorien schon in den Naturwissenschaften, aber vor allem in den Sozialwissenschaften stellen nur *probabilistische Aussagen* auf. Sie behaupten, dass einige Faktoren bestimmte Tendenzen dazu aufweisen, gewisse Entwicklungen auszulösen. Senkungen der Leitzinsen weisen die Tendenz auf, eine Belebung der Wirtschaft hervorzurufen, müssen das aber nicht in jedem Fall tun. Starkes Rauchen verursacht in vielen Fällen Lungenkrebs, aber wir kennen auch etliche Ausnahmen.

Probabilistische Aussagen sind jedoch im strikten Sinne nicht falsifizierbar. Behauptet jemand etwa, dass eine Münze gefälscht ist und eine Wahrscheinlichkeit von 0,6 für Kopf besteht, so sprechen zwar bestimmte Daten gegen diese Hypothese, aber immer wieder nur mit einer bestimmten *Wahrscheinlichkeit*. Nehmen wir etwa an, dass wir 1000-mal die Münze werfen und dabei keinmal Kopf erhalten, so ist dieses Resultat noch nicht logisch inkonsistent mit der Hypothese, dass die Wahrscheinlichkeit für Kopf 0,6 sei. Das Resultat hat nur eben eine geringe Wahrscheinlichkeit, wenn unsere Hypothese stimmt; es ist aber durch die Hypothese nicht logisch ausgeschlossen. Popper sah das Problem und sprach in solchen Fällen von *praktischer Falsifikation*. Sind

die Daten im Lichte der Hypothese sehr unwahrscheinlich, so dürfen wir die Theorie als *praktisch falsifiziert* bezeichnen. Im Zusammenhang mit den Signifikanztests in Kapitel 6 werde ich auf diese Frage zurückkommen und nenne solche Fälle *probabilistische Falsifikationen*. Hier sind allerdings spezielle methodologische Zusatzregeln nötig, die keineswegs harmlos sind und einige seltsame Konsequenzen gegenüber der klassischen Falsifikation aufweisen.

Damit ist aus dem einfachen Popper-Programm der Falsifikation schon ein viel komplexeres geworden, das erstens mit *Hilfsannahmen* zurechtkommen muss, die eine Theorie retten können, und daher nur noch die *weicheren Falsifikationen* im Sinne von Lakatos kennt (s. nächster Abschnitt). Außerdem müssen wir auf weitere methodologische Regeln zurückgreifen, um die wichtigen probabilistischen Theorien einbeziehen und praktisch falsifizieren zu können. Darauf werden wir vor allem in Kapitel 6 zurückkommen.

3.2 Forschungsprogramme und Paradigmen

Der Popperschüler Imre Lakatos (1974) plädiert schließlich für das Ende der popperschen »Sofort-Rationalität«. Wir können nicht sogleich entscheiden, ob eine Theorie zu verwerfen sei oder besser durch bestimmte Hilfshypothesen gerettet werden sollte. Selbst Ad-hoc-Hypothesen im Sinne von Popper können sich als wissenschaftlich wertvoll erweisen. Wir müssen also längere Entwicklungsphasen einer Theorie betrachten, in denen sich zeigt, ob die Hilfsannahmen zu fruchtbaren Weiterentwicklungen führen oder eben nicht. Lakatos entwarf dazu das Konzept der *Forschungsprogramme*. Das sind Abfolgen von Theorien, die dadurch zu einem Forschungsprogramm gehören, dass sie dieselben Grundgesetze (denselben *harten Kern*) aufweisen. Die einzelnen Theorien des Programms unterscheiden sich nur in ihrem »Schutzgürtel« von Hilfshypothesen, der sich mit der Zeit ändert, um auf die neuen Daten zu reagieren.

Die dabei auftretenden Theorieentwicklungen nennt Lakatos *fortschrittlich*, wenn sie neue Vorhersagen machen und wenn einige davon auch noch empirisch verifizierbar sind. Dienen die neuen Hilfsannah-

men dagegen nicht dazu, neue Vorhersagen aufzustellen oder sind die jedenfalls nicht empirisch zu bestätigen, dann gilt das Forschungsprogramm als eher *degenerativ*. Erst wenn sich auf längere Sicht ein Forschungsprogramm degenerativ entwickelt, sollten wir es aufgeben. Dabei sind selbst in diesen Fällen immer noch Wiederbelebungen möglich, wenn ein Forscher eine neue Idee hat, um das Forschungsprogramm wieder in eine progressive Richtung zu lenken. Poppers ursprüngliche Idee ist dagegen zu einfach, wenn er von schnellen Falsifikationen einer Theorie ausgeht. Lakatos' Aufweichung des popperschen Ansatzes hat allerdings den Nachteil, dass es keine klaren und endgültigen Falsifikationen mehr gibt. Wir müssen aber wohl mit den weicheren lakatosschen Falsifikationen auskommen.

Ein solches Forschungsprogramm zeichnet sich nach Lakatos auch noch dadurch aus, dass es eine Art von *Fahrplan* aufweist (»seine positive Heuristik«), wie es weiterzuentwickeln ist. Das ist natürlich wieder eine recht idealisierende Annahme, aber sie hilft zumindest dabei, die Fortschrittlichkeit des Programms besser einzuschätzen. Wir wissen etwa, welche Ressourcen im Hinblick auf mögliche Theoriemodifikationen wir noch zu erwarten haben und können so begründete Vermutungen wagen, ob dabei Fortschritte zu erwarten sind.

Lakatos veranschaulichte dieses Vorgehen am Beispiel des Bohrschen Atommodells. Das ging davon aus, dass die Elektronen sich auf bestimmten Bahnen um den Atomkern bewegen, wie die Planeten um die Sonne. Alle Theorieverbesserungen, die für das Planetensystemmodell erfolgten, sollten dann auch ins Atommodell aufgenommen werden. Sie erwiesen sich zunächst überwiegend als sehr fortschrittlich. Da ist der Übergang von Kreisbahnen zu Ellipsenbahnen zu nennen, der Einsatz der sogenannten reduzierten Massen und schließlich sind relativistische Korrekturen am Modell zu erwähnen. In diesem speziellen Fall lag also ein recht klarer Fahrplan vor, der schon erkennen ließ, wohin sich das Forschungsprogramm entwickeln würde. Doch das ist in den meisten Fällen kaum zu erwarten und uns bleibt dann nur übrig, einfach abzuwarten, wie kreativ die Vertreter des Programms sind und ob sich dabei empirische Erfolge einstellen. Die einfache Poppersche Falsifikationsregel haben wir damit in jedem Fall verloren.

Thomas Kuhn. Der Wissenschaftshistoriker Thomas S. Kuhn (z.B. in 1976) geht dann noch einen Schritt weiter und betont ebenfalls, dass es eine Art von einfachem Algorithmus zur Theorienbewertung nicht geben kann. Außerdem versucht er an Beispielen aus der Wissenschaftsgeschichte nachzuweisen, dass Wissenschaftler keinesfalls darum bemüht sind, falsifizierende Anwendungen ihrer Theorien zu finden. Statt ernsthafter Tests suchen sie in den meisten Phasen der wissenschaftlichen Entwicklung – der sogenannten *Normalwissenschaft* – eher nach erfolgreichen Anwendungen und somit nach Bestätigungen ihrer Theorien. Kuhn spricht in diesem Zusammenhang von *Paradigmen*, die es im Rahmen der *normalen Wissenschaft* zu verteidigen gilt, und erst, wenn sich hartnäckige *Erklärungsanomalien* zeigen, die sich selbst bei wiederholten Versuchen nicht in das Paradigma integrieren lassen, kann es zu einer *Krise* kommen. Zum Paradigma gehört nicht nur die Theorie selbst, sondern auch die dahinterstehende Metaphysik oder Naturphilosophie, eine Mengen von intendierten Anwendungsbeispielen der Theorie und neben bestimmten Bewertungsmaßstäben für Theorien nennt Kuhn für die Paradigmen noch einiges mehr aus dem Umfeld wissenschaftlicher Theorien.

Wenn sich in einer Krise noch ein neues Paradigma findet, das die Anomalien auflösen kann, kommt es nach Kuhn zur *wissenschaftlichen Revolution* und zum Wechsel zum neuen Paradigma sowie zum Verwerfen des alten. Solche Revolutionen sind aber keine Falsifikationen gemäß bestimmten wissenschaftsinternen Spielregeln (wie bei Popper oder Lakatos), sondern werden von Kuhn in die Nähe von *Gestaltwechseln* gerückt und haben womöglich zumindest z.T. wissenschaftspsychologische oder -soziologische Ursachen. Außerdem ist nach Kuhn immer das Auftreten eines neuen Pardigmas für die Ablösung des alten erforderlich.

Allerdings lassen sich in der Wissenschaftsgeschichte auch Preisgaben eines alten Paradigmas finden, die man als weiche Falsifikationen im Sinne von Lakatos deuten kann, ohne dass unbedingt schon ein alternatives Paradigma vorhanden sein musste. Auch die kuhnsche Darstellung der Wissenschaftsdynamik dürfen wir also nicht als striktes Schema verstehen.

Allerdings hat Kuhn eine Reihe bedenkenswerter Punkte genannt, die wir als weitere Probleme einer popperschen Wissenschaftstheorie

ansehen können. Da sind vor allem seine Überlegungen zur *Inkommensurabilität von Theorien* über die Kluft einer wissenschaftlichen Revolution hinweg zu nennen. Das bedeutet, dass es keinen gemeinsamen epistemischen Maßstab für die Bewertung der Qualität solcher Theorien mehr gibt, was natürlich die meisten unserer Vorstellungen von einer langsamen Verbesserung unserer Theorien und einem so gestalteten Fortschritt in der Wissenschaft untergräbt.

Kuhn nennt dazu eine Reihe von Problemen, die beim Paradigmenwechsel auftreten können, die man zumindest ernsthaft in Betracht ziehen sollte (vgl. dazu Bird 2013). Nehmen wir den Fall, dass eine wissenschaftliche Revolution stattfindet, die u.a. den Übergang von einer Theorie T zu einer Theorie T* beinhaltet. Woran kann dann ein Vergleich der beiden Theorien im Hinblick auf ihre Wahrheitsnähe oder im Hinblick auf andere epistemische Qualitäten scheitern?

Zunächst nennt Kuhn die *semantische Inkommensurabilität*. Damit ist gemeint, dass die Theorien vielleicht denselben Term aufweisen, dieser aber trotzdem eine andere Bedeutung in den Theorien haben kann. Wir werden später noch eine Auffassung der Bedeutung theoretischer Terme kennenlernen, wonach diese von den Gesetzen der Theorien und gewissen Brückenprinzipien abhängig ist, und wenn die in den beiden Theorien verschieden sind, kann sich das demnach auf die Bedeutung der Terme in den Theorien auswirken. So bezeichnet der Term »Masse« in der klassischen Mechanik eine Erhaltungsgröße, die durch eine Konstante repräsentiert wird, aber in der relativistischen Mechanik stellt zumindest die »relativistische Masse« eine Funktion der Geschwindigkeit dar und kann in Energie umgewandelt werden.

Diesen semantischen Problemen kann ich hier nicht nachgehen. Die Kritiker Kuhns wenden aber ein, dass sich zwar die Bedeutungen der Terme ändern können, aber die Referenten der Terme (das, worauf sie sich beziehen) dabei gleichwohl erhalten bleiben können. In unserem Beispiel wird man auf Seiten der relativistischen Mechanik dabei insbesondere an die Ruhemasse denken. Wenn die Referenten von »Masse« und »Ruhemasse« dann aber dieselben sein sollten, genügt das womöglich bereits für einen Vergleich der Theorien. Hier wird man in jedem Einzelfall weitere naturphilosophische Analysen anstrengen müs-

sen, um ermitteln, welche Probleme semantischer Inkommensurabilität tatsächlich auftreten.

Ein weiteres Problem sieht Kuhn in der *Beobachtungsinkommensurabilität*, wonach die Vertreter unterschiedlicher Paradigmen auch etwas Verschiedenes sehen, selbst wenn sie eigentlich dieselben Beobachtungen machen. Manchmal spricht Kuhn sogar davon, dass sie in verschiedenen Welten leben. Meines Erachtens übertreibt Kuhn die an dieser Stelle auftretenden Probleme, aber das ist hier nicht mein Thema. Erste Antworten dazu habe ich schon in Kapitel 2 formuliert.

Spannender für uns ist die *methodologische Inkommensurabilität*. Dabei geht es darum, dass unsere Regeln für das induktive Schließen zu keinem eindeutigen Ergebnis führen oder wir sogar noch nicht einmal von bestimmten Regeln sprechen können. Es werden dann Entscheidungen der Wissenschaftlergemeinschaft für bestimmte Theorien nicht mehr anhand wissenschaftsinterner Werte (wie vermutlicher Wahrheitsnähe) getroffen, sondern anhand externer Kriterien, die keinen klaren Zusammenhang zu unseren epistemischen Werten aufweisen. Womöglich entscheiden sich die jüngeren Wissenschaftler in einer Art von Auflehnung gegen die etablierten Wissenschaftler und ihre Theorien und für ein neues Paradigma, ohne dass das durch die epistemischen Vorzüge des neuen Paradigmas epistemisch rational begründet werden könnte.

Wenn solche Phänomene vorkommen, stellt das die Rationalität des wissenschaftlichen Vorgehens und unsere Vorstellungen von wissenschaftlichem Fortschritt natürlich in Frage. Daher wird Kuhn gerne von denjenigen herangezogen, die überhaupt bestreiten möchten, dass es so etwas wie wissenschaftliche Rationalität gibt. Doch Kuhn selbst hat sich immer dagegen gewandt. Er erkennt in der Wissenschaftsgeschichte sehr wohl den wissenschaftlichen Fortschritt, er glaubt nur nicht, dass wir den mit einer einfachen Akkumulationskonzeption von Fortschritt oder mit einfachen Regeln der Theorienwahl richtig beschreiben können. Wo liegen dann die Probleme?

Kuhn hat in seinen Arbeiten viele wissenschaftshistorische Studien durchgeführt, mit denen er belegen möchte, dass bestimmte Formen von Inkommensurabilität auftreten. Diese Fallstudien können wir hier nicht diskutieren. Stattdessen möchte ich nur die grundlegenden Probleme

schildern. Ein wichtiges Problem der methodologischen Inkommensurabilität besteht darin, dass wir mehrere interne Bewertungskriterien für Theorien finden, die in einem Konflikt miteinander stehen können, und dass es für diese Fälle keine klaren Vorrangregeln gibt.

Kuhn nennt als Bewertungskriterien u.a. (1) Präzision, (2) Konsistenz (mit den Daten, aber auch anderen Theorien), (3) Reichweite (die Theorie sollte über die Daten hinausgehen) (4) Einfachheit (bzw. Vereinheitlichungskraft) und (5) (theoretische) Fruchtbarkeit (vgl. Kuhn 1977, 338). So kann die eine Theorie präziser und die andere einfacher sein. Welche ist dann epistemisch besser? Außerdem weisen die Kriterien bestimmte *Vagheitsspielräume* auf. Was ist denn z.B. unter Einfachheit genau zu verstehen? Geht es darum, dass möglichst wenige Entitäten postuliert werden oder eher darum, dass es möglichst wenige und einfache Gesetze gibt?

Für mich sind die kuhnschen Kriterien eher sekundäre Kriterien, die zunächst alle auf die grundlegenden Ziele der Wissenschaften zurückgeführt werden müssen (vgl. Kapitel 2). Einfachheit kann tatsächlich etwas mit Wahrheitsnähe zu tun haben, wie wir in Kapitel 7 noch sehen werden, aber dazu ist zunächst genauer zu klären, wie wir Einfachheit in diesem Sinne zu verstehen haben. (Einfachheit hat für mich vor allem mit der Anzahl der freien Parameter in einer Theorie zu tun, die ihrerseits mit dem Gehalt verknüpft sind. Je mehr Parameter, desto geringer ist normalerweise der Gehalt einer Theorie, und das ist nach Kapitel 2 wiederum mit der Wahrheitsnähe assoziiert.) Explizieren wir Einfachheit dagegen nicht im Hinblick auf die Wahrheitsnähe von Theorien, scheint es sich dabei nur um eine *pragmatische Tugend* zu handeln, die darauf abzielt, dass wir die Theorie leicht anwenden können. Hier wären also weitere Debatten erforderlich, um Kuhns Punkt genauer rekonstruieren zu können.

Doch wir haben in Kapitel 2 bereits gesehen, dass es selbst bei den grundlegendsten Zielen der Wissenschaften *zwei Ziele* gibt, die in einem Spannungsverhältnis stehen und daher in Konflikt geraten können, nämlich der *Gehalt einer Theorie* und ihre *epistemische Wahrscheinlichkeit* oder das Ausmaß, in dem sie begründet ist. Dazu wird vor allem debattiert, wie man die beiden Aspekte miteinander verrechnen kann, um eine einheitliche Konzeption von Wahrheitsnähe zu erhalten.

Es scheint dafür keine einfache Lösung zu geben. Kuhn scheint also auf ein Problem aufmerksam zu machen, dass bereits auf der basalen Ebene unseres erkenntnistheoretischen Projekts zu finden ist. Außerdem werden wir im nächsten Kapitel sehen, dass gerade der hier propagierte Schluss auf die beste Erklärung wiederum ähnliche Fragen aufwirft. Die Qualität von Erklärungen kennt unterschiedliche Dimensionen, die in Konflikt geraten können. Damit ist natürlich noch nicht gesagt, ob solche Konflikte in der Wissenschaftspraxis auch tatsächlich häufiger auftreten. Das können wir nur anhand weiterer Fallstudien zur Wissenschaftsgeschichte entscheiden.

Wir kennen das Problem bereits aus unserem Alltag und es muss dort keineswegs unlösbar erscheinen. Bei der Auswahl einer geeigneten Wohnung werden wir etwa die Kriterien wie deren Größe, den Preis und die Lage und vermutlich noch weitere Kriterien berücksichtigen. Die Kriterien geraten typischerweise in Konflikt miteinander und z.B. das Kriterium der guten Lage zeigt außerdem, dass es nicht nur eine Weise der Explikation geben wird. Höchstwahrscheinlich haben die meisten von uns auch keine klaren Verrechnungsregeln für diese Kriterien. Trotzdem gibt es sicher viele Fälle, in denen eine Wohnung klar die bessere ist, selbst wenn sie der anderen Wohnung nicht in allen Kriterien überlegen ist. So ähnlich kann es uns auch in vielen Fällen der Theorienwahl ergehen, und das Fehlen von Vorrangregeln ist jedenfalls noch kein Beweis für eine Irrationalität bei der Auswahl. Trotzdem müssen wir diese Probleme im Blick behalten und Kuhn zugestehen, dass es dadurch keineswegs trivial erscheint, zu explizieren, was wir mit Fortschritt in der Wissenschaft meinen.

3.3 Die eliminative Induktion

Damit uns die Falsifikationen auch im positiven Sinne voranbringen, müssen wir zumindest einen weiteren Schritt gehen, den Popper abgelehnt hat. Wir müssen zubilligen, dass wir mindestens in manchen Fällen schon Gründe für die Annahme haben, dass eine richtige Darstellung der Ursachen und damit eine wahre Theorie für bestimmte Phänomene bereits in einer endlichen Liste L = $\{T_1,...,T_n\}$ zu finden ist, die uns

vorliegt. In dieser Situation helfen uns Falsifikationen einiger dieser Theorien tatsächlich weiter, denn das spricht für die verbliebenen Theorien. Im Idealfall bleibt nur noch eine Theorie übrig, und die liefert dann die richtige Darstellung unserer Phänomene. Zumindest verfügen wir so insgesamt über gute Gründe für die Behauptung, dass die verbliebene Theorie wahr ist. Damit wird eine Theorie *positiv* ausgezeichnet, was Popper noch nicht möglich war. Dieses Verfahren stammt letztlich von Francis Bacon und man nennt es die *eliminative Induktion*.

Schema der eliminativen Induktion:
(1) Es liegt eine endliche Liste L = $\{T_1,...,T_n\}$ von Theorien vor, für die wir annehmen, dass die gesuchte wahre Theorie für einen bestimmten Bereich sich darin befindet.
(2) Wir können alle Theorien der Liste bis auf eine – etwa T_1 – falsifizieren.
(3) *Schlussfolgerung*: Dann ist diese Theorie T_1 durch unser Verfahren bestätigt und sollte vorläufig akzeptiert werden.

Das ist auch das Verfahren, das Sherlock Holmes seinem Adlatus Dr. Watson gegenüber als seine Methode darstellt. Dabei betont er manchmal, es sei rein deduktiv, und das lesen wir auch immer wieder in der Literatur, doch das ist nicht ganz richtig. Typische induktive Irrtumsrisiken kommen an mindestens zwei Stellen ins Spiel: 1. bei der Auswahl der Liste L und 2. bei den einzelnen Falsifikationen, die eben oft nur weiche (und revidierbare) Falsifikationen sind. Bei Holmes besteht die Liste L aus der Liste aller schwach Verdächtigen, die zumindest ein *Motiv* und die *Gelegenheit* zu der in Frage stehenden Tat hatten. Natürlich können wir dabei jemanden übersehen haben und müssen später unsere Liste eventuell ergänzen. Dann werden im nächsten Schritt diejenigen von der Liste eliminiert, die letztlich doch ein Alibi aufweisen oder psychologisch gesehen nicht zu der Tat in der Lage waren etc. Holmes behauptet schließlich, dass der Übrigbleibende der Täter sein muss, ganz gleich wie unwahrscheinlich das auch sei. Dem sollten wir so nicht folgen. Die letzte verbleibende Theorie sollte möglichst alle Daten zumindest gut

erklären können und so auch selbst eine gewisse Plausibilität aufweisen. Wir werden auf die weiteren Anforderungen im Rahmen des Schlusses auf die beste Erklärung wieder zurückkommen. Doch zunächst einmal dürfte die Grundidee klar sein.

Bloß wie gelangen wir in der Wissenschaft zu einer geeigneten Liste? Popper hält das für unmöglich angesichts der schieren Anzahl an möglichen Erklärungen für ein Phänomen. Doch in der Praxis sieht das wieder ganz anders aus. Wir haben oft schon einen gewissen Überblick über die zur Verfügung stehenden Theorien und andere Ansätze erscheinen deutlich unplausibler. So dürfen wir uns etwa wieder auf bestimmte kausale Rahmenannahmen bzw. anderes Hintergrundwissen stützen, um viele logisch mögliche Erklärungen als zu unplausibel von vornherein auszuschließen. Hier bleibt natürlich immer das Irrtumsrisiko, das uns bei jedem induktiven Schließen plagt, aber das müssen wir eben immer akzeptieren. Schauen wir uns ein Beispiel an, nämlich die Entdeckung der Ursachen von BSE, dem sogenannten *Rinderwahnsinn*. Das sollte das Verfahren durchaus plausibel erscheinen lassen.

Wenn wir heutzutage in der Medizin nach der Ursache einer neuen Krankheit suchen, haben wir normalerweise schon eine Liste von möglichen *Ursachentypen* vor Augen, die etwa Folgendes umfasst: Infektionskrankheiten, Erbkrankheiten, Mangelernährung, Vergiftungen und Verätzungen sowie Verbrennungen, Erkrankungen des Immunsystems, Umwelteinflüsse und Unfälle, die zu äußeren Schädigungen führen, Tumorerkrankungen, psychische Erkrankungen, degenerative Erkrankungen durch Abnutzung etwa bestimmter Gelenke und vielleicht noch einige wenige andere Typen. Die Liste ist zumindest nicht besonders lang. Innerhalb der einzelnen Typen von Krankheiten sind allerdings wiederum einige Untertypen zu unterscheiden. Infektionskrankheiten beinhalten solche durch Bakterien, Viren, Amöben etc. Bei der Untersuchung von Rinderwahnsinn war man schnell zu der Überzeugung gelangt, dass es sich um eine *Infektionskrankheit* handelt: Wenn man das Fleisch infizierter Tiere verfütterte, konnte man nämlich dadurch die Krankheit wiederum bei anderen Tieren auslösen. Das können die anderen Hypothesen praktisch nicht erklären und somit konnten sie relativ rasch eliminiert werden. Dann war die weitere Frage, welche der Untergruppen von Ursachen innerhalb dieser Gruppe sich eliminieren

lässt. Da etwa antivirale und antibakterielle Mittel das Fleisch nicht desinfizieren konnten, wurden weitere Unterfallhypothesen eliminiert. Leider hatten wir damit den speziellen Fall erhalten, dass eigentlich keine Hypothese mehr übrig blieb und unsere Unterliste innerhalb der Infektionskrankheiten schließlich noch einmal erweitert werden musste. Um 1982 wurde deshalb durch Prusiner die Prionenhypothese entwickelt. Pathogene Prionen sind Eiweißkomplexe, die keine Erbinformation mehr enthalten, aber durch ihre bloße Anwesenheit bestimmte Gehirnzellen verändern können. Diese neue Theorie blieb als einzige übrig und stellte so die beste Erklärung dar, der wir nun folgen sollten.

Bird (2011) beschreibt auch das recht bekannte Beispiel der Aufdeckung der Ursachen des Kindbettfiebers durch Semmelweis als ein Beispiel für eine typische eliminative Induktion. Ignaz Semmelweis (1983) versuchte 1846 und 1847 am Wiener Krankenhaus herauszufinden, warum in der Abteilung I der Entbindungsstation deutlich mehr Fälle von Kindbettfieber auftraten als in der Abteilung II. Die Erkrankungsrate in Abteilung I lag etwa dreimal so hoch wie in der anderen Abteilung. Dazu entwickelte er damals schon erstaunlich viele alternative Erklärungshypothesen, die *mögliche Ursachen des Kindbettfiebers* aufzeigen:

(1) Mögliche Überfüllung in Abteilung I.
(2) Epidemische Einflüsse bzw. das Klima.
(3) Ungeschickte Untersuchungen durch die Medizinstudenten in Abteilung I im Unterschied zu dem geschickteren Umgang durch die Hebammen in Abteilung II.
(4) Psychische Effekte durch den Priester, der den sterbenden Frauen die Sakramente erteilte, wozu er durch die Abteilung I hindurchgehen musste.
(5) Die Lage bei der Geburt (auf dem Rücken liegend in Abteilung I)

Die Hypothese (1) konnte den *Unterschied* nicht erklären, zumal nach Bekanntwerden der Probleme der ersten Abteilung die zweite Abteilung die überlaufene war. Auch die zweite Hypothese lieferte zwar eine schon damals plausible Erklärung für das Auftreten von Kindbettfieber, aber es konnte nicht den *Unterschied* zwischen den Abteilungen erklären und wurde daher von Semmelweis für sein spezielles Erklärungsziel als unbrauchbar ausgesondert. Die Hypothese (3) verweist zwar auf

einen Unterschied zwischen den Abteilungen, aber Semmelweis konnte sich selbst davon überzeugen, dass er nur marginal war, und schloss daher aus, dass er so gravierende Folgen haben könnte. Es blieben also die beiden Erklärungsansätze (4) und (5), die Semmelweis anhand von experimentellen Interventionen eliminieren konnte. Zunächst ließ er den Priester den weiteren Weg außen an der Abteilung I vorbei gehen, ohne dass sich dadurch die Erkrankungsraten veränderten. Auch die Veränderung der Lage bei der Geburt hatte keine entsprechenden Effekte. Also musste eine neue Hypothese her. Semmelweis kam darauf, weil ein Kollege sich bei einer Autopsie verletzt hatte und danach unter ähnlichen Symptomen wie dem Kindbettfieber gestorben war. Also vermutete er als Erklärung:

(6) Durch die Untersuchungen der Leichen durch die Studenten, die regelmäßig Autopsien vornahmen, gelangte ein bestimmtes Leichenmaterial an die Hände der Studenten und durch die Untersuchungen in die Frauen und verursachte dort das Kindbettfieber. Die Hebammen in der Abteilung II führten dagegen keine Autopsien durch.

Semmelweis konnte diese Annahme weiter bestätigen, indem er die Studenten dazu brachte, sich vor den Untersuchungen der Schwangeren die Hände mit Chlorkalk zu desinfizieren. Danach sank die Erkrankungsrate schließlich sogar unter die in der zweiten Abteilung. Das war ein ziemlich beeindruckender Erfolg und ein vorbildlicher Forschungsprozess. Trotzdem ließen sich viele Kollegen nicht von seinen Ergebnissen überzeugen, und viele Ärzte weigerten sich, selbst entsprechende Desinfektionsmaßnahmen durchzuführen. Das lag wohl u.a. daran, dass aufgrund manchmal mangelhafter Desinfektionen seine Resultate nicht immer reproduziert wurden, aber ebenfalls an persönlichen Rivalitäten und Feindschaften, für die in diesem Fall viele weitere Frauen mit dem Leben bezahlen mussten, was Semmelweis auch sehr aufgebracht hat (vgl. Zankl 2010).

Wir sehen wieder, dass uns schon die Daten selbst immer wieder Probleme bereiten können. Trotzdem lässt sich an diesem Beispiel sehr schön das Zusammenspiel von eliminativer Induktion und der Suche nach positiven Erklärungen erkennen. Die ersten beiden Hypothesen konnten wir aufgrund mangelnder Erklärungsleistung zumindest für

die beobachtete Differenz eliminieren. Die nächsten drei Hypothesen gaben jeweils Hinweise darauf, welche Beobachtungen oder Experimente nun weiterführend sein würden, um abduktiv schließen zu können, die entsprechenden Experimente führten aber letztlich doch nur zu weiteren Eliminationen. Dann verhalf ein Analogieschluss Semmelweis zu einer neuen Hypothese, die sich schließlich durch weitere Daten bestätigen ließ. Die 6. Hypothese bot nun für die Differenz eine gute Erklärung sowie für das ähnliche Krankheitsbild des Kollegen und schließlich dafür, dass das Desinfizieren der Hände die Differenz zwischen den beiden Abteilungen beseitigte. Diese gezielte experimentelle Intervention war ein entscheidender Hinweis für Semmelweis. So ergab sich eine überzeugende Bilanz für die Hypothese (6). Dass insbesondere diese *Erklärungsleistung* letztlich den Ausschlag geben sollte, wird uns gleich zum sogenannten *Schluss auf die beste Erklärung* führen.

Auch die besonders in den Sozialwissenschaften und der Medizin sehr beliebten *Signifikanztests* lassen sich als Instanzen der eliminativen Induktion verstehen. Dabei geht es häufig um *kontrollierte Experimente*, die wir mit Hilfe von Signifikanztests auswerten. Wenn wir z.B. eine einfache Hypothese H bestätigen möchten, die besagt, dass U eine Ursache für den Effekt E darstellt, wobei U etwa die Gabe eines Medikaments und E die entsprechende Heilung bezeichnet, versuchen wir dazu eine sogenannte Nullhypothese H_0 als Gegenhypothese aufzustellen, die dann besagt, dass U keine Wirkung auf E hat, und versuchen diese als nächstes probabilistisch zu falsifizieren. Durch Falsifizieren der Gegenhypothese haben wir dann unsere Ausgangshypothese H bestätigt.

In der Regel sind natürlich noch weitere Konkurrenzhypothesen H_1,...,H_n denkbar, die andere mögliche Ursachen für E benennen. Die sollen i.A. anhand des Versuchsaufbaus falsifiziert werden. Zu diesem Zweck betrachten wir üblicherweise eine Versuchs- und eine Kontrollgruppe, die so zusammengestellt wurden (z.B. per Zufallsauswahl), dass vermutlich alle für das Auftreten von E relevanten Faktoren bis auf U auf beide Gruppen gleich verteilt sind. U sollte hingegen möglichst nur in der Versuchsgruppe zu finden sein. Stellt sich der Effekt E ebenfalls nur in der Versuchsgruppe ein und nicht in der Kontrollgruppe, haben wir somit einen guten Grund für die Annahme, dass U den Effekt E verursacht hat, da die möglichen anderen Faktoren

als Ursache für diesen Unterschied ausgeschlossen werden können, weil sie nach der Zufallsauswahl in beiden Gruppen in gleicher Weise vorliegen sollten. Das ist gerade der Kern der Differenzmethode, auf die ich in Kapitel 7 noch genauer eingehen werde. Damit sind viele mögliche Konkurrenzhypothesen auf einen Schlag falsifiziert worden.

Allerdings handelt es sich in der Medizin oder den Sozialwissenschaften häufig nur um *probabilistische* Effekte, d.h., in der Versuchsgruppe tritt E etwas häufiger auf als in der Kontrollgruppe, aber eben nicht immer. Dann bleibt eine weitere Gegenhypothese übrig, die es noch zu falsifizieren gilt. Das ist die erwähnte Nullhypothese H_0, die besagt, dass es sich bei unserem Versuchsergebnis bloß um eine *Zufallsschwankung* handelt und U in Wahrheit überhaupt keinen kausalen Einfluss auf E hat. Nur durch Zufall ist in der Versuchsgruppe die Heilungsquote höher als in der Kontrollgruppe. Das kann eben bei Zufallsprozessen geschehen. Um diese Hypothese falsifizieren zu können, muss eine probabilistische Falsifikation her, nämlich die sogenannten Signifikanztests. Die sind jedoch unter Wissenschaftstheoretikern recht umstritten, weil probabilistische Falsifikationen viele Unterschiede zu strikten Falsifikationen aufweisen (vgl. Howson & Urbach 1989 und s.a. Kap. 6). Insbesondere müssen wir bei Zufallsprozessen mit Abweichungen in jeder Größenordnung irgendwann einmal rechnen und können dann die Nullhypothese fälschlicherweise als falsifiziert betrachten.

Dabei wird zu H_0 ein *Zurückweisungsbereich* Z festgelegt. Sollte das experimentelle Ergebnis in den Bereich Z fallen, gilt H_0 als probabilistisch falsifiziert und damit H als indirekt gestützt. Dabei wird Z meist so festgelegt, dass es bei Zutreffen der Hypothese H_0 nur eine Wahrscheinlichkeit von höchstens 5% gibt, dass das Resultat in den Bereich Z fallen wird. Das heißt, wenn ein Ergebnis beobachtet wird, das zu Z gehört, also ein Ergebnis auftritt, das bei Annahme von H_0 als sehr unwahrscheinlich gelten muss, dann gilt H_0 als probabilistisch falsifiziert. Das ist eine wichtige Anwendung einer neueren Variante der eliminativen Induktion. Fehlgeschlagene Signifikanztests (also Tests ohne signifikantes, d.h. falsifizierendes Ergebnis für die Nullhypothese) können demnach eigentlich nicht weiter interpretiert werden. Sie sprechen jedenfalls nicht sogleich gegen H, sondern das Verfahren ist praktisch nur im Sande verlaufen, weil es uns nicht gelungen ist,

alle Hypothesen bis auf eine zu falsifizieren. Das ist typisch für die eliminative Induktion (vgl. dazu Kap. 6).

Schema für die eliminative Induktion in einem Experiment plus Signifikanztest:

(1) Vermutlich ist eine der Theorien H, H_0, H_1,...,H_n eine wahre Beschreibung der Ursachen eines Phänomens E.
(2) Durch den Vergleich von geeigneter Versuchs- und Kontrollgruppe im Experiment wissen wir, dass die Theorien H_1,...,H_n keine wahre Beschreibung der Ursachen von E darstellen.
(3) Eine probabilistische Falsifikation der Nullhypothese H_0 durch ein signifikantes Ergebnis in unserem Experiment schaltet außerdem auch noch H_0 aus.
(4) *Schlussfolgerung*: Unsere Theorie H ist vermutlich eine wahre Beschreibung der Ursachen von E.

Die wichtigste Frage an die eliminative Induktion ist natürlich immer, ob wir wissen und wie wir das begründen können, dass unsere Liste schon alle relevanten Konkurrenzhypothesen enthält. Im Falle der kontrollierten Experimente versuchen wir durch die Randomisierung die Konkurrenzhypothesen auszuschalten, ohne dass wir sie im Einzelnen überhaupt kennen müssen oder aufzählen können. Sobald uns das gelingt, sind wir also fein raus. Doch das Verfahren ist nicht immer unproblematisch, wie wir in Kapitel 7 sehen werden.

In anderen Fällen sind wir auf weiteres Hintergrundwissen angewiesen. Es handelt sich dabei meist um allgemeine *kausale Annahmen* in einem bestimmten Bereich darüber, welche Ursachentypen hier überhaupt am Werke sind. Für die Medizin wurde das in dem oberen Beispiel deutlich. Aus der Physik wissen wir, dass es vier Grundkräfte gibt, die in unserer Welt wirken, die damit einen gewissen Rahmen dafür abstecken, welche Hypothesen noch plausibel in konkreten Fällen sind. Selbst für die Sozialwissenschaften haben wir eine relativ kleine Liste an Ansätzen, welche Faktoren wie etwa intentionale Zustände, Emotionen oder physiologische Zustände auf der Mikroebene wirken. Auch für die Makroebenen lassen sich dann gewisse Einschränkungen

erkennen, die etwa durch die Forderung nach einer Mikrofundierung noch deutlicher werden (vgl. Bartelborth 2007, Kap. V), die uns Hinweise dafür geben können, dass wir die plausibelsten Kandidaten für eine Beschreibung der Ursachen bereits alle aufgelistet haben. John Norton (2003) spricht für seine Konzeption materieller Induktion davon, dass wir auf *lokale materielle Annahmen* für unsere induktiven Schlüsse vertrauen müssen. So können wir diese meist bereichsspezifischen grundlegenden Annahmen beschreiben. Empiristen würden sie vermutlich als metaphysische Annahmen einordnen, da es sich um sehr grundlegende kausale Annahmen darüber handelt, wie unsere Modelle eines Bereichs beschaffen sein müssen.

Wir werden aber sehen, dass letztlich alle Ansätze induktiven Schließens auf derartige Listen angewiesen sind. Für die Bayesianer argumentierte Hawthorne (1993) dafür, dass das bereits im Bayesianismus angelegt ist, während Earman (1992) explizit verlangt, dass die Bayesianer stärker die Techniken der eliminativen Induktion berücksichtigen sollten. Die Bayesianer haben jedenfalls einen sehr einfachen Trick entwickelt, um zu einer vollständigen Liste von Hypothesen zu gelangen. Wenn unsere Liste L = H_1,...,H_n von einander ausschließenden Hypothesen womöglich noch nicht vollständig ist, ergänzen sie sie um die sogenannte »*Catch-all*«*-Hypothese* $H^* \equiv \neg(H_1 \vee \ldots \vee H_n) \equiv \neg H_1 \,\&\, \ldots \,\&\, \neg H_n$). Das ist allerdings genau genommen keine eigenständige *substantielle* Hypothese und sie liefert insbesondere keine objektiven Wahrscheinlichkeiten bzw. Likelihoods für bestimmte Ergebnisse. Daher handelt es sich mehr um einen formalen Trick, der uns nicht wirklich weiterhilft (vgl. dazu auch Kap. 5). Unsere Einschätzung, dass unsere Liste bereits vollständig ist, lässt sich dann auch so ausdrücken, dass wir H^* nur noch eine geringe Wahrscheinlichkeit wahr zu sein zubilligen. Doch das besagt nur, dass eine unserer Hypothesen $H_1,\ldots,H_n$ vermutlich wahr ist und daher $H_1 \vee \ldots \vee H_n$ eine hohe Wahrscheinlichkeit aufweist.

Einen anderen Weg gehen etwa die *Likelihoodisten*, die aus der Not eine Tugend machen. Sie ziehen sich darauf zurück, dass wir immer nur *komparative Bestätigungen* erzielen können. Unsere Daten E liefern also immer nur Aussagen der Form $B(H_1,H_2{:}E)$, d.h., E bestätigt H_1 mehr als H_2 und das wird eventuell noch quantifiziert, aber es wird nicht mit der Behauptung verbunden, dass damit H_1 die am besten

bestätigte Hypothese überhaupt ist. Dabei stützen sie sich (wie der Name schon andeutet) auf die Likelihoodquotienten $P(E|H_1)/P(E|H_2)$ oder den Logarithmus davon. Doch wenn wir uns auf diese komparativen Bestätigungen in konkreten Entscheidungen stützen müssen, bleibt uns letztlich nichts anderes übrig, als nach der Hypothese zu suchen, die gegenüber allen anderen uns gerade bekannten Konkurrenzhypothesen aus diesem Bereich jeweils besser bestätigt ist. Natürlich ist sie unser bester Tipp, wenn dann noch gewisse Randbedingungen erfüllt sind, die wir im Kapitel über den Schluss auf die beste Erklärung erörtern werden.

Überhaupt ist es meine Strategie zu argumentieren, dass die eliminative Induktion eine Unterform und somit ein Teil des abduktiven Schließens darstellt (s. Kap. 4). Bird (2006) hofft dagegen noch, dass mehr oder weniger alle abduktiven Schlüsse dem einfacheren Schema der eliminativen Induktion gehorchen, er übersieht dabei aber, dass die Falsifikationen unserer wissenschaftlichen Theorien keineswegs so einfach sind, wie Popper sich das vorstellte, sondern eher weiche Falsifikationen aufgrund von hartnäckigen Erklärungsanomalien darstellen. Wir kommen also schließlich nicht darum herum, einen genauen Vergleich der Erklärungskraft der einzelnen Theorien vorzunehmen, und damit befinden wir uns mitten im abduktiven Schlussverfahren.

3.4 Die hypothetisch-deduktive Theorienbestätigung

Wenn es uns nicht gelingt, alle Konkurrenztheorien zu falsifizieren oder uns einen Überblick über alle Konkurrenzhypothesen zu verschaffen, zeigt sich ein Problem der eliminativen Induktion. Unsere Daten können intuitiv trotzdem noch eine bestimmte Theorie bestätigen, doch die eliminative Induktion wird diesen positiven Aspekten des induktiven Schließens nicht gerecht, die sich aus dem Eintreffen der Vorhersagen einer Theorie ergeben können. Wir müssen eine Theorie nicht unbedingt indirekt bestätigen, indem wir ihre Konkurrenten schwächen oder eliminieren, sondern können durchaus ebenso ihre zutreffenden Vorhersagen als direkte positive Bestätigungen der Theorie betrachten. Das ist eine erfolgreiche Praxis der Wissenschaften. Aus der allgemeinen Relativitätstheorie folgte z.B. die *überraschende Vorhersage*, dass das

Licht im Schwerefeld der Sonne eine Ablenkung erfährt. Das konnte einige Jahre später im Rahmen einer Sonnenfinsternis tatsächlich bestätigt werden und wurde als enorm wichtiges Indiz für die Richtigkeit der Relativitätstheorie eingestuft. Das zeigt die Grundidee des hypothetisch-deduktiven Bestätigungskonzeptes. Danach wird eine Theorie durch ihre (deduktiv abgeleiteten) Vorhersagen bestätigt, wenn diese sich als wahr erweisen.

Allerdings hatten wir bereits gesehen, dass wir für deduktive Schlüsse aus Theorien auf Beobachtungsaussagen meistens darauf angewiesen sind, weitere *Hilfsannahmen* hinzuzunehmen. Diese Hilfsannahmen sollten ihrerseits bereits induktiv gut gestützt also plausibel sein und natürlich nicht allein schon unsere Beobachtung implizieren. Das zeigt, wie wir die Grundidee der hypothetisch-deduktiven Theorienbestätigung noch ergänzen müssen. Popper könnte ohne die Forderung der Plausibilität sonst einwenden, dass wir z.B. mit entsprechenden Ad-hoc-Annahmen Vielerlei ableiten können, das nicht wirklich Auskunft über unsere Theorie gibt.

Schema der hypothetisch-deduktiven Theorienbestätigung

(1) Wir starten mit einer Theorie T und plausiblen Hilfsannahmen A.
(2) Wir schlussfolgern deduktiv bestimmte beobachtbare Resultate E aus T und A: T & A $\Rightarrow$ E, aber es gilt nicht: A $\Rightarrow$ E.
(3) Wir stellen fest, dass E tatsächlich beobachtet wird.
(4) *Schlussfolgerung*: Dann ist T durch E bestätigt bei gegebenem Hintergrundwissen A: B(T:E;A)

Trifft E nicht ein und wir können definitiv non-E feststellen, dann gilt T natürlich als geschwächt durch diese Beobachtung im Rahmen des hypothetisch-deduktiven Verfahrens. Wir müssen aber nicht gleich von einer Falsifikation sprechen.

Problematisch ist an dem hypothetisch-deduktiven Bestätigungsverfahren vor allem, dass die Ableitung von Vorhersagen noch keinen *inhaltlich relevanten Zusammenhang* zwischen Theorie und Daten herstellen muss. Wenn E aus T herleitbar ist, so ist es ebenso aus der

Konjunktion T & H mit beliebigem H deduzierbar. Diese konjunktive Theorie würde daher ebenso durch E bestätigt. Also würde ein beliebiges H durch E zumindest *mitbestätigt*. Das sollte eigentlich nur der Fall sein, wenn ein bestimmter relevanter Zusammenhang zwischen H und E besteht. Wenn also T eine medizinische Theorie wäre und H eine astrologische, könnten bestimmte Heilungserfolge, die T bestätigen, zugleich unsere astrologische Theorie H mitbestätigen. Das darf natürlich nicht passieren. Hier müssen wir eine weitere Relevanzbeziehung zwischen der bestätigenden Theorie und den Daten einfordern (s.u.).

Im Schluss auf die beste Erklärung bzw. der Abduktion wird zu diesem Zweck später verlangt, dass die bestätigte Theorie H das Datum E *erklären* muss, um durch E bestätigt zu werden. Außerdem können wir an unsere Forderung aus dem ersten Kapitel denken, dass nur *nomische* Konditionale durch ihre Instanzen zu bestätigen sind. Die Abduktion, die im Zentrum von Kapitel 4 stehen wird, vereint schließlich alle Ideen der zuletzt genannten Induktionsverfahren. Doch bevor wir uns den weiteren Problemen des hypothetisch-deduktiven Schließens und schließlich dem Schluss auf die beste Erklärung zuwenden können, möchte ich noch kurz Hempels Bestätigungskonzeption vorstellen, die zumindest einen Ausweg aus einigen Schwierigkeiten des hypothetisch-deduktiven Schließens bieten kann.

3.5 Hempels Instanzenkonzeption der Theorienbestätigung

Die formalen Konsequenzen dieser Bestätigungsansätze sind bereits 1974 durch Wolfgang Lenzen ausführlicher untersucht worden. Insbesondere hat er zahlreiche Anforderungen an qualitative Bestätigungskonzeptionen zusammengetragen und dann bewiesen, dass viele davon nicht wirklich im Paket zusammenpassen. Trotzdem sind einige dieser Anforderungen sicher hilfreich, um sich darüber klar zu werden, was wir von einer solchen Konzeption der Theorienbestätigung möglicherweise erwarten dürfen. Daher möchte ich eine Auswahl davon kurz vorstellen, wobei mit H die Hypothesen bezeichnet werden und mit E die Daten

(auf die Angabe weiteren Hintergrundwissens wird der Übersichtlichkeit halber verzichtet):

Hempels Adäquatheitsbedingungen für Bestätigungsbeziehungen
1. Spezielle Konsequenzbedingung
Wenn B(H:E) und H⇒H*, dann gilt auch: B(H*:E).
2.* Konverse Konsequenzbedingung
Wenn B(H*:E) und H⇒H*, dann gilt auch: B(H:E).
3. Konjunktionsbedingung
Wenn B(H:E) und B(H*:E), dann gilt: B(H & H*:E).
4.* Konverse Konjunktionsbedingung
Wenn B(H & H*:E), dann gilt: B(H:E) und B(H*:E).
5. Äquivalenzbedingung
Wenn B(H:E) und H⇔H*, dann gilt: B(H*:E).
6. Direkte Konsistenzbedingung
Wenn B(H:E), dann gilt nicht B(¬H:E).
7. Folgerungsbedingung
Wenn E⇒H, dann gilt: B(H:E).
8.* Konverse Folgerungsbedingung
Wenn H⇒E, dann gilt: B(H:E).

Die meisten dieser Adäquatheitsbedingungen stammen direkt von Hempel. Eine Ausnahme bilden die Bedingungen (4) sowie die Bedingungen (2) und (8), die von ihm diskutiert aber abgelehnt wurden. Die Ablehnung der Bedingung (8) zeigt sogleich, dass Hempel kein Freund des hypothetisch-deduktiven Ansatzes war, sondern einen dazu eigenständigen Weg vorschlug, der die anderen Adäquatheitsbedingungen erfüllen sollte. Bedingung (1) sieht zunächst einmal sehr plausibel aus, denn mit der Bestätigung einer Hypothese H sollte man natürlich zugleich auch die Abschwächungen H* der Hypothese bestätigt haben. Doch das Problem ist, dass sie nicht zu Bedingung (2) passt, diese aber vom hypothetisch-deduktiven Ansatz in jedem Fall erfüllt wird. Daher mussten die Vertreter dieses Ansatzes die an sich plausiblere Bedingung (1) ablehnen.

Diese Unverträglichkeit wird schnell deutlich, wenn wir annehmen, dass wir für unsere Hypothese H* eine Bestätigung E besitzen (also B(H*:E)) und wir nun eine beliebige Hypothese H' konjunktiv zu H*

hinzufügen. Dann gilt nach (2), dass auch diese Konjunktion durch E gestützt wird: B(H* & H':E). Daraus folgt aber nach (1): B(H':E). Da H' beliebig gewählt war, haben wir nachgewiesen, dass jede beliebige Hypothese durch unser Datum E bestätigt wird. Das sollte natürlich unter keinen Umständen der Fall sein und bedeutet, dass (1) und (2) nicht miteinander verträglich sind. Leider erfüllt der hypothetisch-deduktive Ansatz wie gesagt automatisch (2) und daher müssen wir die eigentlich plausiblere Anforderung (1) zurückweisen oder sie muss entsprechend modifiziert werden, wenn wir am klassischen DH-Ansatz festhalten wollen.

In unserem Beispiel haben wir auch gleich einen Verstoß des hypothetisch-deduktiven Ansatzes gegen (4) zu verzeichnen, der gerade unser Problem auslöst. Verstöße gegen (4) werden auch als »tacking«-Paradox oder als *Problem der irrelevanten Konjunktionen* bezeichnet. Das besagt, ich kann zu einer bestätigten Hypothese einfach beliebige Bestandteile konjunktiv hinzufügen und die resultierende Konjunktion wird immer noch bestätigt, obwohl das noch keineswegs bedeutet, dass beide Konjunkte bestätigt werden. Das wünscht man sich nicht, weil somit eigentlich irrelevante Hypothesen in den Genuss einer Bestätigung kommen; es ist aber für das hypothetisch-deduktive Bestätigungsverfahren offensichtlich der Fall, da die deduktive Logik monoton ist. Klar dürfte jedenfalls wieder unsere Konjunktionsbedingung (3) sein, während (4) wünschenswert erscheint, aber von vielen Bestätigungsverfahren, wie dem hypothetisch-deduktiven, eben leider intuitiv nicht erfüllt wird. Wir werden etwa im Bayesianismus auf das Problem der irrelevanten Konjunktionen wieder zurückkommen.

Die spezielle Konsequenzbedingung (1) ist auch für die Anwendung unserer wissenschaftlichen Theorien von entscheidender Bedeutung. Wenn wir etwa annehmen, wir hätten eine Theorie T gut genug bestätigt, um sie zu akzeptieren und können (anhand unseres akzeptierten Hintergrundwissens) aus T bestimmte Vorhersagen E deduktiv ableiten, dann sollten wir damit auch über gute Gründe verfügen, dass E auftreten wird. Nur so können wir unser wissenschaftliches Wissen auch gewinnbringend einsetzen. Die Rechtfertigung für T sollte sich dafür auf

die Konsequenzen von T übertragen. Diese Idee besitzt zudem einige Anwendungen in der Philosophie (vgl. Moretti & Tommaso 2013).

Allerdings gibt es dazu Gegenbeispiele wie das folgende Zwillingsbeispiel von Crispin Wright:

> Antonia und Berta seien gemäß unserem Hintergrundwissen eineiige, identisch aussehende Zwillinge und es sei: (E) Diese Frau sieht aus wie Antonia. (P) Diese Frau ist Antonia. (Q) Diese Frau ist nicht Berta. E bestätigt in diesem Fall P und Q folgt aus P, aber E scheint nicht Q zu bestätigen. Das liegt daran, dass die beiden Zwillinge identisch aussehen und wir bisher keine Hinweise darauf haben, dass es sich um Antonia und nicht um Berta handelt.

Man kann diesen Fall nun wie folgt beschreiben: P = (a oder b) & (non-b). Tatsächlich bestätigt E nur den ersten Teil von P und nicht den zweiten Teil, der gerade mit Q übereinstimmt. Womöglich haben wir hier wieder einen Fall von irrelevanten Konjunktionen vor uns, in dem nicht alle Teile von P bestätigt werden, in dem also keine genuine Bestätigung von ganz P vorliegt, was uns die Probleme eines Übertragungsfehlers für die Rechtfertigung einbringt. Das sollten wir durch eine Konzeption von genuiner Bestätigung zu vermeiden versuchen, die nur dann von Bestätigung spricht, wenn alle Teile der Hypothese bestätigt werden (vgl. etwa Schurz 1991, Sprenger 2011).

Die Äquivalenzbedingung scheint wiederum recht selbstverständlich zu sein, aber wir werden im sogenannten *Rabenparadox* sehen, dass auch sie nicht unbedingt harmlos ist. Die Konsistenzforderung (6) hilft uns zu vermeiden, dass wir sagen müssen: E spricht *für* H und spricht zugleich *gegen* H. Was aber nicht ausgeschlossen wird, ist, dass ein Datum für zwei konkurrierende Hypothesen gleichzeitig spricht. Nehmen wir etwa an, wir hätten 4 Karten aus einem Kartenspiel gezogen und dazu zwei Hypothesen gebildet: 1. Hypothese H, wonach alle Karten Asse seien, und 2. Hypothese H*, wonach alle Karten Herz seien. Wenn wir nun erfahren, dass eine der Karten ein Herz-Ass ist, so werden dadurch beide Hypothesen ein Stück weit bestätigt, obwohl die Hypothesen einander ausschließen. Das muss also weiterhin gestattet sein, zumal wenn die beiden Hypothesen einen überlappenden »Erfolgsbereich« wie in unserem Beispiel aufweisen.

Die Folgerungsbedingung (7) sieht wieder sehr plausibel aus, denn deduktives Schließen ist eigentlich die stärkere Variante des Schließens und sollte als Grenzfall für das induktive Schließen erhalten bleiben. Die Bedingung (8) stellt dagegen gerade die Idee des hypothetisch-deduktiven Schließens dar und wird daher in dem vorhergehenden Ansatz akzeptiert, aber von Hempel aufgrund der erwähnten Schwierigkeiten abgelehnt.

Hempel (1943, 1945) versuchte nun zumindest die Adäquatheitsbedingungen ohne »*« mit seiner Konzeption zu erfüllen. Seine Grundidee ist leicht erläutert.

> **Hempels Instanzenbestätigung**: Ein Datum E bestätigt eine Hypothese H genau dann, wenn die Einschränkung H_D der Hypothese H auf die Objekte D, die in den Daten wesentlich vorkommen, deduktiv aus E folgt. Das heißt: $B_{Hempel}(H{:}E)$ gdw. $E \Rightarrow H_D$

Nehmen wir etwa an, unsere Datenaussage E sagt etwas über die Objekte aus D={a, b} aus. Dabei sollen nur die *wesentlich vorkommenden* Objekte gezählt werden. Das sind solche, die in jeder logisch äquivalenten Formulierung auch vorkommen. Sonst könnten wir z.B. $E^* \equiv E \,\&\, (Fc \vee \neg Fc)$ als neues Datum wählen, das auch noch c erwähnt, ohne aber etwas Substantielles über c zu sagen. Solche technischen Tricks werden also ausgeschlossen, und man konzentriert sich auf die wesentlich vorkommenden Objekte in D. Dann können wir die Einschränkungen auf D leicht bestimmen, indem wir den Allquantor durch eine Konjunktion ersetzen und den Existenzquantor durch eine Disjunktion. Ist etwa $H \equiv \forall x(Fx)$, so ist $H_D \equiv Fa \,\&\, Fb$ und für $H \equiv \exists x(Fx)$ ist $H_D \equiv Fa \vee Fb$.

Die Grundidee der Instanzenbestätigung ist nun recht einfach. Wenn wir uns auf die wesentlichen Objekte aus unseren Beobachtungen beschränken, dann sollen die Daten die Behauptung der Hypothese wenigstens für diese Daten nachweisen. Dabei zeigen sich schon erste Probleme, wenn wir mit Theorien arbeiten, die theoretische Terme enthalten, die wir in den Datenbeschreibungen so überhaupt nicht wiederfinden. Aber wir beschränken unsere Aufmerksamkeit zunächst mehr auf die einfacheren Hypothesen. Leider gibt es selbst dafür bereits technische Probleme, die Hempel noch umgehen musste.

Betrachten wir etwa die beiden Hypothesen H ≡ ∀x(Fx) und H* ≡ ∀x(Fx) & Fc. Die sind offensichtlich logisch äquivalent und sollten daher nach der Äquivalenzbedingung gleichbehandelt werden, was aber für Hempels Konzeption bisher noch nicht der Fall ist, denn das Datum E ≡ Fa & Fb bestätigt zwar H, aber nicht H*, weil H^*_D nach unserer bisherigen Beschreibung gerade Fa & Fb & Fc ist. Wir müssen also noch solche Hypothesen wie H* verbieten oder die Einschränkung der Hypothese H* geeignet definieren (auf a und b beschränkt) oder einen entsprechenden Weg mit dem gewünschten Ergebnis wählen, der logisch etwas umständlicher aussieht wie etwa bei Lenzen (1974, 72). Diese technischen Details möchte ich nicht weiter verfolgen, da die Idee von Hempel bereits hinreichend klar geworden sein dürfte und sie nicht mehr im Zentrum der heutigen Aufmerksamkeit steht.

Jedenfalls beschreibt die hempelsche Instanzenbestätigung einen Ansatz, der zumindest bestimmte Fälle der oben genannten Probleme mit Konjunktionen vermeiden hilft. Haben wir zwei Hypothesen in einer Konjunktion H ≡ ∀x(Fx) & ∀x(Gx), so wird H nur dann durch ein Datum E bestätigt, wenn beide Konjunktionsglieder durch E bestätigt werden. Gilt etwa E ≡ Fa, dann wird H_E nicht bestätigt, weil dazu H_E ≡ Fa & Ga deduktiv aus E folgen müsste, was nicht der Fall ist. Allerdings könnten immer noch Konjunktionsprobleme auftreten, wenn die beiden Hypothesen in H über unterschiedliche Individuenbereiche sprechen, wenn es sich also z.B. um geeignete Konditionale handelt, deren Antenzedensbereiche sich nicht überschneiden. In Sprenger (2010, 243f.) finden sich entsprechende Gegenbeispiele gegen die Hempel-Bestätigung. Also erfüllt selbst der hempelsche Ansatz die Forderung (4) nur zum Teil. Der Gewinn durch die Instanzenbestätigung ist daher vermutlich nicht so groß, dass er eine Wiederbelebung dieser Konzeption rechtfertigen würde. Daher versuchte man die Konjunktionsprobleme direkt im Rahmen des hypothetisch-deduktiven Ansatzes durch eine Art von Relevanzlogik anzugehen.

3.6 Probleme der hypothetisch-deduktiven Theorienbestätigung

3.6.1 Relevanzprobleme

Das erste Problem des hypothetisch-deduktiven Ansatzes ist bereits genannt worden. Er ist zu liberal, wenn es darum geht, was alles mitbestätigt wird durch ein Datum E. Das versucht man z.B. durch eine neue Form der *Relevanzlogik* zu beheben, nach der nur die Teile einer Hypothese bestätigt werden, die tatsächlich erforderlich sind, um das Datum abzuleiten, die also für die Ableitung tatsächlich relevant sind. Wenn wir die Prämissen als eine Konjunktion von einfachen Hypothesen (und einigen Hilfsannahmen) betrachten, werden dann nur die Hypothesen bestätigt, ohne die eine Ableitung der Datenaussagen nicht mehr gelingt. Diese Prämissen gelten somit als *relevant*.

Der hypothetisch-deduktive Ansatz scheint allerdings auch zu liberal zu sein, was die Daten selbst (also die Konklusionen) angeht. Wenn E aus H ableitbar ist, so ist auch E∨E* mit beliebigem E* aus H ableitbar und damit gilt: B(H:E∨E*). Das neue disjunktive Datum E∨E* ist aber schon dann nachgewiesen, wenn wir zeigen, dass E* wahr ist. Das hieße, dass das Vorliegen von E* bereits genügen würde, um H zu bestätigen: B(H:E*), was natürlich nicht der Fall sein sollte, da E* in überhaupt keiner Verbindung zu H stehen muss. Auch hier kann eine Relevanzlogik helfen, die nur noch solche Daten als bestätigend zulässt, die tatsächlich zu den direkten Konsequenzen der Theorie gehören, bzw. die zum direkten Gehalt der Theorie gehören. Setzen sich die Konklusionen etwa aus einer Disjunktion von einfachen Beobachtungsaussagen zusammen, so sind nur die Beobachtungen relevant und bestätigend, die auch für sich schon aus der Theorie ableitbar sind.

Es geht also zum einen um eine *Prämissenrelevanz* und zum anderen um eine *Konklusionsrelevanz*, die in unseren Ableitungen zu fordern ist. Eine Idee von Schurz (vgl. 1991; 1994; 2005) ist dazu, dass irrelevante Bestandteile dadurch gekennzeichnet sind, dass sie inessentielle Subformeln enthalten. Das sind Formelteile, die durch andere Formelteile ausgetauscht werden können, ohne dass sich an der Gültigkeit der

Ableitungen etwas ändert. Letztlich hat er diese Austauschbarkeitsforderungen allerdings auf die Prädikate in den Formeln bezogen:

> **Relevanz in logischen Schlüssen** (Schurz 2006, 107)
>
> (1) Die Konklusion K eines gültigen Arguments ist *relevant* gdw. es in K kein Prädikat gibt, das an einigen Vorkommnissen simultan durch ein beliebiges gleichstelliges Prädikat ersetzt werden kann und trotzdem der Schluss gültig bleibt.
>
> (2) Die Prämissenmenge P eines gültigen Arguments ist *relevant* gdw. es in P kein Prädikat gibt, das an einem einzelnen Vorkommnis durch ein beliebiges gleichstelliges Prädikat ersetzbar ist und trotzdem der Schluss gültig bleibt.

Eine ähnliche Idee findet sich auch bei Ken Gemes (1998). Das ist sicher ein großer Fortschritt, um solche Irrelevanzprobleme beim hypothetisch-deduktiven Schließen zurückzudrängen. Wir beschränken die Bestätigung einer Theorie T durch ein Datum E [B(T:E)] damit auf den Fall, dass T eine relevante Prämisse in der Ableitung der relevanten Konklusion E darstellt. In neuerer Zeit hat Jan Sprenger (2011, 2014) einen Lösungsansatz entwickelt, der sich zusätzlich noch auf die hempelsche Instanzenbestätigung stützt, um die Relevanzprobleme zu lösen, aber alle Ansätze weisen weiterhin noch bestimmte Probleme auf.

Leider ist der hypothetisch-deduktive Ansatz zudem mit weiteren Schwierigkeiten behaftet. Eine solche Problematik (auch für andere Induktionsansätze) nennt Schurz in (2005). Es stellt sich die Frage, ob sie überhaupt die gesuchte *Induktionseigenschaft* aufweisen (vgl. Kap. 3.6). Rein formal betrachtet lässt sich eine Hypothese der Art $H \equiv \forall xFx$ immer in zwei logisch unabhängige Hypothesen aufspalten: $H_1 \equiv Fa$ und $H_2 \equiv \forall x(x \neq a \rightarrow Fx)$. Das Datum $E \equiv Fa$ bestätigt dann H_1, aber nicht H_2 hypothetisch-deduktiv. Damit bestätigt E in diesem Rahmen nur genau den Teil der Hypothese, der logisch aus E folgt, aber besitzt darüber hinaus keine induktive Kraft, um ebenfalls den Rest der Hypothese zu stützen. Wir haben es dann wieder mit dem Problem der *deduktiven Teilbestätigung* zu tun, das wir schon im ersten Kapitel beschrieben haben.

Für das induktive Schließens erwarten wir aber eine weitergehende Hypothesenbestätigung. Im hempelschen Instanzenansatz wird dieses Problem noch deutlicher. Der Ansatz setzt sogar explizit darauf, dass der Teil der Hypothese bestätigt wird, in dem er aus den Daten abgeleitet wird, der nur über die Daten spricht. Über den *Rest* der Hypothese machen wir damit keine weitere Aussage. Doch gerade das ist für das induktive Schließen die spannende Frage: Inwieweit werden alle Teile der Hypothese durch unsere Daten mitbestätigt? Hier sollte also z.B. sichergestellt sein, dass unsere Hypothesen nicht so einfach zerlegt werden können und dann immer noch entsprechende induktive Stützung erfahren, denn sonst sind es genau genommen keine wirklich induktiven Bestätigungsverfahren mehr. In gewisser Weise geht es dabei wieder um eine Art von Relevanzproblem, denn die Frage ist, welche Relevanz bzw. welchen Zusammenhang bestimmte Daten für eine Hypothese aufweisen. Sagen sie uns nur etwas über den direkt betroffenen Teil der Hypothese oder geht ihre Bedeutung darüber hinaus, weil sie auf einen grundlegenden nomischen Zusammenhang verweisen? Auf diese Frage komme ich immer wieder zurück, da sie leider eine wesentliche Problematik für praktisch alle Induktionsverfahren darstellt.

Eine andere Lösungsidee für das letzte Relevanzproblem ist jedenfalls darin zu sehen, dass wir genaugenommen nomische Konditionale der Form $\forall x(Fx \gg Gx)$ bestätigen müssen (s. Kap. 1.6.1). Damit verbinden wir die Vorstellung, dass die Eigenschaftsinstanzen von Fs notwendigerweise solche von Gs nach sich ziehen oder die beiden Eigenschaften nicht zusammenhängen. Es kann geradezu zu den *Identitätsbedingungen* von F gehören, dass F das kausale Vermögen aufweist, G hervorzubringen. Diese Idee von *Eigenschaften als kausalen Kräften* – die heute von vielen Wissenschaftstheoretikern (u.a. von Esfeld 2008a, 2010, Bird 2007 und Chakravartty 2007) diskutiert wird – kann unsere Induktionsschlüsse womöglich besser repräsentieren als die einfachen Konditionale. Gibt es nämlich viele zusammen auftretende Instanzen von Fs und Gs, so deutet das darauf hin, dass tatsächlich nicht nur ein zufälliges Zusammentreffen vorliegt, sondern eine entsprechende notwendige und kausale Verknüpfung zwischen Fs und Gs besteht. Erst dadurch begründen einige gemeinsame Fs und Gs schließlich die Annahme, dass

wir diesen Zusammenhang auch in anderen Fällen wiederfinden werden (s. dazu Kap. 3.7).

Wenn wir feststellen, dass viele elektrisch geladene Körper Kräfte auf andere elektrisch geladene Körper nach bestimmten Regeln ausüben, dann vermuten wir, dass eine solche notwendige Verbindung vorliegt und wir daher auch in anderen Fällen für elektrisch geladene Körper entsprechende Kräfte finden werden. Eine bloß zufällige Korrelation dieser beiden Eigenschaften wäre dafür aber zu wenig. Leider finden wir im materialen Konditional keine Hinweise auf diese Zusammenhänge. Sie scheinen für das hypothetisch-deduktive Schließen auch nicht gefordert zu werden, worin sich m.E. eine Schwachstelle des Ansatzes zeigt. Einen solchen HD-Ansatz für nomische Konditionale zu formulieren, bleibt aber eine Aufgabe für die zukünftige Forschung.

3.6.2 Das Problem der theoretischen Terme

Eine andere Schwierigkeit des ganzen Bestätigungsverfahrens bilden die sogenannten *theoretischen Terme*. Jede fortgeschrittene wissenschaftliche Theorie führt neue Begriffe ein, die sich auf unbeobachtbare Objekte oder Eigenschaften beziehen wie *Neutrinos* oder *elektromagnetische Felder* oder die *soziale Kohäsion* in einer Gesellschaft bzw. den *Intelligenzquotienten* von Personen etc., um damit bestimmte Phänomene erklären zu können. Wenn wir nun aus einer solchen Theorie einige Aussagen ableiten, wie dass die Emission eines Neutrinos oder ein bestimmtes elektrostatisches Feld zu erwarten ist oder dass die soziale Kohäsion durch bestimmte Maßnahmen gesteigert wird, dann liefern uns diese Ableitungen zunächst keine direkt durch Beobachtung entscheidbaren Aussagen und daher keine Ableitungen von Daten. Wie kommen wir dann zu beobachtbaren Resultaten bzw. Datenaussagen bzw. wie werden wir die theoretischen Terme wieder los?

Hier möchte ich mich auf die *Zweistufentheorie* der theoretischen Terme der logischen Empiristen stützen, die ich kurz erläutern muss. Danach bestehen entwickelte Theorien vor allem aus zwei unterschiedlichen Typen von Aussagen (rein theoretischen Gesetzen und gemischten Brückenprinzipien s.u.), die beide einen gesetzesartigen Charakter haben, aber eine etwas unterschiedliche Funktion. Dazu wird zunächst das

Vokabular in zwei Klassen aufgeteilt, nämlich die *Beobachtungsbegriffe* und die übrigen Begriffe, die als *theoretische Terme* bezeichnet werden. Beobachtungsbegriffe sind von dem Typ, dass sich für einfache Aussagen Ba, die nur den Beobachtungsbegriff B und den Namen für ein Objekt a enthalten, im Prinzip mit Hilfe einer entsprechenden Wahrnehmung entscheiden lässt, ob sie wahr sind oder nicht.

Schurz (2006, 61) verwendet noch eine etwas andere Charakterisierung, wonach Beobachtungsbegriffe dadurch ausgezeichnet sind, dass sie unabhängig von unserem Hintergrundwissen *ostensiv* (also durch hinweisende Gesten) erlernbar sind. Auch wenn diese Abgrenzungen alle etwas vage bleiben, sollten i.w. klar sein, was wir unter Beobachtungsbegriffen verstehen wollen. Sie beziehen sich auf das, was wir direkt beobachten können, ohne dafür bereits Instrumente einzusetzen. Als weitere *Daten* habe ich immer schon interpretierte Beobachtungen zugelassen, bei denen aus den direkten Beobachtungen bereits weitergehende Schlussfolgerungen gezogen wurden, aber wir müssen die grundlegende Problematik dabei zumindest einmal ansprechen (und werden u.a. im Rahmen des Bayesianismus noch einmal darauf zurückkommen, wie wir auch die Zuverlässigkeit der Daten beurteilen können).

Nach Ansicht der frühen logischen Empiristen müssten die theoretischen Terme idealerweise mit Hilfe von Beobachtungsbegriffen *definiert* werden, damit sie eine entsprechende (empirische) Bedeutung für uns besitzen und wir theoretische Aussagen so überprüfen können. Im *sogenannten Operationalismus* wird etwa verlangt, dass sie zumindest eine operationale »Definition« aufweisen, d.h., dass wir zumindest für jeden theoretischen Term ein Messverfahren angeben können, das es erlaubt, den Term in jeder Situation zu messen, also den Wert zu bestimmen, den die Größe annimmt, bzw. für qualitative Terme zu bestimmen, ob der Ausdruck korrekt auf die Situation angewandt werden kann oder nicht.

Doch Carnap (1968, Teil V) hatte schon bald eingesehen, dass derartige Forderungen zu weit gehen und die Realität der Wissenschaften nicht treffen. Eine echte Reduktion auf die Beobachtungsbegriffe ist für die meisten theoretischen Terme nicht zu erwarten und entspricht auch nicht ihrer Funktion in der Theorienbildung. Deshalb entwarf er die deutlich schwächeren Anforderungen der *Zweistufentheorie*, nach der

wir zunächst die zwei Stufen von Begriffen und Aussagen unterscheiden: die *Beobachtungsebene* und die *theoretische Ebene*. Zwischen denen gibt es allerdings Verbindungen in Form sogenannter *Brückengesetze*, die Ausdrücke aus beiden Ebenen enthalten. Ein Beispiel dafür finden wir im Kraftgesetz von Lorentz, nach dem die Kraft auf einen elektrisch geladenen Probekörper mit Ladung q in einem elektrischen Feld E gerade F = qE (als Vektorgleichung) beträgt. Dabei ist das elektrische Feld E unser theoretischer Term und zumindest die Kraft F in dem Gesetz ist beobachtbar oder wenigstens durch weitere Brückengesetze schließlich mit beobachtbaren Größen wie der Auslenkung einer Federwaage verbunden. Vermutlich ist die Ladungsgröße q ebenfalls als theoretischer Term zu betrachten, der u.a. durch das Kraftgesetz, aber vielleicht noch andere Gesetze etwa aus dem Bereich der elektrochemischen Zusammenhänge mit Beobachtungsgrößen verknüpft werden kann.

Die (partielle) *Bedeutung* eines theoretischen Terms ergibt sich im Rahmen der Zweistufentheorie durch seine Stellung sowohl in rein theoretischen Gesetzen (in unserem Beispiel des elektrischen Feldes etwa in den maxwellschen Gleichungen) sowie durch seine Verbindung in Brückengesetzen mit beobachtbaren Größen (in unserem Beispiel also durch das lorentzsche Kraftgesetz). Erst durch solche Brückengesetze sind die theoretischen Größen in bestimmten (einzelnen) Anwendungen der Theorie messbar, oder es lassen sich aus theoretischen Aussagen zumindest mit ihrer Hilfe Beobachtungsaussagen ableiten, die wir dann überprüfen können.

Beispiel: Klassische Elektrodynamik. Betrachten wir das oben genannte Beispiel der klassischen Elektrodynamik (ED) noch etwas genauer. Beschränken wir uns dabei auf den Fall der Elektrodynamik im leeren Raum, um die Komplikationen, die sich für elektromagnetische Felder in Materie ergeben, hier zu vermeiden. Die Gesetze der Theorie – die Maxwellschen Gleichungen – geben die Interaktionen von elektrischem Feld **E**, magnetischem Feld **B** (oder der magnetischen Flussdichte) und den Ladungen der Dichte in Form von partiellen Differentialgleichungen erster Ordnung an. (Vektoren werden dabei fett gesetzt.) Zur besseren Anschaulichkeit seien die Gleichungen kurz (mit vereinfachten Einheiten) angeführt:

Die Maxwellschen Gleichungen

$\text{div}\ \mathbf{E} = \varrho$ (die Ladungen sind die Quellen des elektrischen Feldes)

$\text{div}\ \mathbf{B} = 0$ (es gibt keine magnetischen Monopole)

$\text{rot}\ \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$ (Änderungen des magnetischen Feldes führen zu einem elektrischen Wirbelfeld.)

$\text{rot}\ \mathbf{B} = \mathbf{j} + \frac{\partial \mathbf{E}}{\partial t}$ (Elektrische Ströme j sowie der Verschiebungsstrom führen zu einem magnetischen Wirbelfeld)

Dabei lässt sich der (freie) elektrische Strom j noch definieren als: $\mathbf{j} = \varrho\mathbf{v}$, mit der Geschwindigkeit **v** für die (freien) Ladungen. Dann ist die Frage, welche Größen erst durch die Elektrodynamik eingeführt werden bzw. nicht unabhängig von ihr zu messen sind und damit die ED-theoretischen Größen darstellen. Das sind zunächst **E** und **B**, aber selbst für die Ladungsdichte ϱ ist die Sache nicht so klar. Rechnen wir sie hier aber einfach mal zu den für die Elektrodynamik vorgegebenen Größen und damit nicht-ED-theoretisch (vgl. a. Bartelborth 1987). Wenn wir nun aus der Elektrodynamik beobachtbare Schlussfolgerungen ziehen möchten, können wir uns etwa auf die Kräfte beziehen, die auf Probeteilchen mit Ladung q einwirken, die in ein elektromagnetisches Feld gebracht werden. Die Größe dieser Kraft wird durch das Lorentzkraftgesetz bestimmt, die in unserem Beispiel das wichtigste Brückengesetz darstellt:

Lorentzkraftgesetz: $\mathbf{F} = q\mathbf{E} + q\mathbf{v} \times \mathbf{B}$

Wenn wir annehmen, dass wir die Kraft **F** schon vortheoretisch zur Elektrodynamik messen können, liefert uns das Kraftgesetz nun eine schöne Verbindung zwischen den zwei theoretischen Größen **E** und **B** und zwei beobachtbaren Größen **F** und **v**.

Es zeigt sich zunächst, dass die theoretischen Größen nicht definierbar sind anhand der nichttheoretischen Größen. Allerdings findet man in manchen Physikbücher (z.B. Halliday et al. 2003, 644) den Vorschlag $\mathbf{E} = \mathbf{F}/q$ als Definition zu betrachten. Das bezieht sich dann aber schon nur noch auf die speziellen Situationen der Elektrostatik, in der $\mathbf{B} = 0$ ist, und kann daher nicht als Definition betrachtet werden. Außerdem liefert der Vorschlag auch nur Werte für die Stellen, an denen wir tatsächlich ein

Probeteilchen eingesetzt haben. Eine Definition verlangt dagegen, dass der definierte Ausdruck überall salva veritate durch den definierenden Ausdruck ersetzt werden kann. Davon sind wir hier weit entfernt.

Liefert der Vorschlag denn wenigstens ein *Messverfahren* für einfache Situationen der Elektrostatik – wenn er schon keine reduzierende Definition zustande bringt? Auch das ist keineswegs so unproblematisch. Das erste Problem haben wir schon benannt. Wir können bestenfalls einzelne Werte eines elektrischen Feldes bestimmen, das schließlich auf allen Raumpunkten eines Gebietes festgelegt wird. Ein weiteres Problem ist das folgende: Durch das Einbringen des Probeteilchens wird das vorliegende elektrische Feld bereits verändert. In welchen Fällen das Ergebnis trotzdem als approximativ korrekte Messung betrachtet werden kann, kann wiederum nur mit Hilfe der Elektrodynamik selbst bestimmt werden. Das Brückengesetz führt also in unserem Fall immerhin zu einem *approximativen Messverfahren* für bestimmte einfache Situationen.

Zur Funktion der Brückengesetze. Auf solche Brückengesetze sind wir also im Rahmen der hypothetisch-deduktiven Theorienbestätigung angewiesen, aber auch im Rahmen anderer induktiver Begründungsverfahren. Sie stellen erst die Verbindung von der rein theoretischen Ebene zu bestimmten Beobachtungen her. Wir können sie als Teile der Theorie betrachten oder zu den Hilfsannahmen rechnen, sollten uns aber auf jeden Fall ihrer besonderen Rolle bei Theorientests bewusst sein. In Kapitel 5.4.6 werden wir noch eine konkrete Anwendung der Zweistufenkonzeption auf den Begriff der *Propensität* kennenlernen, bei dem auch deutlich wird, dass wir noch weitere Unterbestimmtheiten für theoretische Terme zu akzeptieren haben.

Wir sollten die zwei Stufen (Beobachtungsebene–theoretische Ebene) durch mehrere Stufen ersetzen und davon ausgehen, dass der Aufbau der theoretischen Begriffe durch mehrere Theorien schrittweise geleistet wird, wie ich das für den Begriff des elektrischen Feldes schon angedeutet hatte: elektrisches Feld in der Elektrodynamik — Kraftbegriff in der Mechanik — Auslenkung einer Federwaage als basales Messverfahren. Das wird etwa im Rahmen der strukturalistischen Theorienauffassung so gesehen (vgl. Gähde 1983), soll hier aber nicht ausführlich erörtert werden.

Genauer gesagt müssten wir dafür ein Konzept von *T-Theoretizität* einführen, wonach ein Term theoretisch relativ zu einer bestimmten Theorie T ist, wenn er nicht bestimmbar ist innerhalb von Vortheorien zu T, die selbst näher an der empirischen Basis liegen, als es für T der Fall ist. Die Kraftfunktion ist dann zumindest mit Hilfe der mechanischen Vortheorien zur Elektrodynamik bestimmbar (vgl. a. Bartelborth 1987) und damit keine theoretische Größe für die Elektrodynamik.

Das Problem der theoretischen Terme hat jedenfalls wiederum deutlich gemacht, dass wir neben den Kerngesetzen einer Theorie auf weitere Annahmen angewiesen sind, um Datenaussagen ableiten zu können. Wie schon im Falsifikationismus spielen die Hilfsannahmen allgemein eine wichtige Rolle bei der Ableitung von Beobachtungsaussagen. Der Vorteil des hypothetisch-deduktiven Bestätigungsansatzes gegenüber der konservativen Induktion oder hempelschen Instanzenbestätigung ist jedoch, dass es überhaupt möglich ist, Theorien mit theoretischen Termen zuzulassen. Dabei bleibt wieder das Problem, dass wir auf *plausible* Hilfsannahmen angewiesen sind und sich mit der Hilfe geeignet ausgedachter Ad-hoc-Annahmen potentielle Misserfolge der Theorie zu einfach in erfolgreiche Bestätigungen verwandeln lassen.

Hier wird u.a. zu fordern sein, dass zumindest die betrachteten zusätzlichen Hilfshypothesen – wie die Brückenprinzipien selbst – wieder einen gesetzesartigen Charakter haben. Es zeigt sich zugleich die Problematik, dass es uns auch im Zusammenhang mit Brückengesetzen nicht immer gelingen muss, Beobachtungsaussagen aus einer Theorie abzuleiten. Das wurde nur im Operationalismus sichergestellt, aber nicht mehr innerhalb der Zweistufenkonzeption. Hier bleibt sicher Raum für weitere Debatten und mögliche Verschärfungen der Konzeption, wonach es uns zumindest in einzelnen Anwendungsfällen gelingen sollte, mit Hilfe der Brückengesetze eine Messung der theoretischen Terme durchzuführen.

3.6.3 Unterbestimmtheit und probabilistische Theorien

Ein weiteres grundlegendes Problem der hypothetisch-deduktiven Theorienbestätigung besteht darin, dass wir zwar deduktiv aus den Theorien (und gewissen Hilfsannahmen) auf die Daten schließen, aber aus den

Daten nur induktiv auf die Theorien. Es scheint zwar intuitiv naheliegend zu sein, dass erfolgreiche Vorhersagen einer Theorie für diese Theorie sprechen, andererseits bleibt ein wichtiges Problem außen vor: Wenn gilt T⇒E, so sagt das noch nichts darüber, ob es nicht viele andere Theorien T* gibt, für die das Gleiche gilt: T*⇒E. (Die Hilfsannahmen lassen wir hier einmal weg. In ihnen können sich unsere beiden Fälle natürlich ebenfalls noch unterscheiden.) An dieser Stelle tritt eine *Unterbestimmtheit der Theorie* durch die Daten besonders deutlich zu Tage. Unsere Bestätigung der Theorie T durch E sagt uns noch nichts über die Konkurrenz. Man kann sagen, ihr fehlt das *komparative Element*.

Haben wir etwa 20 plausible, untereinander inkompatible Theorien, die alle diese Ableitung gestatten, werden wir wohl kaum sagen können, dass T in diesem Fall wesentlich bestätigt wurde. Es fehlt an dieser Stelle eine Liste L von Theorien, bei der wir über Gründe verfügen, anzunehmen, dass sie alle plausiblen Konkurrenten enthält, die es ebenso gestatten, E abzuleiten. Dann könnten wir z.B. zumindest Vergleiche anstellen, welche der Theorien auf die gewagteren Hilfsannahmen angewiesen ist. Allerdings bleibt dann immer noch das Problem, dass es im hypothetisch-deduktiven Ansatz keinen guten Vergleichsmaßstab dafür gibt, welche der Theorien durch unsere Daten *besser* bestätigt werden. Das zumindest müsste ergänzt werden, wenn wir auf diesem Weg ein umfassendes Induktionsverfahren gewinnen wollen. Oder wir müssten auf den glücklichen Fall treffen, dass unsere Liste von vornherein nur eine Einer-Liste ist.

Das Unterbestimmtheitsproblem verschärft sich noch weiter, wenn wir probabilistische Theorien betrachten. Zunächst haben wir dort das Problem, dass keine konkreten Beobachtungsaussagen mehr ableitbar sind. Natürlich lassen sich gewisse probabilistische Aussagen (im Extremfall die Hypothese selbst) deduktiv ableiten. Aber das sind keine Datenaussagen. Wir können die Wahrheit dieser »Vorhersagen« nicht durch einfache Beobachtungen oder Messungen entscheiden. Wir müssen unser Verfahren an dieser Stelle also schon erheblich abwandeln.

Man könnte daran denken, ein Datum als Bestätigung einer probabilistischen Theorie zu betrachten, wenn es in den sogenannten *Akzeptanzbereich* A der Theorie fällt, auf den wir in Kapitel 6 noch genauer eingehen werden. Nehmen wir an, unsere Hypothese H besage,

dass die Wahrscheinlichkeit für Kopf bei einer bestimmten Münze gerade 0,6 sei. Dann konstruieren wir im der Nullhypothesen-Signifikanztests einen *Zurückweisungsbereich* Z – wie oben schon erwähnt – als Menge der Resultate, die zusammengenommen eine Wahrscheinlichkeit von unter 5% aufweisen, wenn wir einmal H als wahr annehmen: P(Ergebnis in Z|H). Für 10 Würfe wäre das etwa die Menge Z = {0,1,2,3} ∪ {8,9,10}. Dann sei A die dazu komplementäre Menge A = {4,5,6,7}. Man könnte sagen, dass ein Resultat aus Z klar gegen unsere Hypothese spricht. So hatten wir das oben schon als probabilistische Falsifikation gedeutet, während ein Resultat aus dem Bereich A mit unserer Hypothese *gut verträglich* ist. Doch wir werden in Kapitel 6 darauf zurückkommen, dass ein Resultat aus A bestenfalls eine *sehr schwache Stützung* für H darstellt, insbesondere, weil es hier sehr viele Hypothesen gibt, die ebenso zu bestimmten Resultaten aus Z passen. Stellen sich etwa 5 Köpfe ein, sind natürlich auch die Werte 0,45; 0,48; 0,55 etc. als Wahrscheinlichkeiten für Kopf geeignet. Jedes Mal liegt das Resultat wieder im Annahmebereich der Theorie. Wir erhalten so eine größere Unterbestimmtheit, weil sehr viele Theorien zu unseren Daten passen.

Während wir für probabilistische Theorien also ein probabilistisches Falsifikationsverfahren kennen (was jedoch auch problematisch ist, wie wir in Kapitel 6 sehen werden), gibt es kein sinnvolles zum hypothetisch-deduktiven Ansatz vergleichbares Bestätigungsverfahren für probabilistische Aussagen. Die Unterbestimmtheit wird zu groß, wenn wir alles als Bestätigung betrachten, was im Akzeptanzbereich der jeweiligen Theorien liegt. Man könnte daher zunächst sagen, dass sich für den Fall probabilistischer Hypothesen die Poppersche These von der Asymmetrie von Falsifikation und Verifikation wiederfinden lässt.

Der Likelihoodismus. Als Ersatz für eine direkte positive Bestätigung probabilistischer Hypothesen bietet sich somit eher ein *komparatives Verfahren* an, auf das ich in Kapitel 5 zurückkommen werde. Wenn wir zwei probabilistische Theorien T und T* haben, die beide E mit einer gewissen Wahrscheinlichkeit vorhersagen, dann können wir annehmen, dass der Likelihoodquotient der beiden Theorien bzgl. der Daten am ehesten das Ausmaß wiedergibt, indem E für T gegenüber T* spricht:

$$lq(T,T^*{:}E) = P(E|T) \;/\; P(E|T^*)$$

Das ist die Idee der *Likelihoodisten*, die dann noch gern zum Logarithmus des Likelihoodquotienten übergehen, da wir damit ein Maß symmetrisch um den Nullpunkt herum erhalten, das für bestimmte Daten sogar additiv ist.

$$LQ(T,T^*:E) = \log [P(E|T) \,/\, P(E|T^*)]$$

Besteht E etwa aus einer Konjunktion von Daten $E \equiv E_1 \,\&\, \ldots \,\&\, E_n$, die relativ zu unseren Theorien stochastisch unabhängig sind, dann gilt:

$$LQ(T,T^*:E) = LQ(T,T^*:E_1)+\ldots+LQ(T,T^*:E_n)$$

Für L(T,T*:E) gilt somit, dass es genau dann größer als null ist, wenn die Theorie T die Daten E »mehr voraussagt« als es T* tut, wobei hier schwächende Daten und stützende Daten sogar auf einfache Weise gegeneinander aufgerechnet werden. (Auch die Bayesianer setzen statt auf Signifikanztests auf den Bayes-Faktor, der für die objektiven Likelihoods mit dem Likelihoodquotienten übereinstimmt. Mehr dazu in Kapitel 6.4.2.)

Schurz spricht allgemein von der *Likelihood-Intuition* und wir werden in Kapitel 5 noch das sogenannte Likelihood-Gesetz kennenlernen, nach dem ein Datum umso mehr für T spricht, je größer die Wahrscheinlichkeit ist, die es dem Datum beilegt, je größer also P(E|T) ist. Der deduktive Fall ist dann der Grenzfall, dass $T \Rightarrow E$ gilt und daher P(E|T)=1 ist.

Gerade wenn wir an das Ideal *wissenschaftlichen Wissens* denken, lässt der hypothetisch-deduktive Ansatz noch zu viele Fragen offen. Vor allem die Unterbestimmtheitsproblematik wird hier nicht hinreichend gelöst. Man könnte sagen, die Existenz von Konkurrenztheorien, die das Datum ebenfalls abzuleiten gestatten, *unterminiert* unsere hypothetisch-deduktive Theorienbestätigung. Obwohl der Ansatz also zunächst intuitiv plausibel erscheint und viele Wissenschafter ihre Vorgehensweise so ähnlich darstellen, fehlt doch noch etwas zu einem plausiblen Induktionsverfahren. Das wird m.E. erst im abduktiven Schließen zu finden sein.

3.7 Die Induktionseigenschaft

Ein Problem für das hypothetisch-deduktive Schließen war darin zu sehen, dass nicht klar ist, wie es die Induktionseigenschaft gewährleisten kann. Hier geht es uns speziell um die Bestätigung von *Allhypothesen* ∀x(Hx) und die *Induktionseigenschaft* (IE), nach der sich die Bestätigung einer Hypothese H von bestimmten Objekten auf andere überträgt:

> **Die Induktionseigenschaft (IE)**: Es gibt für eine Hypothese H verschiedene Objekte a und b, für die gilt: B(Hb:Ha)

Es sollte doch zumindest für bestimmte Objekte a und b unserer Hypothese der Fall sein, dass ein Nachweis, dass für a die Hypothese gilt (Ha), eine Bestätigung dafür bietet, dass auch für andere Objekte b unsere Hypothese zutrifft (Hb). Wäre das nicht der Fall, läge keine genuin *induktive* Bestätigung unserer Hypothese mehr vor, sondern jedes Datum würde nur den Teil der Hypothese stützen, der deduktiv aus dem Datum folgt. Der Punkt ist der: Ist unser Datum etwa E ≡ Ha & Hb & Hc, so können wir eine Hypothese T ≡ ∀xHx zerlegen in T_1 und T_2 mit: T_1 ≡ Ha & Hb & Hc und T_2 ≡ ∀x(x∉{a,b,c} → Hx). Offensichtlich wird nur T_1 durch E bestätigt und das heißt, dass jede Stärkung von T_2 entfällt. Also findet eigentlich kein genuin induktives Schließen statt.

Die Überlegungen zur Metaphysik induktiven Schließens (s. Kap. 1.6.1) machten schon deutlich, dass wir die Induktionseigenschaften nicht in jedem Fall erwarten dürfen, sondern nur für bestimmte nomische Hypothesen, die wir als projizierbar einstufen. Das sind typischerweise solche Hypothesen, die ein *nomisches Konditional* T ≡ ∀i(Fa_i≫ Ga_i) beschreiben, bei dem es eine gesetzesartige Verbindung zwischen den Eigenschaften F und G gibt. Nur wenn unsere Hypothese also einen gesetzesartigen Charakter hat, können wir die bisherigen Daten als Indiz dafür werten, dass gerade zwischen F und G tatsächlich ein entsprechender nomischer Zusammenhang vorliegt und nur in dem Fall dürfen wir erwarten, dass der dann ebenfalls für andere Objekte wiederzufinden ist, die wir noch nicht untersucht haben.

Oder es muss zumindest eine Erklärungsverbindung im Sinne von Achinstein (2001) vorliegen, der für eine Bestätigungsbeziehung verlangt,

dass entweder Fa_i erklärt, warum Ga_i der Fall ist oder umgekehrt oder es zumindest eine Hypothese gibt, die beides zusammen erklärt. Die Anforderung war, dass wir zumindest im Hintergrund eine *gesetzesartige Beziehung* annehmen müssen, die wir anhand von einzelnen beobachteten Instanzen und weiterem Hintergrundwissen als begründet vermuten können und dann auf neue Instanzen anwenden dürfen.

Es sei F etwa die Eigenschaft, dass etwas ein Stück Metall sei und erhitzt werde, während G die Eigenschaft darstelle, dass sich der betreffende Gegenstand ausdehnt. Bisherige Beobachtungen erhitzter Metalle und unser Hintergrundwissen über die grundlegenden Gemeinsamkeiten von Metallen begründen die Vermutung, dass Metalle eine bestimmte *gesetzesartige Dispositionseigenschaft* aufweisen, nämlich sich auszudehnen, wenn sie erhitzt werden. Das ist eine charakteristische Eigenschaft für Metalle, die aus ihrer molekularen Struktur resultiert. Erst wenn ein solcher gesetzesartiger Zusammenhang bzw. ein entsprechender kausaler Mechanismus gegeben ist, begründen bisherige Beobachtungen unsere Vermutung, dass auch die nächsten Metalle, die wir erhitzen, sich wieder ausdehnen werden.

Nur in den Fällen, in denen wir also gewisse *nomische Muster* vermuten, bieten die bisherigen Daten zugleich einen Grund für eine Extrapolation. Sonst könnte bisher einfach nur ein zufälliges Zusammentreffen einiger Objekte, die zugleich F und G sind, beobachtet worden sein. Zufällig waren alle Ingenieure, die wir kennengelernt haben blond. Dann extrapolieren wir trotzdem nicht einfach auf die Haarfarbe der nächsten Ingenieure, die uns begegnen werden. Jedenfalls dann nicht, wenn wir nicht einen Grund zu der Vermutung haben, dass im Hintergrund ein kausaler Mechanismus wirksam ist, der etwa die blonden Männer geeigneter für den Ingenieursberuf macht, als es Männer mit anderen Haarfarben sind. Das zeigt wieder, dass die induktive Rechtfertigung keine rein syntaktische Beziehung zwischen Datenaussagen und Hypothesen darstellt, sondern eine weitergehende Beurteilung verlangt, ob es sich um eine gesetzesartige (und damit erklärende) Hypothese handelt oder nicht. Das meinten wir u.a. damit, dass Induktion *dreistellig* ist. Erst für nomische Konditionale sollten wir die Instanzenkonzeption oder den hypothetisch-deduktiven Ansatz in seiner Relevanzvariante als vertretbare Induktionsverfahren betrachten.

Wir können an dieser Stelle schon erste Hinweise erkennen, wie der Ansatz des abduktiven Schließens bzw. der Schluss auf die beste Erklärung (den wir uns als nächstes Induktionsverfahren genauer ansehen werden) das Problem der Induktionseigenschaft lösen kann. Damit eine Theorie T ein Datum E *erklärt*, müssen bestimmte metaphysische Anforderungen der genannten Art erfüllt sein. Es muss in T ein gesetzesartiger kausaler Zusammenhang beschrieben werden, der zusammen mit den konkreten Ursachen aufzeigt, wie es zu E kam. Ein solcher gesetzesartiger Zusammenhang (ein nomisches Muster) ist aber nicht mehr einfach in Teile zerlegbar, denn die stellen selbst kein nomisches Muster mehr dar, sondern haben nur noch den Charakter zufälliger Zusammentreffen. Erst das Vorliegen der generellen Zusammenhänge erklärt und nur diese nomischen Muster können auch als Ganzes durch die Daten gestützt werden. Das zeigt, wieso die Induktionseigenschaft in das abduktive Schließen geradezu eingebaut ist.

Die Argumentation für eine Extrapolation lautet dann so: Unser Hintergrundwissen sagt uns, dass, wenn überhaupt eine Beziehung zwischen F und G vorliegt, sie nomischen Charakter hat. Unsere bisherigen Daten sprechen in dem Fall dafür, dass die Eigenschaft F die Eigenschaft G in systematischer (kausaler) Weise hervorruft. Daher erwarten wir beim nächsten Auftreten von F, dass wiederum G vorliegt. Wir stützen uns also darauf, dass gewisse grundlegende Zusammenhänge zwischen bestimmten Eigenschaften in unserer Welt stabil bestehen bleiben und daher extrapolierbar sind, während es daneben viele zufällige Zusammentreffen gibt, die keinen Anlass zu einem solchen Fortschreiben geben. Das entspricht dem Unterschied zwischen zufälligen Korrelationen und kausalen Beziehungen. Nur die letzteren können wir für gute Vorhersagen einsetzen.

Wir suchen dazu in unserem Alltag ganz gezielt nach solchen nomischen Mustern, die auch Erklärungskraft besitzen. Wir tendieren sogar dazu, vorschnell solche Muster zu »erkennen«. Selbst in der Wissenschaft neigen wir dazu, aus Korrelationen vorschnell auf Kausalzusammenhänge zu schließen. Wir alle kennen die vielen Meldungen aus der Medizin, wo man bestimmte korrelative Zusammenhänge zwischen Ernährung und Krankheiten festgestellt hat, wonach etwa diejenigen, die Rotwein trinken, weniger Herzinfarkte haben als die Abstinenzler und dann bald

Rotwein als wirksamer Herzschutz verkauft wird. Psychologen haben diese Neigung zum Auffinden erklärender Muster genauer untersucht und konnten etliche Fehlschlüsse darauf zurückführen. Ihr Resümee ist:

> To live, it seems, is to explain, to justify, and to find coherence among diverse outcomes, characteristics, and causes. With practice we have learned to perform these tasks quickly and effectively.
> (Gilovich 1991: 22)

Unser in vielen Hinsichten recht erfolgreicher Umgang mit unserer Umwelt zeigt aber auch, dass wir dabei oft richtig liegen und durchaus beachtliche Vorhersageerfolge verbuchen können. Jedenfalls ist das hypothetisch-deduktive Schließen zumindest um eine geeignete metaphysische Komponente zu ergänzen. Erst wenn wir Grund zu der Annahme haben, dass gewisse nomische Muster vorliegen, sollten wir die Vorhersageerfolge zugleich als Indikatoren für die Wahrheit unserer Hypothesen betrachten und in weiteren Extrapolationen darauf (mit der nötigen Vorsicht bei induktiven Schlüssen) vertrauen.

Übrigens findet sich hier ebenso eine erste Antwort auf den Induktionsskeptiker. Wieso glauben wir, in bestimmten Fällen extrapolieren zu dürfen? Weil wir annehmen, dass es in unserer Welt stabile nomische Muster gibt und uns manche Daten Hinweise auf solche Muster liefern, so dass wir danach diese Muster für Vorhersagen einsetzen können. Die Überlegungen zur konservativen Induktion hatten schon gezeigt, dass wir nicht immer auf einfache Weise extrapolieren dürfen, aber in den Fällen, in denen wir eine kausale gesetzesartige Verbindung zwischen bestimmten Eigenschaften aufgedeckt haben, ist genau das unsere Berechtigung für zukünftige Fälle davon auszugehen, dass diese Verbindung weiterhin Bestand hat und wir daher schließen dürfen, wenn ein weiteres Objekt F aufweist, dass es dann ebenfalls G aufweist.

Es bleibt natürlich noch die schwierige Frage zu beantworten, wie wir Kenntnis der kausalen gesetzesartigen Zusammenhänge erlangen. Letztlich sind wir dafür natürlich wiederum auf bestimmte induktive Schlüsse angewiesen. Die haben allerdings einen speziellen Charakter und deshalb werde ich mich diesen Schlüssen im Kapitel 7 gesondert zuwenden. Wir sind auf ganz spezielle Annahmen über unsere

Welt angewiesen, um hier zumindest in einem ersten Schritt einfache Kausalbeziehungen aufdecken zu können, auf die wir uns dann in weiteren induktiven Schlüssen stützen dürfen, um zu weitergehenden Kausalhypothesen aufsteigen zu können.

Natürlich wird das alles einen hartleibigen Skeptiker nicht überzeugen können. Er antwortet darauf z.B.: Woher willst Du wissen, dass diese gesetzesartige Beziehung (selbst wenn wir einmal akzeptieren, dass sie heute gilt) weiterhin und insbesondere morgen Bestand hat? Sie könnte nur bis heute gelten, und ab morgen gelten ganz andere Naturgesetze. Und natürlich haben wir keine zwingenden Gründe für diese Art von Vorhersageverfahren und für das Fortbestehen der Gesetze bzw. der nomischen Muster in unserer Welt. Das sind grundlegende Annahmen unseres Modells der Welt, die sich bewährt haben, die aber gegen einen radikalen Skeptiker nicht zu verteidigen sind.

Besonders ein konsequenter Empirist wird sie als zu metaphysisch ablehnen, doch damit verliert er die Möglichkeit, zwischen sinnvollen Extrapolationen und unsinnigen zu unterscheiden in seinem Modell der Welt. Er kann so unsere Induktionspraxis nicht mehr angemessen erklären oder verstehen. Das sollte einen guten Grund für uns darstellen, ein metaphysischeres Bild der Welt zu akzeptieren, das besser zu unserer erfolgreichen wissenschaftlichen Praxis passt (vgl. dazu Esfeld 2008a). Oder wir sollten uns überhaupt nicht erst darauf einlassen, uns von Empiristen diktieren zu lassen, was als metaphysisch und damit problematisch zu gelten hat. In der Wissenschaft und bereits im Alltag gehen wir in unseren Annahmen über unsere Umgebung immer schon über das hinaus, was sich direkt durch Beobachtungen nachweisen lässt, und das zeigt in unserem Umgang mit unserer Umwelt große Erfolge. Insbesondere sind wir für Erklärungen (vgl. Bartelborth 2007) und jetzt auch für induktive Schlüsse immer wieder auf Annahmen über die kausale Grundstruktur in unserer Welt angewiesen, die selbst nur sehr indirekt empirisch zu testen sind.

Der humesche Induktionsskeptiker wird natürlich das Argument nicht gelten lassen, dass wir dieses Verfahren weiter einsetzen sollten, weil es uns in der Vergangenheit so gute Dienste geleistet hat. Unsere beste Erwiderung ist dann, dass es immer noch besser ist, einem Verfahren zu vertrauen, das in der Vergangenheit gut funktioniert hat und das uns sagt,

wie wir aus gemachten Erfahrungen lernen können, um damit zukünftige Entwicklungen vorherzuplanen, als sich den Standpunkt des Skeptikers zu eigen zu machen, wonach wir keinem Verfahren vertrauen dürfen und wir dann ganz ohne Leitschnur für unsere Entscheidungen dastehen. Immerhin *könnten* wir ja zumindest Recht behalten mit unserer Annahme, dass die Naturgesetze Konstanten in unserer Welt darstellen, die eine Extrapolation in die Zukunft erlauben. Für uns selbst besitzt diese Idee daher eine gewisse Überzeugungskraft und rundet unser Modell der Außenwelt in kohärenter Weise ab. Dafür können wir uns kaum die Anforderungen des radikalen Skeptikers zum Maßstab nehmen, sondern müssen an weniger radikale Herausforderungen denken.

Sind diese starken Annahmen aber tatsächlich erforderlich? Wir schließen doch selbst in den Fällen induktiv, in denen wir keine derartigen Annahmen voraussetzen. Denken wir an folgende kritische Beispiele: Wir schließen vom fallenden Barometer auf den herannahenden Sturm oder von gelben Fingern auf ein erhöhtes Lungenkrebsrisiko oder vom Fieber und Gliederschmerzen auf die Grippe. In all diesen Fällen stellt die erste Eigenschaft keine Ursache der zweiten dar. Allerdings stehen bestimmte kausale Beziehungen im Hintergrund dieser Schlüsse. Der fallende Luftdruck verursacht sowohl das Fallen des Barometers wie das Aufziehen des Sturms. Es gibt also eine erklärende Beziehung zwischen unseren beiden Tatsachen. Außerdem verlangen diese Schlüsse nach entsprechenden speziellen Voraussetzungen. Wir schließen anhand solcher *indirekten nomischen Muster* nur, wenn diese Voraussetzungen nach unserem besten Wissen tatsächlich vorliegen. Es darf z.B. keine *Intervention* am Barometer vorliegen, d.h., es darf nicht der Fall sein, dass das Barometer vom Luftdruck »abgekoppelt wurde«, indem es schlicht von Hand auf einen bestimmten Wert eingestellt wurde. Sollte das der Fall sein, würden wir es nicht mehr für entsprechende Vorhersagen des Sturms einsetzen.

Echte nomische Muster sind hingegen sogar stabil unter vielen Interventionen. So hilft es zwar nicht, die gelben Finger zu verhindern, um sein Lungenkrebsrisiko zu senken, aber es hilft natürlich schon, wenn wir das dazu führende Rauchverhalten ändern und am besten ganz mit dem Rauchen aufhören. Eine Intervention im Hinblick auf das Rauchverhalten lässt den Zusammenhang zwischen Rauchverhalten

und Lungenkrebs unverändert bestehen und lässt so entsprechende Vorhersagen zu, welche Auswirkungen derartige Interventionen mit sich bringen werden. Sie führen schließlich zu Veränderungen des Lungenkrebsrisikos. Entscheidend sind in all diesen Fällen die vorliegenden nomischen Muster bzw. das Vorliegen stabiler kausaler Mechanismen und nicht nur die auftretenden Regularitäten. Nur auf diese Weise ist unser Wissen für praktische Zwecke nutzbar.

Im dritten Fall haben wir es schließlich mit einer umgekehrten Kausalbeziehung zu tun. Die Grippe verursacht bestimmte Symptome, anhand derer wir sie wieder im Sinne des abduktiven Schließens erschließen dürfen. Auch hier können Interventionen an den Symptomen zu falschen induktiven Schlüssen führen. Korrektes induktives Schließen setzt also wiederum eine im Wesentlichen korrekte Kenntnis der kausalen nomischen Zusammenhänge voraus. Das zeigt sich schließlich ebenso in den Fällen, in denen wir nicht einmal mehr indirekte nomische Zusammenhänge vermuten. Wenn z.B. bisher immer ein Mensch mit dem 3. Buchstabe A im Vornamen den Raum betreten hat, sobald Hanna eine 3 gewürfelt hat, so werden wir nun keineswegs schließen, dass das auch beim nächsten Wurf von Hanna zu erwarten ist. Sobald F und G sich auf Mengen beziehen, die keine *natürlichen Arten* darstellen, erhalten wir normalerweise keine projizierbaren Aussagen mehr, die wir durch Beobachtung von Instanzen weiter begründen können. Das ist auch nicht zu erwarten, denn die Objekte weisen in solchen Fällen keine wesentlichen gemeinsamen Merkmale mehr auf, d.h., sie zeigen kein ähnliches kausales Verhalten, das wir für die Projektion in die Zukunft benötigen. Das gilt etwa für Prädikate wie »Gegenstände, die Franz gehören« oder auch Prädikate vom »grue«-Typ, denen wir in Goodmans Grue-Paradox begegnen werden.

Sogar in den Fällen, in denen zwei Prädikate für vermutlich natürliche Arten verknüpft werden, die zu *verschiedenen Bereichen* gehören, dürfen wir nicht sogleich auf Projizierbarkeit setzen. Manche Philosophen wie Davidson argumentieren etwa dafür, dass wir keine gesetzesartigen Zusammenhänge zwischen mentalen Eigenschaften und neuronalen Eigenschaften erwarten dürfen, selbst wenn es zumindest einfache generelle Zusammenhänge in jedem der beiden Bereiche gibt. Es müssen

also zusätzlich – wie in Kapitel 1 beschrieben – nomische Konditionale vorliegen, wenn wir Projizierbarkeit annehmen wollen.

Deutlicher wird das noch für einen Mix chemischer Konzepte und solcher aus dem menschlichen Bereich – etwa für Artefakte. Wenn bisher alle uns bekannten Löffel aus Metall waren, muss sich das keineswegs so fortsetzen. Wir dürfen für Löffel vielleicht eine gewisse Grundform oder bestimmte basale Funktionalitäten extrapolieren, aber viel mehr ist nicht zu erwarten. Insbesondere finden wir hier keine echten nomischen Muster für solche »gemischten« Aussagen. Nun möchte ich allerdings nicht gleich argumentieren, dass es sich bei Löffeln überhaupt nicht mehr um eine natürliche Art handelt und daher überhaupt keine Projizierbarkeit vorliegt. Die Übergänge scheinen mir eher graduell zu sein und das überträgt sich auch auf den Begriff des *nomischen Musters*. Doch was ist damit genau gemeint?

3.8 Nomische Muster

Das Konzept der nomischen Muster ist gedacht als eine Abschwächung bzw. als Ersatz für das Konzept von *Naturgesetzen*. Dabei sollen nomische Muster zum einen die philosophischen Aufgaben übernehmen, die wir Naturgesetzen zugedacht haben, und zugleich einige der schwierigsten Probleme lösen, die wir mit dem Begriff der Naturgesetze verbinden. An zahlreichen Stellen stützen sich Philosophen auf eine bestimmte Konzeption von Naturgesetzen und versuchen mit ihrer Hilfe philosophische Probleme zu klären. Dabei sind vor allem zwei Probleme für Naturgesetze aufgetreten, für die das Konzept der nomischen Muster eine Lösung anbieten soll. Das erste notorische Problem ist die Auszeichnung von Naturgesetzen und ihre *Abgrenzung* gegenüber anderen Aussagen insbesondere Allaussagen wie:

(1) Zu allen Zeiten fließt in allen Flüssen insgesamt mehr Wasser als Coca-Cola.

Diese Aussage ist sicher kein Naturgesetz, besitzt aber den logischen Charakter einer Allaussage ähnlich wie typisch naturwissenschaftliche Behauptungen wie z.B. das newtonsche Gravitationsgesetz.

Das zweite Problem betrifft die Anwendungen unseres Konzepts der Naturgesetze. Häufig wird als ein wesentliches Merkmal der Wissenschaften (etwa in der Abgrenzungsdebatte mit dem Kreationismus) genannt, dass die Wissenschaften Naturgesetze ermitteln und dass sie mit deren Hilfe bestimmte Phänomene erklären und vorhersagen. Doch in vielen Wissenschaften wie etwa den Sozialwissenschaften und der Biologie lassen sich nur schwer *strikte Gesetze* ausmachen. Vielmehr stoßen wir auf einfache Generalisierungen, die wir trotzdem zum Erklären einsetzen:

(2) Je mehr Wasser und Dünger eine bestimmte Pflanze erhält, umso stärker wächst sie. Oder:

(3) Je höher die Intelligenz einer Person ist, umso erfolgreicher ist sie im Beruf.

Derartige Allaussagen stellen (selbst wenn sie wahr sind) keine strikten Naturgesetze dar, denn von solchen Gesetzen erwarten wir, dass sie ausnahmslos gelten und nicht auf bestimmte Anwendungsbereiche eingeschränkt sind. Eine Hilfskonstruktion war immer zu sagen, dass es sich um *Ceteris-paribus-Gesetze* handelt. Doch es bleibt meist unklar, was ein CP-Gesetz CP(G) genau behauptet. CP(G) sollte jedenfalls mehr besagen als: G gilt, außer es handelt sich um einen Ausnahmefall. Hier droht die völlige Inhaltslosigkeit von CP(G) (vgl. Bartelborth 2007).

Kritiker haben zusätzlich eingewandt, dass CP-Gesetze überhaupt keine *Gesetze* mehr sind und wir durch die Namensgebung eigentlich eine Irreführung begehen. Insbesondere bleibt jedenfalls die Frage zu klären, was CP-Gesetze von akzidentellen Aussagen vom Typ (1) unterscheidet. Nomische Muster sollen hier Abhilfe schaffen und zugleich die Debatte vermeiden, ob es sich noch um echte Gesetze handelt, die von unterschiedlichen Autoren unterschiedlich beantwortet wird. Man gibt nun zu, dass es sich um etwas anderes handelt, wobei allerdings als Grenzfall weiterhin die Naturgesetze enthalten sind.

Nomische Muster sind Generalisierungen, die typischerweise kausale Abhängigkeiten beschreiben und dabei eine spezielle *Invarianz* bzw. *Stabilität* aufweisen. So behaupten wir in (2), wenn wir es als ein nomisches Muster betrachten, dass wir bestimmte gezielte Manipulationen

(oder Interventionen) an dem jeweiligen System durchführen können, unter denen die Behauptung (2) weiterhin gilt bzw. invariant ist. Den Begriff der Intervention finden wir genauer (und zumindest semiformal) erläutert in den Kausaldebatten (etwa bei Woodward 2003), werden ihn aber sogleich intuitiv erklären. Um deutlicher zu machen, welche Invarianzen hier entscheidend sind, werden die Abhängigkeiten häufig durch eine einfache Gleichung für ein System s dargestellt, wie etwa in:

(2*) Pflanzenwachstum(s) = a × Wassermenge(s) + b × Dünger(s) + c

Hierbei sind a, b und c einfach reelle Zahlen und es wird ein linearer Zusammenhang von Wassermenge und Dünger zum Pflanzenwachstum angenommen. Entscheidend ist nun, dass ein derartiger funktionaler Zusammenhang zumindest für einige Interventionen (in einem gewissen Wertebereich) stabil bleibt, damit (2*) ein nomisches Muster beschreibt. Es muss zumindest einen nichttrivialen Bereich für die Wassermenge und einen für die Düngermenge geben, in dem wir durch gezielte Veränderung nur dieses einen Parameters die entsprechenden Veränderungen im Pflanzenwachstum herbeiführen können, die von (2*) vorhergesagt werden. Also muss es zum Beispiel Situationen geben, in denen die gezielte Vergrößerung der zugeführten Wassermenge bei einem Konstanthalten aller anderen Einflussgrößen zu einem vermehrten Pflanzenwachstum führen würde. Das entspricht auch den typischen Fragestellungen in *idealen Experimenten*: Was würde passieren, wenn wir ganz bestimmte Größen gezielt (als Intervention) verändern würden? Auf einige dieser *kontrafaktischen Fragen* muss ein nomisches Muster eine Antwort geben. Je mehr dieser Fragen es korrekt beantwortet, umso *stärker* ist das Muster. Diese Stabilität gilt außerdem in einem bestimmten Bereich von Randbedingungen, d.h. die Muster zeigen diese Invarianz in einem bestimmten Bereich, den wir für die Bestimmung der Stärke eines Musters mit berücksichtigen müssen (vgl. Kap. 4.3.4).

Was hier verlangt wird, kann durch einen Vergleich mit einem nichtnomischen Muster verdeutlicht werden. Nehmen wir einmal an, es gilt tatsächlich der folgende Zusammenhang:

(4) Je gelber die Finger einer Person sind, umso höher ist ihr Lungenkrebsrisiko.

Dieser Zusammenhang besteht, weil er durch das Rauchen vermittelt wird, das zum einen die Finger gelb einfärbt und zum anderen unser Krebsrisiko erhöht, und könnte sogar zu einer entsprechenden Gleichung führen, die für einen gewissen Bereich Gültigkeit besitzt. Doch weder der einfach beschriebene Zusammenhang noch die entsprechende Gleichung wären stabil unter den entsprechenden Interventionen. Würden wir die Gelbfärbung der Finger vermindern, indem wir sie säubern, würde daraus keine Änderung des Lungenkrebsrisikos resultieren, jedenfalls dann nicht, wenn wir alle anderen Parameter (insbesondere das Rauchverhalten der Person) konstant halten. Das ließe sich sogar in einem kontrollierten Experiment überprüfen. Mit der vermutlich wahren Generalisierung (4) erhalten wir also kein nomisches Muster. Sie ist daher nicht für Erklärungen und auch nur bedingt für Vorhersagen geeignet. Sie erlaubt bestimmte Vorhersagen, solange wir nicht in das System gezielt eingreifen, aber gerade das interessiert uns in vielen Fällen. Wir möchten wissen, mit welchen Eingriffen wir bestimmte Veränderungen eines Systems herbeiführen können. Das beantworten nomische Muster, aber nicht bloße Korrelationen wie die in (4) genannte. Mit nomischen Mustern können wir daher auch *erklären*, warum ein bestimmtes Pflanzenwachstum auftritt, während die gelben Finger keineswegs das Lungenkrebsrisiko erklären.

Nomische Muster stellen damit die *Beschreibungen* bestimmter grundlegender Dispositionen in unserer Welt dar, oder wir können mit »nomische Muster« auch diese Zusammenhänge selbst meinen, also in unserem Fall die zugrundeliegenden stabilen Dispositionen. Ähnlich mehrdeutige Verwendungsweisen finden sich natürlich ebenso im Fall des Gesetzesbegriffs und der Kontext sollte jeweils deutlich machen, welche Lesart intendiert ist.

Insbesondere physikalisch sehr grundlegende Dispositionen erzeugen relativ stabile kausale Zusammenhänge mit den geforderten kontrafaktischen Abhängigkeiten. Wenn etwas *zerbrechlich* ist, dürfen wir normalerweise folgern, dass es zerbrechen würde, wenn wir es unter geeigneten Randbedingungen fallen lassen würden. Das sind die kontrafaktischen Zusammenhänge, die für die Invarianzbedingung erforderlich sind. Die spezielle Disposition der Zerbrechlichkeit kann aber natürlich auf grundlegendere Muster zurückgeführt werden. Überhaupt zeigt sich

hier (wie schon in den Überlegungen zur Invarianz), dass es sich beim Konzept der nomischen Muster um einen graduellen Begriff handelt. Es gibt grundlegendere Muster mit einem größeren Invarianzbereich und solche mit einem kleineren. Damit variiert zugleich ihre Erklärungskraft. Außerdem gibt es relativ strikte Dispositionen, bei denen sich eine Manifestation bei Vorliegen der Auslösebedingung und gewissen Randbedingungen praktisch immer einstellt und daneben probabilistische Dispositionen (manchmal Propensitäten genannt), die sich nur in einer Erhöhung von bestimmten Wahrscheinlichkeiten manifestieren.

Statt des Satzes (4) hätten wir also den Satz (5) »Je mehr eine Person raucht, umso höher ist ihr Lungenkrebsrisiko«, zu Erklärungszwecken wählen sollen, der die geforderte Invarianz aufweist, denn nach allem, was wir darüber wissen, verändert eine Veränderung unseres Rauchverhaltens tatsächlich unser Lungenkrebsrisiko in der entsprechenden Richtung. In (5) beschreiben wir eine probabilistische Disposition von Menschen, auf die Zufuhr von Zigarettenrauch mit einem Lungenkrebs zu reagieren. Doch auch hier kann die zukünftige Forschung vielleicht präzisere Muster liefern, die die Manifestationsbedingungen dieser Disposition genauer beschreiben. Das wären stärkere Muster mit einer höheren Erklärungsleistung (vgl. dazu Kap. 4.3.4).

Diese unterschiedlich guten Erklärungen kennen wir aus der Praxis unseres Erklärens im Alltag und in der Wissenschaft. Das wird durch eine graduelle Konzeption von nomischen Mustern recht gut nachgezeichnet. Der klassische Ansatz kennt dagegen nur die strikte Abgrenzung von Gesetzen und Nicht-Gesetzen, und damit war es oft nur ein kleiner Schritt, einer Disziplin wie der Biologie oder den Sozialwissenschaften ganz abzusprechen, dass sie echte Gesetze formulieren können. Damit verloren sie zugleich jegliche Erklärungskraft. Doch das scheint uns in vielen Fällen weit übertrieben zu sein und die tatsächliche Erklärungskraft dieser Disziplinen zu unterschätzen. Selbst wenn diese Erklärungen häufig nicht die gleiche Gestalt und Stärke wie physikalische Erklärungen aufweisen, erkennen wir doch zumindest an, dass auch hier kausale Zusammenhänge aufgedeckt werden, die im Einzelfall durchaus einige Eigenschaften von bestimmten Objekten oder Systemen erklären können. Man denke z.B. an die Evolutionstheorie oder einfache ökonomische Zusammenhänge wie die von Angebot und Nachfrage,

die sicher eine gewisse Erklärungskraft besitzen. Mit dem Konzept der nomischen Muster kann man diese Übergänge angemessen beschreiben und kann darüber hinaus zeigen, welche Rolle solche gesetzesähnlichen Zusammenhänge in Erklärungen spielen.

Über die Frage, was die Stärke der Muster ausmacht und damit ihre Erklärungskraft bestimmt, gibt es allerdings unterschiedliche Auffassungen. Woodward (2003) und andere beziehen sich explizit nur auf die *funktionale Invarianz* der Muster für ein konkretes einzelnes System. Damit ist gemeint, dass für ein einzelnes Objekt möglichst große Invarianzbereiche für den beschriebenen Zusammenhang gelten sollten. Bartelborth (2007, 2008) hält dagegen ebenso die *Bereichsinvarianz* für ein erklärungsrelevantes Merkmal der Muster. Dabei geht es (im Sinne der klassischen Vereinheitlichungskonzeption von Erklärung) darum, dass für ein solches Muster ebenfalls zählt, auf welche und vor allem wie viele Typen von Objekten sie zutreffen. Je weiter verbreitet ein Muster in unserer Welt ist, umso basaler ist es und umso besser ist es geeignet, bestimmte Phänomene zu vereinheitlichen und so zu erklären. Das ist gerade die Art von Invarianz, die von den Vertretern einer Vereinheitlichungskonzeption von Erklärung in den Vordergrund gestellt wird. Die beiden Formen der Invarianz stehen typischerweise in einem Spannungsverhältnis zueinander (vgl. Bartelborth 2008). Für die *Erklärungskraft* der Muster sind sogar noch weitere Eigenschaften zu berücksichtigen, die aber auch damit zusammenhängen, wie strikt die jeweiligen nomischen Muster sind (vgl. auch Kap. 4.3.4).

Die Idee, sich genauer mit diesen Formen von Invarianz zu beschäftigen, stammt ursprünglich aus der Kausalitätsdebatte. Startpunkt waren die sogenannten *Manipulationstheorien der Kausalität*, nach denen für uns ein kausaler Zusammenhang dort am deutlichsten wird, wo wir aktiv etwas verändern können. Kausalität sollte natürlich nicht nur auf menschliches Manipulieren unserer Umwelt beschränkt bleiben und deshalb entwickelte sich der sogenannte *interventionalistische Ansatz*, nach dem A dann als Ursache von B gelten darf, wenn durch eine gezielte Intervention an A eine Veränderung an B herbeigeführt wird. Diese Grundidee bedurfte zunächst einer Explikation, was mit einer gezielten Intervention gemeint ist.

Das Konzept scheint uns zwar intuitiv verständlich zu sein, seine Präzisierung führt aber doch zu etlichen Schwierigkeiten (vgl. Woodward 2003). Aufdecken können wir derartige Kausalbeziehungen jedenfalls nur, wenn diese nicht nur im Einzelfall bestehen, sondern ein allgemeineres Muster darstellen, das sich in kontrollierten Experimenten überprüfen lässt. Für Erklärungen sind normalerweise singuläre Kausalbeziehungen unzureichend (wenn es sie den gäbe und wir sie ermitteln könnten). Erst der Einblick in allgemeinere Zusammenhänge – nämlich die nomischen Muster – kann uns verständlich machen, warum bestimmte Ereignisse aufgetreten sind. Doch diese allgemeinen Zusammenhänge müssen nicht gleich den Charakter von strikten Naturgesetzen haben, sondern oft finden wir nur *gesetzesähnliche* Muster, die durchaus Erklärungskraft besitzen.

3.9 Die Rabenparadoxie und das »grue«-Paradox

Bevor wir uns dem Schluss auf die beste Erklärung zuwenden, möchte ich noch eine weitere Problematik induktiver Schlüsse ansprechen, die schon im ersten Kapitel ein Thema war. Wir haben etwa *Nicods Kriterium* (NR) kennengelernt, das eine Grundform induktiven Schließens zu sein scheint. Einzelne positive Instanzen bestätigen demnach eine Aussage über eine größere Gruppe von Objekten. Allerdings haben wir inzwischen erfahren, dass Rechtfertigung genau genommen dreistellig ist und daher (NR) eigentlich entsprechend zu ergänzen wäre. Dazu haben sich verschiedene Autoren (wie z.B. Good 1983) Beispiele für ein bestimmtes Hintergrundwissen ausgedacht, so dass das einfache Nicod Kriterium nicht mehr plausibel erscheint.

Es sei wiederum unsere Hypothese R: *Alle Raben sind schwarz*. Man brachte gegen (NR) nun folgende Situation ins Spiel: Nehmen wir an, wir wüssten schon, dass es entweder (i) nur insgesamt 100 schwarze Raben gibt oder (ii) eine Million schwarze Raben und 10 weiße. Nur im ersten Fall ist unsere Hypothese wahr. Aber wenn ich überhaupt einen oder mehrere schwarzen Raben finde, dann spricht das eher dafür, dass der zweite Fall vorliegt, da wir in dem Fall viel häufiger auf schwarze

Raben stoßen würden. Also gilt (NR) nicht für jedes Hintergrundwissen, sondern nur für bestimmte Fälle von »normalem« Hintergrundwissen.

Wir sehen diesen Zusammenhang zu weiterem Hintergrundwissen in ähnlicher Weise in anderen Fällen. Ist unsere Hypothese, dass zwei Philosophieprofessoren im Raum sind, und wir erfahren als Datum, dass Professor Wittgenstein im Raum ist, so muss das nicht unbedingt unsere Hypothese stützen. Wissen wir z.B., dass Professor Wittgenstein keine anderen Philosophieprofessoren mag und sofort den Raum verlässt, sollte einer von ihnen dazukommen, so spricht unser Datum gegen unsere Hypothese.

Viele weitere Beispiele belegen, dass (NR) nicht ganz so harmlos ist, wie es auf den ersten Blick scheint. Trotzdem wird es sicher häufig sinnvoll anzuwenden sein und ist eine Grundidee hinter unseren anderen Verfahren. Besonders deutlich wird das im Falle der hempelschen Instanzenbestätigung, aber eigentlich auch im Falle des hypothetisch-deduktiven Schließens. Allerdings führt uns (NR) manchmal sogar in Paradoxien. Da ist vor allem die sogenannte *Rabenparadoxie* zu nennen, die jedoch ebenso die meisten anderen induktiven Schlussformen plagt.

Das Rabenparadox. Das folgende Beispiel haben wir schon recht kurz in Kapitel 1.6.3 behandelt: Es gehe wieder um unsere Rabentheorie R in ihrer einfachen Interpretation durch ein materiales Konditional. Außerdem betrachten wir noch die dazu logisch äquivalente Theorie K: »*Alle nicht-schwarzen Dinge sind Nicht-Raben.*« Das ist logisch gesehen die Kontraposition der Rabentheorie und daher logisch mit dieser äquivalent. Es ist nun eine naheliegende und kaum bestrittene Forderung, dass logisch äquivalente Aussagen durch dieselben Daten bestätigt werden. Das war Hempels Äquivalenzforderung. Nehmen wir jedoch (NR) hinzu, geraten wir sogleich in Schwierigkeiten. Sehe ich z.B. eine weiße Küchenmaschine (E), so handelt es sich dabei um einen nicht-schwarzen Nicht-Raben. Also wird nach (NR) unsere kontraponierte Rabentheorie K durch E bestätigt. Die Theorie K ist logisch äquivalent zur Rabentheorie R, also sollte R ebenfalls durch E bestätigt werden. Danach bestätigt die weiße Küchenmaschine unsere Vermutung, dass alle Raben schwarz sind. Doch dieses Ergebnis (Quine spricht hier von *indoor ornithology*) ist kaum plausibel.

Carl Gustav Hempel, der das Paradox entwickelte und dessen Instanzenbestätigung davon ebenfalls betroffen war, war schließlich bereit, dieses unangenehme Resultat zu schlucken. Auch die modernen bayesianischen Ansätze (s. Fitelson & Hawthorne 2010), die ebenso davon betroffen sind, haben sich letztlich dazu bekannt (vgl. Kap. 5.6.4). Sie beweisen mit Hilfe einer Reihe von empirischen Zusatzannahmen wie der, dass es viel mehr nicht-schwarze Dinge als Raben gibt, dass zumindest der Grad der Bestätigung durch einen schwarzen Raben für R deutlich höher ist, als der durch die weiße Küchenmaschine. Das ist ihre »Lösung« des paradoxen Resultats.

Das kommt mir nicht wirklich akzeptabel vor. Man stelle sich vor, ein Biologe möchte unsere Rabentheorie begründen und sagt dazu, dass er nun besonders viele nicht-schwarze Dinge in irgendwelchen Küchen daraufhin untersuchen möchte, ob sie zugleich Nicht-Raben sind. Da das jeweils nur eine schwache Bestätigung der Rabentheorie bietet, möchte er das durch größere Anzahlen wettmachen. Untersucht er nur genügend nicht-schwarze Dinge, käme er so zu einer guten Bestätigung der Rabentheorie, wenn Hempel und die Bayesianer Recht hätten. Wir können über diesen Ansatz wohl nur lachen, denn er untersucht vermutlich niemals Raben und daher erfährt er auch nichts über deren Farbe, unabhängig davon, ob es viele Raben gibt oder nicht. Wir sollten also darauf beharren, dass die weißen Küchenmaschinen uns keine Informationen über Raben liefern, gleichgültig wie viele wir davon untersuchen. Der Forschungsantrag des Biologen gehört abgeschmettert (vgl. a. Siebel 2004 und Kap. 5.6.4).

Wie können wir aber dann dem Rabenparadox entkommen? Wir müssen auf eine Lösung zurückgreifen, die schon im ersten Kapitel kurz genannt wurde und bereits in Vorformen bei Quine (1969) zu finden ist. Wir müssen (NR) auf die Fälle beschränken, in denen unsere fragliche Hypothese *gesetzesartigen Charakter* hat. Die Prädikate unserer Theorie müssen sich auf *natürliche Arten* beziehen und die Hypothese muss eine (kausale) gesetzesartige Behauptung (ein nomisches Muster oder wenigstens ein nomisches Konditional mit Erklärungswert) darstellen. Die sind nicht leicht zu erkennen und man benötigt weiteres Hintergrundwissen dazu (vgl. Bartelborth 2007). Das wird jedenfalls der Lösungsvorschlag des *Schlusses auf die beste Erklärung* sein. Demnach gilt (NR) nur in den

Fällen, in denen die Theorie die Instanzen auch *erklärt*. Dazu benötigen wir mehr, als dass das Datum eine bloße Instanz unserer Hypothese ist. Insbesondere kommt es auf unser weiteres Hintergrundwissen über die Art der vorliegenden Zusammenhänge an. Die Beispiele des letzten Abschnitts legen diese neue Art von Instanzenbestätigung nahe (vgl. a. Kap. 5.5.23).

Das wirft allerdings die Frage auf, ob unsere ursprüngliche Rabentheorie überhaupt auf diese einfache Weise bestätigt werden kann. Das hängt davon ab, wie die Theorie genau zu verstehen ist. Ich habe sie hier immer schon so verstanden, dass ein bestimmter gesetzesartiger Zusammenhang behauptet wird. Es gehört demnach zur genetischen Ausstattung von Raben dazu, dass sie schwarz sind. Wird das durch unsere biologischen Erkenntnisse gestützt, können schwarze Raben unsere Rabenhypothese stützen.

Ein weiteres Problem für die Regel (NR) ist daneben die Frage, was eigentlich eine *Instanz* einer Generalisierung wie H ≡ ∀x(Rx→Sx) (Alle Raben sind schwarz) ist. Da kommen neben A ≡ Ra&Sa auch andere Aussagen wie z.B. Sa in Betracht, weil daraus bereits Ra → Sa ableitbar ist. Jedenfalls bestätigt A die Hypothese H nicht im Sinne des hypothetisch-deduktiven Bestätigungskonzeptes und darin wird schon eine bestimmte Asymmetrie erkennbar. So gilt einerseits H&Ra ⇒ Sa, d.h., es gilt B(H:Sa;Ra), also wird H durch Sa in einer bestimmten Situation (mit Hintergrundwissen Ra) bestätigt, in der schon bekannt ist, dass es sich bei a um einen Raben handelt. Das wird manchmal auch so dargestellt, dass hier ein zweistufiges Verfahren vorliegt: In einem ersten Schritt wird ein Rabe (zufällig) ausgewählt und erst im zweiten Schritt daraufhin überprüft, ob er schwarz ist. Einige Wissenschaftstheoretiker suchen nach der Lösung der Rabenparadoxie in dieser Richtung. Da die zeitliche Reihenfolge aber letztlich in unseren bisherigen Ansätzen nicht zum Tragen kommt, scheinen diese Ansätze nicht wirklich erfolgversprechend zu sein. Es bleibt aber zumindest die oben erwähnte Asymmetrie, denn andererseits gilt:

H & ¬Sa ⇒ ¬Ra, also B(H:¬Ra;¬Sa).

Also haben wir in diesem Fall ebenfalls ein kontraintuitives Ergebnis. Wenn wir bereits wissen, dass etwas nicht-schwarz ist, so könnte

allerdings die Entdeckung, dass es sich nicht um einen nicht-Raben handelt, intuitiv schon eher dazu beitragen, die Hypothese H zu stützen, denn es wird ein möglicher Falsifikator von H ausgeschlossen. Wenn es uns gelingt, alle Falsifikatoren auszuschließen, wird dadurch schließlich ebenfalls die Hypothese H bewiesen. Allerdings könnte sie dann noch auf triviale Weise wahr sein, indem es etwa überhaupt keine Raben gibt. Doch wenn wir von diesem Spezialfall einmal absehen, wirkt das Verfahren nicht mehr ganz so paradox, sondern nur noch sehr ineffektiv. Es wären extrem viele nicht-schwarze Gegenstände zu untersuchen, ob sie Nicht-Raben sind, bevor wir auf diesem Wege einen beachtlichen Teil der potentiellen Falsifikatoren von H ausgeschlossen hätten. Dazu kommt, dass die Rabenhypothese implizit auch noch über unendlich viele Zeitpunkte spricht. Sie besagt nicht nur, dass alle Raben schwarz sind, sondern ebenfalls, dass sie das zu allen Zeitpunkten dauerhaft sind. Wir müssten also alle Objekte immer wieder testen, ob die Hypothese noch gilt, bzw. die nicht-schwarzen Gegenstände immer wieder testen, ob sie nicht doch inzwischen Raben geworden sind, wenn wir alle Gegeninstanzen auf diesem Weg ausschließen wollen.

Das scheint ähnlich wie die eliminative Induktion abzulaufen. Aber der entscheidende Schritt fehlt hier leider. Für die eliminative Induktion müssen wir zunächst eine *endliche* Liste von substantiellen Konkurrenzhypothesen entwerfen, die hier nicht in Sicht ist. Die Anzahl der möglichen Falsifikatoren ist potentiell unendlich, wenn wir noch bedenken, dass auch zukünftige nicht-schwarze Gegenstände und alle möglichen Zeitpunkte zählen. Daher ist nicht erkennbar, wie es uns auf diesem Wege gelingen sollte, tatsächliche Fortschritte in der Bestätigung der Hypothese H zu erzielen. Der Anteil der nicht-schwarzen Gegenstände und der jeweiligen Untersuchungszeitpunkte, die wir tatsächlich untersuchen können, dürfte immer recht klein bleiben. Der Forschungsantrag des Biologen scheint somit doch nicht zu retten zu sein. Das bedeutet, dass das Rabenparadox vermutlich weiterhin ein gravierendes Problem der Regel (NR) und ähnlicher Regeln darstellt und wir zu seiner Auflösung auf stärkere inhaltliche Anforderungen wie natürliche Arten, nomische Konditionale und die Erklärungsstärke von H zurückgreifen müssen (vgl. Kap. 1.6.3).

Zur Erinnerung: Der Lösungsvorschlag war, dass wir mit der Regel (NR) nur nomische Konditionale H bestätigen können, und wir auch nur nach solchen Konditionalen suchen. Die sind aber weder logisch äquivalent zu ihrer Kontraposition, noch ist die Kontraposition K selbst wieder ein nomisches Konditional. Dann ist K auch nicht mehr durch seine Instanzen zu bestätigen und die Rabenparadoxie ist nicht mehr reproduzierbar. Das ist ganz im Sinne des Schlusses auf die beste Erklärung, den wir in Kapitel 4 kennenlernen werden.

Das grue-Paradox. Auf ähnliche Probleme stoßen wir auch im sogenannten »grue«-Paradox. Dabei ist »grue« ein Kunstwort, das aus den Ausdrücken »green« und »blue« gebildet wird. Im Deutschen können wir entsprechend von dem »graun«-Paradox sprechen. Das Paradox stammt ursprünglich von Nelson Goodman und wir können es in etwas abgewandelter Form wie folgt rekonstruieren: Zwei Juweliere streiten sich über die Farbe von Smaragden. Juwelier X spricht deutsch und behauptet:

(1) Alle Smaragde sind grün.
(2) Das wird durch alle bisher untersuchten Smaragde bestätigt.

Juwelier Y spricht dagegen die Sprache »graunblün«. Die ist eng mit dem Deutschen verwandt und kennt allerdings nur statt der Farbprädikate »grün« und »blau« die Farbprädikate »graun« und »blün«. Die lassen sich wie folgt ins Deutsche übersetzen:

x ist graun bedeutet: x ist grün bis zum Jahr 2020 und blau danach.
x ist blün bedeutet: x ist blau bis zum Jahr 2020 und grün danach.

Juwelier Y vertritt nun die beiden Behauptungen:

(1*) Alle Smaragde sind graun.
(2*) Das wird durch alle bisher untersuchten Smaragde bestätigt.

Nehmen wir an, beide Juweliere haben dieselben Erfahrungen E mit Smaragden im Laufe ihrer Tätigkeit erworben. Außerdem stützen sich beide auf eine einfache Instanzenkonzeption von Rechtfertigung. Danach begründen die bisher beobachteten Smaragde, die alle nach

unserer Ansicht grün waren, die Behauptung (1). Doch Juwelier Y beschreibt das etwas anders. Für ihn waren die Smaragde graun und sie begründen daher seiner Meinung nach seine Behauptung (1*). Zumindest für den Zeitraum nach 2020 stellen beide Juweliere jedoch unterschiedliche Behauptungen auf. Stützt die Instanzenbegründung nun beide Behauptungen in gleicher Weise oder womöglich doch nur die erste von beiden? Wir glauben natürlich das Letztere, aber können wir das auch so begründen, dass es einen gewissen Grund sogar für Y bieten könnte, an seiner Auffassung zu zweifeln? Dafür ist leider nicht leicht zu argumentieren.

Juwelier Y könnte etwa behaupten, die Prädikate graun und blün seien deshalb so seltsam, weil sie aus blau und grün und einem zusätzlichen Zeitpunkt so künstlich konstruiert wurden. Damit wird ein Farbwandel im Jahr 2020 angenommen, für den es keine empirischen Anhaltspunkte gibt. Das sieht Juwelier Y aber ganz anders. Für ihn sind die Prädikate blau und grün seltsam konstruiert:

x ist grün bedeutet: x ist graun bis zum Jahr 2020 und blün danach.
x ist blau bedeutet: x ist blün bis zum Jahr 2020 und graun danach.

Für ihn ist es Juwelier X, der einen Farbwandel (von graun zu blün) im Jahr 2020 prognostiziert. Je nachdem, was wir also als Grundprädikate betrachten, werden wir andere induktive Schlüsse für natürlich halten. Welche Prädikate tatsächlich projizierbar sind und mit welchen wir tatsächlich nomische Muster beschreiben, die wir dann für Erklärungen und Prognosen heranziehen dürfen, lässt sich leider nicht durch ein einfaches Kriterium beantworten (vgl. Kap. 1.6.3).

Es gibt nur im Rahmen unserer Theorienbildung Hinweise darauf, welche Prädikate wir wählen sollten. So gibt es systematische Zusammenhänge zwischen unserer Farbwahrnehmung und den Wellenlängen des reflektiertes Lichts. Für die Entstehung dieser Reflektionen haben wir Erklärungen, die sie auf die Stoffeigenschaften der reflektierenden Gegenstände zurückführen etc. In diesem größeren Rahmen wird sich die »graunblün«-Auffassung erst noch zu bewähren haben. Wird etwa die Wellenlänge des reflektierten Lichts zu 2020 sich ändern? Oder was passiert dann aus Sicht von Juwelier Y? Solange der »graunblün«-Ansatz keine umfassende Konzeption mit Antworten auf solche Fragen

anzubieten hat, haben wir gute Gründe, uns Juwelier X anzuschließen. Erst der größere Rahmen und die darin angesiedelten Erklärungszusammenhänge entscheiden somit darüber, welche Prädikate wir nutzen sollten, um zu guten Vorhersagen zu kommen. Letztlich müssen wir wieder zwischen nomischen Konditionalen (wie den von Juwelier X) und den nicht-nomischen Generalisierungen (wie denen von Juwelier Y) unterscheiden und dafür gibt es leider kein einfaches Patentrezept.

3.10 Fazit

Die qualitativen induktiven Schluss- bzw. Begründungsverfahren stellen schon die Grundideen für die Bestätigung von wissenschaftlichen Theorien durch bestimmte Daten bereit und zeigen uns einige der grundlegenden Probleme induktiven Schließens auf. Eine Theorie T wird typischerweise durch ihre Instanzen bestätigt oder zumindest indirekt dadurch bestätigt, dass sich ihre Vorhersagen E als wahr herausstellen. Stellen wir dagegen fest, dass dann non-E auftritt, betrachten wir die Theorie als geschwächt oder sogar falsifiziert. Genauso testen wir unsere empirischen Theorien im Normalfall. Leider ist es nicht so ganz einfach, diese intuitiven Ideen in einem präziseren Verfahren zu erfassen. Die logische Ableitbarkeit von E aus T genügt für eine induktive Bestätigung von T durch E leider noch nicht. Wir benötigen einen stärkeren *inhaltlichen Zusammenhang* zwischen T und E, und es nicht klar, ob wir den mit überwiegend logischen Hilfsmitteln allein beschreiben können, oder ob wir (wie Achinstein es annimmt: vgl. Kap. 5.8.8) auf Erklärungsbeziehungen und damit letztlich auf kausale Zusammenhänge angewiesen sind. Außerdem stoßen wir auf das Problem der *Hilfsannahmen*, die meist erforderlich sind, um aus einer Theorie beobachtbare Vorhersagen ableiten zu können. Damit unser Verfahren nicht willkürlich wird, müssen wir Ad-hoc-Annahmen ausschließen und nur *plausible Annahmen* zulassen. Die müssen dann aber schon anderweitig induktiv begründet sein. Doch damit verschieben wir unser Problem nur. Ganz spezielle Hilfsannahmen im Sinne von Brückenprinzipien benötigen wir schließlich, wenn wir Theorien mit theoretischen Termen empirisch testen wollen. Deren Status ist erstens

nicht ganz klar (sind sie etwa rein analytischer Art, wie das Carnap früher vermutete, oder stellen sie doch gewisse empirische Behauptungen dar?), und zweitens ist keineswegs sichergestellt, dass sie immer das leisten, was wir uns von ihnen erhoffen, nämlich eine Messbarkeit der theoretischen Terme in bestimmten Situationen. Im Falle der Propensitäten treten z.B. ganz besondere Probleme auf (vgl. Bartelborth 2011 und Kapitel 5.4).

Überhaupt stellen probabilistisch formulierte Theorien eine besondere Herausforderung für das qualitative Induktionsverfahren dar, aber auch für die Verfahren, die selbst mit Wahrscheinlichkeiten arbeiten. Das erste Problem ist schon darin zu sehen, was diese Theorien genau behaupten. Schließlich müssen wir in unseren empirischen Tests der Theorie diese Behauptungen auf die Probe stellen. Doch dazu später mehr. Außerdem bleibt in unseren bisherigen Verfahren immer die Frage offen, welche Rolle *komparativen Überlegungen* zukommt. Zu unserer Theorie T existieren vielleicht Konkurrenztheorien T*, die E ebenso abzuleiten gestatten. Schwächt das in jedem Fall schon unsere Bestätigung von T durch E? Das werden wir weiter im Blick behalten müssen.

Insbesondere scheinen die Verfahren nur dann die erwünschte Induktionseigenschaft aufzuweisen, wenn wir uns in ihrer Anwendung auf nomische Muster beschränken, die wiederum eng mit bestimmten Kausalkonzeptionen und Erklärungstheorien zusammenhängen. Anderenfalls lassen sich auch Paradoxien wie die der Raben nicht mehr auf intuitive Weise lösen. Auch der inhaltliche Aspekt, dass wir nach möglichst erklärungsstarken Theorien suchen, wurde bisher in den Induktionsverfahren noch nicht berücksichtigt. Damit sind schon die Grundprobleme benannt, mit denen sich jede Konzeption induktiven Schließens beschäftigen muss. Der Schluss auf die beste Erklärung soll unser weiterer Wegweiser sein, um die genannten Probleme anzugehen.

4 Der Schluss auf die beste Erklärung

4.1 Die Ursachen der Cholera

Beginnen möchte ich mein Plädoyer für den Schluss auf die beste Erklärung bzw. das abduktive Schließen mit einem historischen Beispiel. Dabei handelt es sich vermutlich um eine der wichtigsten Entdeckungen der Menschheit und den Beginn der modernen epidemiologischen Forschung. Die wird sicher gern mit der modernen Statistik assoziiert, aber das Beispiel zeigt schon, dass es sich vor allem um eine Form abduktiven Schließens handelt, und das war auch das Resümee des Statistikers David Freedman in seinem spannenden Artikel (1999) zu diesem Thema. Cholera war eine der schlimmsten Plagen der Menschheit, die vermutlich aus Indien nach Europa gelangte und vor allem im 19. Jahrhundert gerade in den großen Städten laufend zu neuen Epidemien mit Tausenden von Toten führte. Es war vor allem dem Einsatz des Londoner Arztes John Snow zu verdanken, dass die Ursachen bzw. die Überträger der Cholera identifiziert werden konnten, was – allerdings erst einige Zeit nach dem Ableben Snows – schließlich zu einer Eindämmung dieser Epidemien führte. Snow veröffentlichte seine Theorie und seine Daten dazu in seinem Buch »On the Mode of Communication of Cholera«, das er zunächst 1849 veröffentlichte, aber mit deutlich mehr Belegen in einer zweiten Auflage 1855 herausgab.

Die allgemein akzeptierte Theorie über Cholera war zu dem Zeitpunkt die sogenannte *Miasma-Theorie*, nach der spezielle giftige Ausdünstungen (also vor allem die schlechte Luft in den Städten) zu den Cholera-Epidemien führten. Deshalb wurden während der Epidemien manchmal bestimmte Fabriken vorübergehend geschlossen, um die Luft wieder zu verbessern. Die Theorie der Übertragung durch direkten Kontakt war demgegenüber zurückgetreten. Zu der Miasma-Theorie entwickelte Snow eine Konkurrenztheorie, die seiner Meinung nach die Fakten im

Zusammenhang mit der Cholera viel besser erklären konnte, die er z.T. selbst in mühevoller Kleinarbeit recherchieren musste.

Unterstützt wurde er bei seiner Datensammlung aber auch von dem Leiter einer Untersuchungskommission zum Cholera-Ausbruch 1854 in London von Dr. William Farr, der allerdings selbst ein überzeugter Vertreter der Miasma-Theorie war. *Snows Infektionstheorie* (wie ich sie jetzt modern nennen möchte), behauptete, dass es sehr kleine und daher nicht sichtbare, lebende Erreger der Krankheit gäbe, die mit dem Wasser oder der Nahrung in den Körper gelangen, sich dort vermehren, was die unangenehmen Symptome (besonders Durchfall) hervorruft, und den Körper dann mit den Ausscheidungen wieder verlassen. Schließlich gelangen sie vor allem über das Wasser wieder in andere Menschen, was zu den epidemischen Auswüchsen der Cholera führt. Diese für die damalige Zeit doch recht phantasievolle Geschichte konnte sich zu Lebzeiten Snows leider nicht durchsetzen, aber Dr. Farr war bereits 1866 von Snows Theorie überzeugt, nachdem er dessen Werk von 1855 studiert hatte. Das stellte einen schönen Fall abduktiver Rechtfertigung dar.

Snow hatte dazu die folgenden Fakten vor allem vom Beginn der Epidemie 1854 in London zusammengetragen (vgl. Freedman 1999):

(1) Es gab ein *Zeitintervall* zwischen der Infektion (dem Kontakt mit entsprechend verunreinigtem Wasser) und dem Ausbruch der Krankheitssymptome von etwa 2–3 Tagen. Das war nach Snow die Zeit, die die Erreger benötigten, um sich im Körper zu vermehren. Das könnte die Miasma-Theorie natürlich auch noch dadurch erklären, dass sie den Zeitpunkt des ersten Kontakts anders bestimmt. Wenn man auf diesen Zusammenhang allerdings erst einmal aufmerksam wird, bietet Snows Theorie auf jeden Fall die gehaltvollere Erklärung.

(2) Die Cholera breitet sich entlang der *menschlichen Handelswege* aus. Das passte gut zur Infektionstheorie, stellte hingegen eine klare *Erklärungsanomalie* für die Miasma-Theorie dar. Es war kaum anzunehmen, dass der Wind die schlechte Luft genau entlang der Handelswege wehte. Außerdem hätte sich die Wirkung bald verflüchtigen müssen.

(3) Die Seeleute auf Schiffen im Hafen erkrankten nur an der Krankheit, wenn sie mit den Menschen der Hafenstädte in *Kontakt* kamen. Die Luft übertrug sich aber natürlich in jedem Fall auch auf die Schiffe.

(4) Für den Ausbruch der Cholera in London 1848 konnte Snow den Seemann (John Harnold) identifizieren, von dem die Krankheit ihren *Ursprung* in London nahm. Die nächsten Opfer hatten nach Harnold in dessen Zimmer gewohnt.
(5) Während der Startphase der Epidemie 1854 zeichnete Snow auf einer Karte die Wohnorte aller Opfer als detaillierte Strichlisten auf der Karte ein und konnte eine eindeutige *Häufung der Opfer* um die Pumpe in der Broad Street zeigen.
(6) Leider konnte er im Wasser der Pumpe *keine Erreger* beobachten. Trotzdem ging er davon aus, dass es sie darin geben müsse und sie nur zu klein für eine Beobachtung seien.
(7) Snow konnte sogar erklären, warum ganz bestimmte Personen in der Broad Street trotzdem nicht sogleich erkrankten. Sie verfügten nämlich – wie z.B. eine Brauerei in der Broad Street – über *eigene Brunnen* für ihr Trinkwasser.
(8) Er konnte auch einen Zusammenhang (wir würden heute sagen eine Korrelation) zwischen den Todesraten von Cholera und der *Wasserqualität* in bestimmten Gegenden von London nachweisen.
(9) Eine Ausnahme bildete die »Chelsea Water Company«, die zwar zunächst auch kontaminiertes Wasser verkaufte, dann aber besondere Methoden der *Wasserreinigung* einsetzte, wie Sandfiltration und Bestrahlung mit Sonnenlicht u.a. Bei ihren Kunden waren die Todesraten dann geringer als bei der Konkurrenz.

Die meisten dieser Fakten (2–5 und 7–9) konnte nur seine Theorie gut erklären, während sie für die Miasma-Theorie eindeutig Erklärungsanomalien darstellten. Seine Auswertung der Korrelationen erfolgte rein intuitiv anhand einer Grafik und ohne den Einsatz von statistischen Hilfsmitteln. Die Karten (im Internet an verschiedenen Stellen zu finden) zeigten den Zusammenhang sichtbar auf und die Frage war vor allem, wie er sich erklären lässt. Im Vergleich der Erklärungskraft schneidet hier die snowsche Infektionstheorie so deutlich viel besser ab als die Miasma-Theorie, dass das ein klarer Hinweis darauf war, dass die Infektionstheorie die Zusammenhänge richtig *erklärt*. Von einer wirklichen Erklärung können wir allerdings nur sprechen, wenn wir annehmen, dass Snows Theorie auch wahr ist und die vermuteten

Erreger nicht nur fiktive Gegenstände sind, sondern sich tatsächlich ungefähr in der beschriebenen Weise verhalten.

Das wurde schließlich weiter gestützt durch die bald folgende Entdeckung von großen Zahlen von Bakterien im Darminhalt der Choleraopfer, die allerdings immer noch nicht sogleich zum Durchbruch für die Infektionstheorie führten, da die Miasma-Theorie noch zu stark in den Köpfen der Mediziner verankert war. Aufhalten ließ sich die neue Theorie aber dann nicht mehr lange. Und auch im Rückblick scheint die Argumentation für die Infektionstheorie sehr überzeugend zu sein. Dabei sind keine quantitativen Wahrscheinlichkeiten im Spiel und es wäre auch wohl sehr schwierig, für die beteiligten Phänomene Wahrscheinlichkeiten oder Likelihoods zu vergeben, auf die etwa der Bayesianer angewiesen ist, den wir noch kennenlernen werden. Hier haben wir es vielmehr mit einem qualitativen *Schluss auf die beste Erklärung* zu tun. Wir sammeln dazu zunächst alle bekannten Fakten zu einem bestimmten Thema und erheben vielleicht noch neue. Dann stellen wir eine Liste von möglicherweise diese Fakten erklärenden Theorien auf und vergleichen schließlich, welche der Theorien die meisten der Fakten erklären kann und im Notfall auch noch, welche der Erklärungen besser sind und welche Theorie die besten Erklärungen abgibt. Auf die Theorie setzen wir sodann unsere größten Hoffnungen und akzeptieren sie vorläufig.

Das Relevanzproblem des hypothetisch-deduktiven Schließens soll beim abduktiven Schließen dadurch gelöst werden, dass wir nicht nur verlangen, dass die Daten ableitbar sind, sondern dass die Theorie die Daten auch tatsächlich *erklärt*. Der Schluss auf die beste Erklärung versucht so, den bisher angeführten Überlegungen Rechnung zu tragen. Er eliminiert bestimmte Hypothesen von unserer Liste durch ihre Erklärungsanomalien und bezieht letztlich die metaphysischen Aspekte des induktiven Schließens mit ein, denn wir werden sehen, dass sie gerade bestimmte Aspekte wissenschaftlicher Erklärungen darstellen. Die Eliminationen bestimmter Hypothesen sind meist *weiche Falsifikationen* in dem Sinne, dass es sich nur um *hartnäckige Erklärungsanomalien* für diese Hypothesen handelt, die dann zu ihrer Elimination führen. Wir müssen jedenfalls nicht behaupten, dass unsere Daten sich logisch zwingend gegen die Hypothesen selbst richten.

4.2 Das Grundschema des Schlusses auf die beste Erklärung

Der Schluss auf die beste Erklärung besteht demnach schematisch zunächst aus zwei Schritten ähnlich der eliminativen Induktion. Im ersten Schritt versuchen wir eine Liste von *potentiellen Erklärungshypothesen* aufzustellen. Diese sollte zum einen nicht zu groß sein, sondern noch übersichtlich bleiben, zum anderen sollte sie jedoch keine relevanten Möglichkeiten übersehen. Unser Hintergrundwissen erlaubt es uns in der Praxis häufig, solche Auswahllisten zu erstellen. Das belegen schon die genannten Beispiele und ebenso die noch folgenden. Im zweiten Schritt versucht man eine Selektion möglichst nur einer Theorie vorzunehmen. Zunächst durch Elimination einiger Hypothesen und dann durch Vergleich der restlichen Hypothesen im Hinblick auf ihre Erklärungsstärke für die vorliegenden Daten. Das heißt, der zweite Schritt beinhaltet genaugenommen wiederum zwei unterschiedliche Verfahren. Die Elimination muss keineswegs eine strikte deduktive Falsifikation sein, sondern es genügt oft schon, wenn eine Hypothese eine Reihe von *Erklärungsanomalien* aufweist, während mindestens eine Konkurrenzhypothese deutlich mehr Phänomene erklären kann.

Schurz (2008) hat eine hilfreiche Typologie abduktiver Schlüsse aufgestellt, die sich vor allem daran orientiert, welche Art von Konklusion wir jeweils erzielen möchten und welche Art von Hintergrundwissen wir dabei voraussetzen. Wenn wir bereits wissen, dass Alkoholgenuss zu Kopfschmerzen führt, können wir unter bestimmten Umständen aus vorliegenden Kopfschmerzen auf die singuläre Tatsache zurückschließen, dass vermutlich ein Alkoholgenuss vorausgegangen ist. Das können wir etwa als bloße *Faktenabduktion* beschreiben. Informatiker denken bei ihren Charakterisierungen des abduktiven Schließens meist an diesen Typ von Schluss. Wir können aber auch aus mehreren Episoden von Alkoholkonsum und darauf folgendem Kopfschmerz auf die allgemeine Regel schließen, dass Alkohol zu Kopfschmerzen führt. Das können wir als eine *Theorienabduktion* kennzeichnen. Das sind für die Wissenschaft normalerweise die spannenderen Schlüsse, denn dabei geht es darum, die nomischen Muster zu ermitteln, die einem Phänomenbereich zugrundeliegen, weshalb sie hier im Vordergrund stehen.

Schurz unterscheidet vor allem zwischen rein *selektiven Schlüssen* und *kreativen* Formen der Abduktion. In den rein selektiven Fällen wählen wir nur noch in einer gegebenen Auswahlliste den besten Kandidaten. Man könnte sagen, der erste Schritt der Abduktion wird hier bereits durch unser Hintergrundwissen vorgegeben. In den kreativen Abduktionen führen wir dagegen sogar *neue Begriffe* und mit deren Hilfe neue Hypothesen ein. Allerdings kann man auch den Schritt zu ganz neuen Hypothesen (wenn auch mit den alten Begriffen) bereits als eine Form von Kreativität bezeichnen. Wir können also von *begrifflicher Kreativität* und reiner *Hypothesenkreativität* sprechen. Fallstudien zeigen uns, dass abduktives Schließen in der wissenschaftlichen Forschung oft ein längerer Prozess ist, in dem beide Elemente (die bloße Auswahl und kreative Aspekte) eng miteinander verwoben sind. Wir dürfen also nicht erwarten, dass die genannten Schritte in fester Reihenfolge und dann endgültig stattfinden, sondern es handelt sich vielmehr um einen längeren komplexen Abwägungsprozess, der insbesondere zu neuen Datenerhebungen oder neuen Hypothesen führen kann.

Gerade in der Wissenschaft befinden wir uns häufiger in der epistemisch komfortablen Situation, dass eine Auswahl ganz bestimmter Hypothesen bereits »vorgegeben« ist und wir nur noch entscheiden müssen, welche davon die beste ist. Wenn wir heutzutage z.B. in der Medizin nach der Ursache einer neuen Krankheit suchen, haben wir bereits eine Liste von möglichen Ursachentypen im Hinterkopf, die wir bereits im Rahmen der eliminativen Induktion in Kapitel 3 beschrieben haben. In dem geschilderten Fall der Untersuchung von Rinderwahnsinn war man durch einen rein selektiven Prozess schnell zu der Überzeugung gekommen, dass es sich um eine Infektionskrankheit handelt. Die entsprechenden Phänomene (das Fleisch infizierter Tiere war ansteckend) können die anderen Hypothesen praktisch nicht erklären und konnten so eliminiert werden. Dann aber erfolgte ein kreativer Schritt (begrifflicher Art und Hypothesenkreativität) in Form der Prionenhypothese. Der kreative Prozess wurde hier durch die eliminative Abduktion vorbereitet (man wusste schon, in welchem Hypothesenbereich die Lösung liegen sollte: Infektionskrankheit), aber der kreative Schritt selbst ist nicht durch ein spezielles Schlussverfahren als Entdeckungsverfahren zu erreichen. Die Abduktion diente vielmehr der nachträglichen Rechtfertigung der

Prionenhypothese als der besten verfügbaren Erklärung für die bekannten Fakten. Je genauer sie die Details der Krankheit zu erklären weiß, um so besser wird sie dadurch bestätigt. Ganz ähnlich lässt sich das Beispiel der Entdeckung des Kindbettfiebers als eine Form von Abduktion beschreiben, die selektive und kreative Elemente enthält.

Welche Faktoren unseres Hintergrundwissens die Auswahlliste bestimmen, ist von Fall zu Fall verschieden. In der Basisdisziplin Physik nehmen wir an, die vier Grundkräfte unserer Welt zu kennen. Daran drohen etwa astrologische Erklärungen zu scheitern, weil sie nicht erklären können (nicht einmal andeutungsweise), wie der Einfluss der Gestirne bei unserer Geburt auf unsere Gene so vonstattengehen soll, dass jeweils ganz bestimmte Charaktereigenschaften dabei herauskommen. So kann u.a. die Physik uns helfen, die Listen nicht zu sehr ausufern zu lassen, weil viele Hypothesen sich als inkompatibel zu unseren Rahmentheorien erweisen. Das bedeutet aber natürlich nicht, dass sie nicht später noch eine zweite Chance erhalten können. Das ist etwa dann möglich, wenn alle Kandidaten aus unserer Liste inzwischen eliminiert werden konnten.

Die Liste eines Detektivs wird von einfachen ersten Überlegungen zu *Motiv* und *Gelegenheit* bestimmt. Man könnte den ersten Schritt als ersten groben Eliminationsschritt bezeichnen. Von allen möglichen Hypothesen bleiben nur noch die übrig, die zum einen gemäß unserem Hintergrundwissen eine Erklärung für die Tat erbringen könnten und zum anderen nicht als zu utopisch (d.h. als nicht zu inkohärent zu unserem Hintergrundwissen) sofort ausgeschlossen werden. Wie gesagt: Dieser Schritt ist irrtumsgefährdet, und er ist auch immer wieder aufrollbar. Prusiner hat seine neue Hypothese erst entwickelt, als die anderen Hypothesen schon versagt hatten. Die Schritte sind also nicht als endgültige Verfahrensschritte in einer bestimmten Reihenfolge zu denken. Der erste Schritt ist nur ein *grober Kohärenzfilter* plus kreativen Anteilen, während dem zweiten die Feinarbeit überlassen bleibt, konkurrierende Erklärungen genau miteinander zu vergleichen. Beide Schritte können immer wieder zum Einsatz kommen. Auch wenn wir hier die Induktionsschlüsse gerne als fast algorithmische Verfahren darstellen, sollten wir doch immer im Blick behalten, dass sie das nicht sind, sondern nur Schritte in einem komplexen Abwägungsprozess,

für den wir nur die Ziele und bestimmte Teilschritte sowie gewisse Bewertungskriterien angeben können.

Im zweiten Schritt starten wir dann die Detailarbeit mit einer Liste von Hypothesen H_1,...,H_n mit ihren jeweiligen Erklärungen der Daten d_1,...,d_m. Zunächst suchen wir nach Hypothesen bzw. ihren zugeordneten Erklärungsansätzen, bei denen einige der Daten nicht erklärt werden können, obwohl die Theorien diese Daten oder Phänomene erklären können sollten. Wenn eine Gravitationstheorie bestimmte Planetenbewegungen nicht erklären kann, so ist das eine eindeutige *Erklärungsanomalie* für diese Theorie. Deswegen muss sie nicht sofort eliminiert werden, wenn es keine klar besseren Konkurrenten gibt. Wir erinnern uns an Lakatos und sein Motto vom Ende der Sofortrationalität. Die newtonsche Gravitationstheorie war in ihrer Anfangszeit mehrfach in dieser Situation. Da sie sehr viele Phänomene erklären konnte und keine ernstzunehmenden Konkurrenten besaß, wurde sie sinnvollerweise beibehalten, da sie zumindest einige Phänomene gut erklären konnte. Sollte eine Theorie allerdings viele Erklärungsanomalien aufweisen, und es gibt erkennbar bessere Konkurrenten, sollte sie eliminiert werden.

Diese Eliminationen und weitere Erklärungsvergleiche verlangen letztlich einen komplexen holistischen Gesamtvergleich der Hypothesen (s.u.). Insbesondere müssen wir dabei auch die positiven Erklärungsleistungen der Konkurrenten berücksichtigen. Damit gehen wir über den ersten Schritt hinaus. Es beginnt ein kleinteiliger Vergleich der einzelnen Erklärungsleistungen. Ein großer Vorteil der Stufen-Konzeption ist dabei, dass wir uns auf dieser Stufe auf *paarweise* Vergleiche zwischen endlich vielen potentiellen Erklärungen beschränken können. Das ist oft sehr viel einfacher, als eine absolute Erklärungsstärke bestimmen zu wollen. Vor allem Paul Thagard (2000) und seine Mitstreiter haben in zahlreichen Fallstudien aus der Wissenschaft, aber ebenso für Fälle in Gerichtsverfahren, gezeigt, wie ein solcher Vergleich vorgenommen werden kann.

In der Wissenschaft müssen wir abwägen, welche Theorie mehr Phänomene mit weniger Hilfsannahmen mindestens gleichgut erklären kann, um unseren Favoriten zu bestimmen. Dazu werden wir uns im nächsten Kapitel mit den unterschiedlichen Aspekten der Erklärungsstärke beschäftigen. Ob wir die am besten erklärende Theorie

dann akzeptieren, hängt noch davon ab, wie groß der Vorsprung zu ihren Konkurrenten ist und für wie gut wir die Erklärungsleistung der Theorie insgesamt erachten. Hier sollten wir also Sherlock Holmes widersprechen, der meinte, man müsse den nach der Elimination übrig bleibenden Kandidaten akzeptieren, ganz gleich wie unwahrscheinlich er sei (für eine kritische Darstellung des abduktiven Schließens s.a. Klärner 2003). Die Erklärungsleistung der besten Hypothese muss jedoch noch einen deutlich positiven Beitrag zur Gesamtkohärenz unseres Meinungssystems bieten, sonst sollten wir lieber Agnostiker bleiben und keine der Hypothesen akzeptieren.

Grundschema des abduktiven Schließens

(1) Wir beobachten bestimmte Phänomene d_1,...,d_m aus einem bestimmten Bereich.
(2) Wir wählen eine möglichst umfassende Liste von potentiell erklärenden Hypothesen H_1,...,H_n dazu aus, die noch einigermaßen plausibel erscheinen bzw. keine größeren Inkohärenzen zu unserem weiteren Hintergrundwissen aufweisen (grober Kohärenzfilter).
(3) Wir eliminieren die Hypothesen, die einige der Daten nicht erklären können (Erklärungsanomalien).
(4) Wir vergleichen die verbleibenden Hypothesen im Hinblick auf ihre Erklärungsqualität für möglichst viele der Phänomene d_1,...,d_m und erhalten als Sieger etwa H_1.
(5) H_1 liefert deutlich bessere Erklärungen als seine Konkurrenten H_2,...,H_n.
(6) H_1 trägt (erheblich) positiv zur Gesamtkohärenz unseres Überzeugungssystems bei.
(7) *Schlussfolgerung*: H_1 ist als vermutlich wahr zu akzeptieren.

Der *Schluss auf die beste Erklärung* oder auch das *abduktive Schließen* stellt eine Weiterentwicklung der deduktiv-hypothetischen Theorienbestätigung und der eliminativen Induktion dar und vereint so die bestätigenden und falsifizierenden Aspekte einer Theorienwahl in einem Verfahren. Ich unterscheide hier der Einfachheit halber wiederum nicht

weiter zwischen abduktiven *Schlüssen* und abduktiven *Rechtfertigungen* und gehe außerdem davon aus, dass alle Abduktionen auch Schlüsse auf die beste Erklärung sind. Andere Arten von Abduktionen werden hier nicht betrachtet. (Für mögliche Beispiele solcher Abduktionen s. Gabbay & Woods 2006, §2.)

Dazu einige weitere Beispiele, die das Verfahren weiter illustrieren können: Sehen wir vor uns Fußabdrücke im Schnee, so schließen wir sogleich, dass dort ein Mensch entlang gegangen sein muss. Dieser Schluss lässt sich am besten als Schluss auf die beste Erklärung rekonstruieren. Bei normalem Hintergrundwissen verfügen wir kaum über eine andere Erklärung für die beobachteten Spuren. In solchen Fällen verkürzt sich also das beschriebene Verfahren und solche Fälle finden wir an vielen Stellen unseres Alltags. Kommt jemand nass von draußen herein, nehmen wir sogleich an, dass es draußen regnet, denn das wäre eine gute Erklärung für seine Nässe. Allerdings fallen einem in Ausnahmefällen noch andere Erklärungen ein, die wir unter Umständen genauer zu betrachten haben.

Speziell Detektive verfahren typischerweise so. Wenn Sherlock Holmes anhand von Indizien (dazu sollen auch Augenzeugenberichte gehören) nach dem Mörder von Fritz sucht, so stellt er in einem ersten Schritt eine Liste der potentiellen Täter zusammen. Das sind etwa diejenigen, die über ein Motiv und die Gelegenheit verfügten, Fritz zu töten. Vorsichtshalber sollte man diese Liste eher größer halten und vielleicht alle aus dem Umfeld des Ermordeten hinzunehmen, die über kein Alibi verfügen, auch wenn wir noch kein Motiv erkennen können. Jeder Verdächtige führt zu einer entsprechenden Hypothese über die Ursachen der Tat. In einem zweiten Schritt versucht Holmes dann, möglichst viele dieser Hypothesen zu eliminieren. Gegenüber Watson stellt er das manchmal so dar, als ob es sich um ein deduktives Verfahren handeln würde: Eliminiere alle Verdächtigen bis auf einen, dann ist derjenige der Täter, ganz gleich wie unwahrscheinlich das auch ist.

Diese Darstellung ist aber aus gleich drei Gründen nicht ganz richtig: Erstens ist das Verfahren nicht deduktiv. Es bleibt bei der Auswahl der Verdächtigen und ebenso im Verfahren der Elimination immer ein Irrtumsrisiko. Wir können etwa einen Verdächtigen übersehen haben oder einen fälschlicherweise ausschließen. Zweitens muss der

Übrigbleibende bzw. die entsprechende Hypothese tatsächlich die Indizien gut erklären können. Wenn etwa Fingerabdrücke auf der vermutlichen Tatwaffe gefunden wurden, die nicht von ihm stammen, sollten wir das zumindest gut erklären können. Drittens kann natürlich leicht mehr als ein Verdächtiger übrig bleiben und wir müssen dann trotzdem nach einer Entscheidung suchen. Das geschieht typischerweise so, dass wir dann in einem dritten Schritt die verbleibenden Kandidaten daraufhin vergleichen, wer am besten zu einer Erklärung der Indizien beiträgt. Das ist typisch für Abduktionen. Wir hatten außerdem schon gesehen, wie schwierig strikte Falsifikationen von Hypothesen sind, daher ist es oft besser, etwas vorsichtiger zu formulieren, dass eine Hypothese bestimmte Daten nicht gut erklären kann und deshalb zurückgewiesen wird. Das passt daher eher in das Abduktionsverfahren.

Das sei noch an einem weiteren Beispiel aus der Wissenschaft erläutert, das von Paul Thagard (1999) ausführlich untersucht wurde. Für die Entstehung von Gastritis und Magengeschwüren gab es zu Beginn der 1980er Jahre nur eine akzeptierte Theorie, nach der sie durch Stress entstehen. Dann kam die Hypothese hinzu, dass sie durch das Bakterium Helicobacter-Pylori verursacht werden. Die neue Hypothese konnte u.a. erklären, wieso eine Neu-Infektion mit dem Bakterium zu entsprechenden Magenbeschwerden führte (Selbstversuch der Ärzte), wieso eine Antibiotikabehandlung so erfolgreich gegen die Erkrankung war und weshalb die bisherigen Behandlungen mit säurebindenden Medikamenten nur kurzfristige Erfolge zeitigten. Daneben gab es jedoch weiterhin große Erklärungslücken, wie etwa die, dass viele Träger des Bakteriums keine Beschwerden aufwiesen.

Aufgrund der größeren Erklärungskraft und weiterer Befunde setzte sich aber die Bakterienhypothese durch (vgl. Thagard 1999), obwohl die Mediziner es zunächst noch einige Zeit für ausgeschlossen hielten, dass überhaupt Bakterien in der sauren Umgebung des Magens überlebensfähig wären. Hier wurde also eine sorgfältige Abwägung der Erklärungsmöglichkeiten der beiden Ansätze erforderlich. Tatsächlich konnte die Stresshypothese die genannten Phänomene kaum oder nur mit gewagten Hilfsannahmen erklären (vgl. die quantitativen Abwägungen in Kap. 4.4). Trotzdem müssen wir natürlich immer wieder die Frage stellen, ob es nicht ganz andere Erklärungen gibt (also andere erklärende Hypo-

thesen oder Theorien) und ob unser Auswahlkandidat tatsächlich eine zufriedenstellende Erklärung für alle bekannten Phänomene aufweist. Damit haben wir das Schema des Schlusses auf die beste Erklärung kennengelernt.

4.3 Erklärungen und Erklärungsstärke

4.3.1 Das hempelsche DN-Schema der Erklärung

Doch was sollen wir unter einer (wissenschaftlichen) Erklärung verstehen, und wie lassen sich Erklärungsstärken bestimmen? Dazu habe ich an anderer Stelle (Bartelborth 2007) für einen Vorschlag argumentiert, den ich in seinen Grundideen kurz vorstellen möchte. Er greift zunächst die hempelsche Idee auf, dass eine Erklärung eines Explanandum-Ereignisses E ein *gutes Argument* dafür sein sollte, dass E auftritt (vgl. Hempel 1965). Das Argument soll uns verständlich machen, warum E stattfand. Es beseitigt unsere Verwunderung darüber, dass das Ereignis E stattfand. Das Argument beschreibt, wie die Welt funktioniert und welche Umstände vorlagen, die somit zu E geführt haben. Dabei soll sich eine *wissenschaftliche* Erklärung auf *Naturgesetze* stützen, d.h. sie zeigt, inwiefern E *nomisch erwartbar* war, wenn man nur die vorliegenden Randbedingungen kennt.

Die hempelsche Konzeption von Erklärung führt zum sogenannten DN-Schema der Erklärung, bzw. der *deduktiv-nomologischen Erklärung*. Gemäß diesem DN-Schema wird ein Ereignis E dadurch erklärt, dass wir es aus wahren Naturgesetzen G und vorliegenden Rand- oder Anfangsbedingungen A *deduktiv ableiten* können. Das Schema soll zeigen, wie sich aus bestimmten Umständen A zwangsläufig E entwickelt hat. Die Gesetze, die diesen Übergang beschreiben, müssen in der Erklärung explizit angegeben werden.

Das deduktiv-nomologische Erklärungsschema

(1) *Naturgesetze*:	G_1,...,G_r und
(2) *Randbedingungen*:	A_1,...,A_m, woraus deduktiv folgt:
(3) *Explanandum-Satz*:	Das Ereignis E findet statt.

Dabei wird verlangt, dass für eine *wahre* und nicht nur *potentielle* Erklärung die Explanans-Bedingungen (1) und (2) wahr sind und dass der Explanandum-Satz E deduktiv aus (1) und (2) folgt. Außerdem müssen die empirischen Naturgesetze für diese Ableitung auch tatsächlich erforderlich sein, denn sonst spielen sie in der Erklärung genau genommen keine wirkliche Rolle, sondern sind nur überflüssiges Beiwerk.

Diese Grundidee ist schon recht plausibel, wenn sie auch Erklärungen noch etwas zu stark idealisiert. Ein einfaches Beispiel sieht wie folgt aus:

(Gesetz)	Alle Metalle dehnen sich aus, wenn sie erhitzt werden.
(Randbedingung)	Das Stück Metall a wurde erhitzt.
(Explanandum Satz)	Das Stück a dehnte sich aus.

So plausibel das Schema auf den ersten Blick wirkt, so viele Fragen und Probleme wirft es leider auf den zweiten Blick auf. Doch bevor wir die diskutieren, möchte ich noch die probabilistische Variante dazu einführen. Die stammt ebenfalls von Hempel und wird als *induktiv-statistisches Erklärungsschema* oder IS-Erklärung bezeichnet. Der Unterschied zum DN-Schema besteht darin, dass die Gesetze nun nicht mehr deterministischer Art sein müssen, sondern auch probabilistische Gesetze sein können.

Wenn wir etwa erklären möchten, warum Franz einen Lungenkrebs entwickelt hat, und wir vermuten, dass sein langjähriges Rauchen dafür verantwortlich ist, dann ist das zugrundeliegende Gesetz nur noch probabilistischer Natur und besagt etwa: »Wer viele Jahre raucht, hat eine hohe Wahrscheinlichkeit, einen Lungenkrebs (schon unter 70) zu entwickeln.« Das können wir nun so abkürzen: »Für alle x gilt: P(Lx|Rx) = 0,9«. Hier geht es um eine bedingte Wahrscheinlichkeit dafür, dass Lx der Fall ist, wenn wir Rx als gegeben annehmen. Dabei steht Rx dafür, dass x viele Jahre geraucht hat, und Lx dafür, dass x einen Lungenkrebs (unter 70) entwickelt. Gleichzeitig haben wir das probabilistische Gesetz nun präzisiert und sprechen von einer 90% Wahrscheinlichkeit für Lungenkrebs.

Das IS-Schema sieht ganz analog wie das DN-Schema aus, nur dass die Gesetze probabilistisch sein können und der Explanandum-Satz, der das Explanandum-Ereignis beschreibt, nur mit einer bestimmten

Wahrscheinlichkeit oder in einem bestimmten Grad aus den Explanans-Bedingungen folgen. Damit wird aus unserer Beispielerklärung:

(Gesetz) Für alle x gilt: P(Lx|Rx) = 0,9.
(Randbedingung) (Rf) Franz hat viele Jahre geraucht. (folgt im Grad 0,9)
(Explanandum Satz) (Lf) Franz entwickelte einen Lungenkrebs.

Das allgemeine IS-Schema sieht im Übrigen aus, wie das DN-Schema mit denselben Anforderungen plus einer weiteren Forderung, die sich daraus ergibt, dass bloße Wahrscheinlichkeitsschlüsse *nichtmonoton* sind. Das heißt, dass eine zusätzliche Prämisse aus einem plausiblen Schluss einen unplausiblen machen kann. Nehmen wir für Franz die zusätzliche Information auf, dass er sehr gute Gene hat (das »Helmut Schmidt Gen«), so sinkt seine Wahrscheinlichkeit für einen frühen Lungenkrebs doch ganz erheblich und das obige Argument verliert seine Gültigkeit. Um dem Rechnung zu tragen, müssen wir das IS-Schema um die Forderung ergänzen, dass alle Informationen über Franz, die für das Explanandum relevant sind, auch in dem Schema angeführt und berücksichtigt werden. Diese Forderung nach *maximaler Spezifität des Explanans* kann sich natürlich nur auf die Eigenschaften von Franz beziehen, die uns bekannt sind und für die wir Gründe haben anzunehmen, dass sie auch probabilistisch relevant für das Explanandum E sind.

Das induktiv-statistische Erklärungsschema

(1) *Naturgesetze*:	$G_1,...,G_r$ (möglicherweise probabilistisch)	
(2) *Randbedingungen*:	$A_1,...,A_m$	(mit hoher Wahrscheinlichkeit folgt:)
(3) *Explanandum-Satz*:	Das Ereignis E findet statt.	

Zu den Anforderungen gehört nun zusätzlich zu denen, die wir vom DN-Schema kennen, die nach *maximaler Spezifität* des Explanans bezüglich dem Vorliegen von E. Außerdem gilt nach Hempel, dass die IS-Erklärung umso besser ist, je höher die Wahrscheinlichkeit ist bzw. je höher der Grad ist, in dem das Explanandum aus dem Explanans folgt. Damit liegen die Grundideen für wissenschaftliche Erklärungen auf dem Tisch. Darum gab es viele Debatten, etwa über die Frage, ob wir

in den Sozialwissenschaften und der Biologie überhaupt Naturgesetze finden oder ob wir nicht in diesen Bereichen und vor allem in der Geschichtswissenschaft auf andere Arten von Handlungserklärungen – wie etwa Gründe-Erklärungen – angewiesen sind und in der Biologie etwa auf funktionale Erklärungen, die anders gestaltet sind. Für diese umfangreichen Fragestellungen muss ich auf Bartelborth (2007) verweisen. Es gibt aber einige klare Fehler in den bisherigen Erklärungsansätzen, die wir in jedem Fall noch hier behandeln müssen, zumal sie sich relativ gut beheben lassen.

4.3.2 Relevanzprobleme

Leider kann die hempelsche Forderung der Deduktion des Explanandums E aus dem Explanans G & A die Idee des Erklärens noch nicht korrekt umsetzen (vgl. dazu Bartelborth 2007). Für die logischen Empiristen war die Relevanz des Explanans für das Explanandum durch die Bedingung der deduktiven Ableitbarkeit (bzw. einer probabilistischen Erwartbarkeit) gegeben. Doch schon ein einfaches Beispiel von *Vorwegnahme* (»preemption«) von Achinstein (1983) belegt, dass das nicht ausreicht:

Gegenbeispiel zum DN-Schema
(1) *Gesetz*: Jeder Mensch, der ein Pfund Arsen zu sich nimmt, stirbt innerhalb von 24 Stunden.
(2) *Randbedingung*: Jones aß ein Pfund Arsen.
(3) *Explanandum Satz*: Jones starb innerhalb von 24 Stunden.

Das Beispiel genügt dem deduktiv nomologischen Schema und das Explanans liefert auch einen guten Grund, das Explanandum-Ereignis zu erwarten, und würde sicherlich eine gute Erklärung abgeben, wenn Jones tatsächlich an der Einnahme des Arsens gestorben wäre. Doch es gibt keine Erklärung für das Ableben von Jones an, wenn dieser stattdessen von einem Bus überfahren wurde, ehe das Arsen anfing zu wirken. Der Bus kam dem Arsen zuvor. Daher trägt das Arsen zur Erklärung von Jones Tod auch nichts bei. Es handelt sich zwar weiterhin um einen logisch gültigen Schluss, aber das Explanans beschreibt insbesondere nicht die *tatsächlichen Ursachen* des Explanandums und stellt somit nicht mehr

die korrekte Erklärung des Todes dar. Um unser Erklärungs-Schema zu verbessern, können wir die Zusatzforderung aufstellen, dass im Explanans eine oder die *Ursache* des Explanandumereignisses genannt werden muss und die genannten Gesetze (als *Kausalgesetze*) zeigen sollen, wie diese Ursache zu E geführt hat.

Für das IS-Schema lassen sich ähnliche Preemption-Beispiele und noch problematischere Fälle anführen. Nehmen wir an, wir hätten zwei Todesschützen S1 und S2, die unabhängig voneinander auf den Diktator D feuern. Dabei handele es sich jeweils um einen indeterministischen Prozess und beide Schützen hätten eine objektive Trefferwahrscheinlichkeit von 0,8. Dass keiner trifft, hat dann (bei kausaler und damit probabilistischer Unabhängigkeit) die Wahrscheinlichkeit 0,04 (als Produkt der Gegenwahrscheinlichkeiten 0,2 ·0,2) und die Todeswahrscheinlichkeit für den Diktator ist somit 0,96. Nehmen wir aber an, dass de facto S1 trifft und S2 danebenschießt. Wir wollen nun erklären, warum D gestorben ist (E). Das Schießen von S2 *erhöhte* zwar die Wahrscheinlichkeit für das Ableben des Diktators (von 0,8 auf 0,96), und es führte zu einer insgesamt *hohen Wahrscheinlichkeit von E*, und der Diktator ist auch gestorben, aber der Schuss von S2 war letztlich doch unbedeutend für seinen Tod und damit auch *erklärungsirrrelevant*. Der Schuss von S2 liefert also ein gutes Argument für E aber keine Erklärung. Beide Schüsse erhöhen zwar in derselben Weise die Wahrscheinlichkeit für den Tod des Diktators, aber da der zweite Schütze de facto danebengeschossen hat, ist nur S1 erklärungsrelevant und nicht S2. In den objektiven Wahrscheinlichkeiten finden sich diese Zusammenhänge aber nicht wieder. (Wir kennen solche Problemfälle aus der Debatte um eine Explikation probabilistischer Kausalzusammenhänge.) Eine Wahrscheinlichkeitserhöhung ist demnach nur dann erklärend, wenn wir das erhöhende Ereignis auch als *Ursache* des Explanandums-Ereignisses betrachten können.

Wesley Salmon hoffte einige andere Gegenbeispiele gegen das IS-Schema dadurch ausschalten zu können, dass er vor allem eine *probabilistische Relevanz* des Explanans für das Explanandum E forderte: P(E|Explanans) > P(E). Doch das Todesschützenbeispiel zeigt zugleich, dass uns das in diesem Fall auch nicht rettet. Es sind inzwischen viele weitere Problemfalltypen bekannt, die nach einer Verbesserung der Konzeption verlangen.

Ein typisches Beispiel findet sich in Korrelationen, die keine direkten Kausalbeziehungen darstellen. Wenn das Barometer fällt, erhöht das die Wahrscheinlichkeit für einen Sturm. Nehmen wir an, diese Korrelation wäre sehr stark. Dann liefert das Fallen des Barometers gute Gründe, an das Auftreten eines Sturmes zu glauben, aber natürlich keine Erklärung dafür. Erklären können das Auftreten des Sturmes nur seine tatsächlichen Ursachen.

4.3.3 Eine verbesserte Erklärungskonzeption

Wir benötigen in jedem Fall eine Relevanzbeziehung, die über die logischen Beziehungen in unserer Darstellung der Erklärung hinausgeht. Wir sind außerdem gezwungen, uns auf einen *objektiv* vorliegenden Zusammenhang zu berufen (vgl. Bartelborth 1996, Kap. VIII.D.1). Das geht wiederum über den Rahmen der klassischen empiristischen Ansätze hinaus. Ein naheliegender Vorschlag, der zumindest die meisten Formen von Ereigniserklärungen abdeckt, ist für die Relevanzbeziehung zu verlangen, dass bestimmte Randbedingungen im Explanans *genuine Ursachen* des Explanandum-Ereignisses sind und das Gesetz als *Kausalgesetz* den entsprechenden kausalen Zusammenhang angibt, der hier realisiert wurde.

Das hat auch etwas damit zu tun, dass die erklärenden Ereignisse oder Eigenschaften typischerweise *Unterschiedsmacher* für das Auftreten des Explanandum-Ereignisses E sind. Ohne sie wäre E nicht eingetreten. So setzt z.B. auch Michael Strevens (2011) in seiner neuen Erklärungskonzeption ganz auf diesen Aspekt unseres kausalen Netzes als Relevanzbeziehung für Erklärungen. Wir werden in Kapitel 7 lernen, dass Ursachen vor allem Unterschiedsmacher sind und daher eine entscheidende Rolle in solchen Ereigniserklärungen spielen sollten.

Oder wir können noch etwas allgemeiner verlangen (und damit Bird 2005 und Bartelborth 2007 folgen), dass das vorliegende Gesetz hier *tatsächlich instantiiert* wurde (ähnlich wie eine einfache Eigenschaft instantiiert wird). Hier finden sich bereits die metaphysischen Aspekte des Erklärens, um die wir jedoch nicht herumkommen, wenn wir eine angemessene Erklärungstheorie entwickeln wollen. Sie müssen daher in

unserem neuen Erklärungsschema berücksichtigt werden (vgl. Bartelborth 2007).

Wir müssen außerdem vom DN-Schema noch darin abrücken, dass wir keine *strikten* Naturgesetze verlangen können. In den Sozialwissenschaften oder der Medizin werden wir keine strikten ausnahmslosen Gesetze finden, sondern bestenfalls *Normalfallhypothesen* (Schurz 2004) bzw. Generalisierungen mit einer bestimmten Stabilität, die ich *nomische Muster* genannt habe (vgl. Kap. 3.8). Das verbesserte DN-Schema sieht dann so aus:

Das verbesserte deduktiv-nomologische Erklärungsschema

(1) *Nomische Muster*:	G_1,...,G_r
(2) *Randbedingungen*:	A_1,...,A_m
(3) *Zusatzforderung 1*:	Mindestens ein A_i ist eine *Ursache* von E.
(4) *Zusatzforderung 2*:	Mindestens ein G_i beschreibt als *kausales Muster*, wie A_i zu E führt.
(5) *Explanandum-Satz*:	Das Ereignis E findet statt.

Um sich nicht nur auf einfache Ursachenerklärungen von Ereignissen festzulegen, können wir auch noch eine etwas allgemeinere Variante der Erklärungstheorie formulieren, die für den einfachen Fall nur eines beteiligten nomischen Musters wie folgt beschrieben werden kann (vgl. Bartelborth 2007):

Neues Erklärungsschema

A $\Longrightarrow$ E	(es gilt ein nomisches Muster: As führen zu Es)
Am	(eine Instantiierung m von A liegt vor)
Am $\Rrightarrow$ Es	(die Instanz s von E wurde *verursacht* von Am bzw. das Muster A$\Rightarrow$ E wurde für s und m *instantiiert*)
Also: Es	(eine Instantiierung s von E liegt vor)

Dabei ist es wieder erkennbar von Bedeutung, dass *nomische* Muster vorliegen und nicht nur irgendwelche Generalisierungen. Es gibt zwar eine Vielzahl möglicher Prädikate, aber nur einige davon beziehen sich

auf erklärende (substantielle) Eigenschaften. Wir würden z.B. »grue«-artigen Prädikaten (im Sinne von Nelson Goodmans »grue«-Paradox) nicht zubilligen, dass sie auf erklärende Eigenschaften verweisen.

Intrinsische Eigenschaften. Auch viele andere relationale Eigenschaften eignen sich nicht für wissenschaftliche Erklärungen wie z.B. die Eigenschaft von Franz »größer zu sein als eine bestimmte Ute«. Wir erwarten nicht, dass wir einen Wachstumsschub von Franz so erklären können: Er war immer größer als Ute und die hatte einen Wachstumsschub, daher hatte er ebenfalls einen. Selbst wenn Franz immer größer war als Ute, erklärt das nicht *seinen* Wachstumsschub. Franz' Größer-sein-als-Ute ist normalerweise keine solche Eigenschaft, die eine intrinsische stabile kausale Eigenschaft von Franz darstellt. Derartige Eigenschaften können sich schnell ändern, ohne dass sich Franz verändert, was ihrer Stabilität schon im Wege steht.

Typischerweise erklären wir das Verhalten eines Systems S anhand seiner *intrinsischen Eigenschaften* (hier also etwa anhand der hormonellen Situation von Franz), natürlich durchaus in Reaktion auf äußere Einflüsse. Aber die verwendeten Eigenschaften sollten selbst zunächst Eigenschaften des jeweiligen Systems S sein, die es dann disponieren, in bestimmter Weise auf solche Einflüsse zu reagieren. Die erklärenden Eigenschaften müssen jeweils gewisse stabile und natürliche Charakteristika der Systeme darstellen. Das erwarten wir von intrinsischen Eigenschaften dieser Systeme, jedoch nicht von kontingenten relationalen Beziehungen zu anderen Objekten. Die Frage, was wir unter der gesuchten Invarianz bzw. Stabilität genau zu verstehen haben, habe ich im Abschnitt über nomische Muster schon beantwortet. Man könnte nun hinzufügen, dass nomische Muster stabile Zusammenhänge zwischen (basalen) intrinsischen Eigenschaften sind (bzw. die Repräsentationen dieser Zusammenhänge, wenn wir darunter eher die Beschreibungen als die Zusammenhänge selbst verstehen).

Beispielerklärungen. Betrachten wir als Beispiel ein bestimmtes nomisches Muster G, in dem wir es mit zwei oder mehr Zufallsvariablen A und E zu tun haben und diese unsere beteiligten kausalen Faktoren darstellen (vgl. Woodward & Hitchcock 2003, Halpern & Pearl 2005). Dabei kann es sich durchaus auch um dichotome Variablen handeln, die qualitative

Prädikate repräsentieren, aber genauso gut um quantitative Größen. Wir müssen uns jedenfalls zunächst um eine genaue Charakterisierung dieser Funktionen bemühen. Mein Vorschlag dazu ist: Es seien A(x,s,t) und E(x,s,t') Funktionen, die bestimmten Objekten oder Systemen x in einer bestimmten Situation s (unter bestimmten Bedingungen) zu einem Zeitpunkt t bzw. t' jeweils eine reelle Zahl zuordnen, die die Ausprägung der Eigenschaften A bzw. E darstellen. Durch die Wahl von t' möchte ich die Möglichkeit eröffnen, dass die Wirkung zeitlich nach der Ursache eintritt. Außerdem soll die Funktion f darstellen, wie die Faktoren (A könnte mehrere Faktoren beinhalten) zusammenwirken, um E hervorzubringen. Zusammen mit einem Faktor für mögliche Abweichungen oder Fehler bzw. nicht beachtete Restfaktoren U erhalten wir so die typische Darstellung solcher kleinen Theorien (vgl. etwa Pearl 2000, Kap. 5; Woodward 2003, Kap. 7):

(G) $E(x,s,t') = f(A(x,s,t)) + U(x,s,t)$

Denken wir als Beispiel an den Fall, dass ein größerer Funke (F) einen Brand (B) in einer Scheune verursacht. Dazu kommen die Faktoren Sauerstoff (S) und die Abwesenheit von Nässe (N) und alle Faktoren seien dichotom und nehmen nur die Werte 0 und 1 an. Dann erhalten wir (wobei wir U im Folgenden meist weglassen):

(1) $B = F \cdot S \cdot (-N)$

Hier müssen drei Faktoren zusammenwirken, um die Wirkung zu erzielen. Die Variablen können wie in unserem Beispiel dichotome Größen aber ebenfalls quantitative Größen darstellen. Die Darstellung als Funktionsgleichung legt das keineswegs schon fest.

Wählen wir als zweites Beispiel einen Zusammenhang zwischen der Menge der eingenommenen Medizin M und der Anzahl N der Nebenwirkungen des Medikaments. Außerdem sei $a > 0$ eine reelle Zahl sei. Dann hat in einem bestimmten Bereich Γ unsere Gleichung vielleicht die Gestalt:

(2) $N = aM$

Das heißt, die Anzahl der Nebenwirkungen oder der Umfang der Nebenwirkungen (N) steigt linear mit der Menge des eingenommenen Medikaments (M). Unsere kleine Theorie sollte genau genommen neben der Gleichung (2) immer die Angabe des Definitionsbereichs Γ für die Funktion f enthalten, für den die Gleichung gilt, sonst ist sie unvollständig.

Die Idee der *nomischen* Muster war nun, dass in (G) A nur dann eine Ursache von E darstellt, wenn es eine *Intervention* an A relativ zu E gibt, so dass, wenn A bei ansonsten gleicher Situation s einen anderen Wert annähme, damit auch der Wert von E ein anderer wäre. [Hinweis zur Notation: »A« und »E« werden hier sowohl für die generischen Ereignisse oder Sachverhalte wie auch für ihre Instanzen eingesetzt und der Kontext sollte klären, was genau gemeint ist.] Eine Erklärung für das Auftreten von Nebenwirkungen in einem bestimmten Umfang N* könnte dann so aussehen:

Erklärung für das Auftreten der Nebenwirkungen N*
N = aM + U ist ein nomisches Muster und die Person bekam das Medikament in der Dosierung M* mit N* = aM*, deshalb waren Nebenwirkungen vom Umfang N* (mit gewissen möglichen Abweichungen, die durch ein geeignetes U zu beschreiben wären) zu erwarten.

U ist dann eine Zufallsvariable, die eine bestimmte Wahrscheinlichkeitsverteilung für bestimmte Abweichungen aufweist. Darauf werden wir noch im Kapitel 7 genauer eingehen. Das ist natürlich nur dann eine korrekte Erklärung, wenn die Nebenwirkungen N* tatsächlich durch die Gabe des Medikaments verursacht wurden und N = aM den kausalen Mechanismus korrekt beschreibt. Zu der Erklärung gehört also die Angabe des instantiierten nomischen Musters bzw. des tatsächlichen kausalen Mechanismus sowie die Angabe der Randbedingungen, die vorliegen müssen, damit der Mechanismus gerade die zu erklärende Wirkung herbeiführt.

4.3.4 Dimensionen der Erklärungsstärke

Welche Stabilität sollte nun ein nomisches Muster bzw. eine erklärende Generalisierung aufweisen, um möglichst *erklärungsstark* zu sein? Für

Woodward und Hitchcock (2003) stehen die oben genannten Interventionen ganz im Vordergrund. Nur sie entscheiden demnach, ob unser nomisches Muster (G) eine gewisse Erklärungskraft besitzt. Dazu muss es als Minimalforderung zumindest *eine* Intervention geben (eine entsprechende Änderung der Werte von A), die zu neuen Werten der Funktion E führen würde. Das ist in Beispielen gut nachvollziehbar. Damit der Funke tatsächlich Ursache des Brandes ist, muss gelten: Hätten wir den Funken verhindert, wäre es auch nicht zu dem Brand gekommen. Ist der Bereich von Interventionen, unter denen (G) für das gerade untersuchte System stabil bestehen bleibt (bzw. die funktionalen Zusammenhänge weiterhin richtig beschreibt), größer, so entscheidet das über die Erklärungsstärke von (G) im Falle eines bestimmten Systems x.

Für die klassische Konzeption der Gesetzesartigkeit und der *Vereinheitlichung* zählt dagegen nur, ob die Gleichung ebenso für andere Objekte x Bestand hat und darüber hinaus, ob sie unter anderen Umständen bzw. für andere Situationen s bestehen bliebe. Damit (G) eine möglichst große Erklärungskraft aufweist, muss es auf möglichst viele andere Systeme x unter möglichst vielen Bedingungen s anwendbar bleiben. Wer hat hier Recht? Meines Erachtens haben beide Konzeptionen ein Stück weit Recht, vor allem wenn es um die Bestimmung von Aspekten der Erklärungsstärke geht, und beide sind dann im Unrecht, wenn sie die Invarianzforderungen der anderen Seite ganz ablehnen (vgl. zum Folgenden Bartelborth 2008).

Die Woodward und Hitchcocksche Stabilitätsforderung möchte ich als *funktionale* oder manchmal auch als *lokale Invarianzforderung* bezeichnen (vgl. Kap. 3.8), weil sie verlangt, dass die funktionale Gleichung (G) erhalten bleibt, bei einer Abänderung des Wertes von A für *ein und dasselbe Objekt oder System* unter denselben Randbedingungen. Die minimale funktionale Invarianzforderung ist die nach Invarianz unter wenigstens einer solchen *Testintervention*. Testinterventionen sollen den Unterschied zwischen einer bloßen Korrelation und echten Kausalbeziehungen aufdecken. Es mag zwar so sein, dass gelbe Finger immer mit einer erhöhten Lungenkrebsrate einhergehen, aber sie sind nicht die Ursache dafür. Wenn wir die anderen Einflussfaktoren wie das Rauchen konstant halten, und nur die Eigenschaft gelbe Finger

auf null setzen (d.h. bei Rauchern etwa die Finger schützen oder säubern), dann sinkt trotzdem ihre Lungenkrebsrate nicht; d.h., dass in diesem Fall die Gleichung nicht mehr korrekt Auskunft darüber gibt, was wir zu erwarten haben. Die Gleichung ist also nicht invariant unter dieser Intervention. Die funktionale Invarianz der Gleichung (G) unter mindestens einer Intervention ist damit geradezu eine Voraussetzung dafür, dass (G) überhaupt kausal interpretierbar ist. Sonst könnte (G) noch eine reine Korrelationsgleichung sein oder sogar nur ein einzelnes Datum darstellen.

Als *globale Invarianzforderung* können wir demgegenüber zunächst an die *Situationsinvarianz* denken, wobei wir verlangen, dass unsere Gleichung ebenso unter anderen Randbedingungen s gilt. Das sollte eigentlich für jede Generalisierung für bestimmte Situationen s erfüllt sein. Es können sich etwa Rahmenbedingungen ändern, die weit entfernt liegen und für den betreffenden Zusammenhang vermutlich völlig irrelevant sind. Das sind die uninteressanten Fälle der Situationsinvarianz. Es gibt aber auch spannendere Fälle und in einem Vergleich zweier Hypothesen kann die Situationsinvarianz daher trotzdem den Ausschlag geben. Erweist sich die Darstellung der Wirkung eines Medikaments als nicht mehr stabil unter einer Änderung der Außentemperaturen, ist diese Wirkung schwächer als eine mit der betreffenden Stabilität.

Die Gleichung sollte außerdem für möglichst viele Objekte (in derselben Situation s) gelten, und wir erwarten damit eine weitere globale Invarianz, nämlich eine gewisse *Objektinvarianz* bzw. *Vereinheitlichung* durch (G). Was wäre, wenn (G) nur für ein einziges Objekt a gelten würde? Das passte nicht zusammen mit unserer Konzeption, wie Eigenschaften wirken, nämlich gleichartig unter gleichen Umständen. Es könnte natürlich de facto so sein, dass bestimmte Umstände nur einmal vorliegen, aber wenn sie sich wiederholen, erwarten wir auch gleiche Wirkungen von denselben Eigenschaften (vgl. Chakravartty 2007).

Außerdem erwarten wir natürlich auch eine starke *Zeitinvarianz* für Paare <t,t'> und <r,r'>, die durch eine bloße zeitliche Verschiebung auseinander hervorgehen, für gesetzesartige Generalisierungen. Wenn sich bestimmte Zusammenhänge zwischen Eigenschaften nur in einem bestimmten Zeitraum zeigen, aber nicht in einem anderen, obwohl die Situationen im Übrigen dieselben sind, so handelt es sich typischerweise

nur noch um eine zufällige Korrelation und nicht um ein gesetzesartiges Muster.

Um diese Invarianzen präziser darstellen zu können, so dass wir uns in der Explikation der Erklärungsstärke darauf stützen können, möchte ich die folgenden Konzepte einführen: Ein System m = <x,s,t,t'> bestehend aus einem Objekt x in einer bestimmten Situation s (die als ein Komplex von gemeinsam instantiierten Eigenschaften oder Randbedingungen zu verstehen sind) und zwei Zeitpunkten t und t' mit t≤t' nenne ich ein *potentielles Modell* unserer Minitheorie G (mit Definitionsbereich Γ). Die potentiellen Modelle bilden praktisch den Definitionsbereich für unsere Variablen A und E, die die betrachteten Eigenschaften wiedergeben. Wenn m dazu die Gleichung G und zugleich eine gewisse funktionale Invarianzforderung erfüllt, so nennen wir m ein *Modell* der Theorie (G).

Die Modelle von G
M(G) = {m; set(A(m)=r) ⇒ E(m) = f(r)} (wobei A(x,s,t,t') gerade dem alten A(x,s,t) und E(x,s,t,t') dem früheren E(x,s,t') entsprechen sollen).

Dabei bedeutet »set(A(m)=r)«, dass die Größe A für das System m per auf den Wert r gesetzt wird (vgl. dazu etwa Pearl 2000, Kap. 5; Woodward 2003, Kap. 7). Wenn in diesem speziellen Fall also die Gleichung (G) erfüllt ist, handelt es sich bei m um ein Modell unserer kleinen Theorie (G). Dann erhalten wir für unseren *lokalen Invarianzbereich* Γ(G) die Charakterisierung:

$$\Gamma(G) = \{r \in \mathbb{R}; \exists m \in M(G) \,\&\, A(m) = r\}$$

Die minimale Forderung nach funktionaler Invarianz besagt nun, dass Γ(G) mindestens zwei Werte r_1 und r_2 enthalten muss mit $f(r_1) \neq f(r_2)$. Man beachte, dass f einfach nur eine mathematische Funktion ist, Γ(G) aber eine empirisch zu bestimmende Größe.

Die Menge M(G) der Modelle unserer kleinen Theorie (G) gibt nun an, auf welche Systeme unsere Theorie erfolgreich anwendbar ist. Damit liefert sie uns im Wesentlichen die globale Vereinheitlichung, zu der unsere Theorie (G) imstande ist. Je größer also M(G) ist, umso vereinheitlichender ist unsere Theorie (G) und umso besser sind ihre Erklärungen. Das beruht vor allem darauf, dass eine große Menge M(G) anzeigt, dass

das nomische Muster (G) ein in unserer Welt grundlegendes und sehr stabiles Muster darstellt.

Erklärungen sollen unsere Erscheinungen möglichst auf solche zentralen Muster in unserer Welt zurückführen. Man könnte sogar sagen, je größer M(G) ist, umso gesetzesartiger ist ceteris paribus (G). Damit haben wir schon einen entscheidenden Parameter für die Erklärungsstärke durch (G) gefunden. Je größer M(G) ist, umso basaler ist (G) und umso besser ist dann eine Erklärung durch (G) als eine Zurückführung auf grundlegende Kausalgesetze in unserer Welt. Je grundlegender der beschriebene kausale Mechanismus in unserer Welt verankert ist bzw. je grundlegender die in (G) beschriebenen Eigenschaften sind, umso besser wird auch die Erklärung durch (G). Das lässt sich am leichtesten erkennen, wenn wir einfache Beispiele dafür betrachten.

Nehmen wir an, wir vergleichen die Erklärungen, die zwei Theorien für bestimmte Aspekte der Umlaufbahn des Mars bereitstellen. Die eine ist die newtonsche Gravitationstheorie und die andere eine reine Marstheorie, die nur für den Mars eine Bahnbeschreibung liefert und so für jeden späteren Zeitpunkt E jeweils die Marsposition angibt. Wollen wir nun erklären, warum der Mars auf einer bestimmten elliptischen Bahn um die Sonne läuft, würde die Marstheorie einfach die Bahnkurve dazu angeben und keine weitergehenden Erklärungen oder Vereinheitlichungen anbieten. Genau genommen könnte man hier wirklich nur noch von einer einfachen Beschreibung und nicht mehr von einer Erklärung sprechen. Die newtonsche Gravitationstheorie bietet dagegen eine Beschreibung des zugrundeliegenden Mechanismus bzw. der zugrundeliegenden dispositionellen Eigenschaft der Gravitationskraft und führt die Bahn des Mars wie auch vieler andere Phänomene auf dieses grundlegende Muster zurück. Erst dadurch erhalten wir eine *Erklärung* der Marsbahn. Diese Form der Vereinheitlichung ist also unabdingbar für das Erklären. Nomische Muster sind daher vor allem dadurch ausgezeichnet, dass sie (im Unterschied zu den bloß lokal invarianten Generalisierungen bei Woodward) intrinsische dispositionale Eigenschaften oder Kräfte beschreiben, die immer wieder dieselben anderen Eigenschaftsvorkommnisse hervorbringen, sobald sie in einer geeigneten Umgebung auftreten. Erst diese globale Invarianz zeigt,

dass es sich hier um ein nomisches Muster handelt, das auch einen bestimmten kausalen Mechanismus in unserer Welt beschreibt.

Allerdings ist die Situation für die Vereinheitlichung und Erklärungsstärke noch etwas komplizierter, denn wir finden ein komplexeres Abhängigkeitsverhältnis zwischen M und Γ(G). Wenn wir uns mit kleineren Bereichen M(G) zufriedengeben, erhalten wir unter Umständen größere Bereiche für Γ und damit einen größeren funktionalen Invarianzbereich. Hier werden also gegenseitige Verrechnungen der zwei unterschiedlichen Formen von Vereinheitlichung möglich.

Betrachten wir die einfache Theorie, dass die Menge eines eingenommenen Medikaments (Variable A) linear wachsend eine Menge an Nebenwirkungen (Variable E) jeweils auf einer Skala von 0 bis 1 verursacht:

(g) $E = cA$, mit $c \in \mathbb{R}$ (hier ist also $f(x)=cx$)

Nun gebe es aber zwei Typen von Menschen, die *Robusten* (r) und die *Sensiblen* (s), und es gebe Menschen in unterschiedlichen Situationen: diejenigen, die noch *andere Medikamente* (m für »mit«) einnehmen und diejenigen, die ansonsten *keine weiteren Medikamente* (o für »ohne«) einnehmen. Dann kann Folgendes der Fall sein: Für die *Robusten ohne weitere Medikamente* (M_{ro}) gilt (g) auf dem ganzen Intervall [0,1], für die *Sensiblen ohne weitere Medikamente* (M_{so}) ergeben sich ab der Dosierung ½ bereits deutlich höhere Raten von Nebenwirkungen als nur linear anwachsende (ebenso für die *Robusten mit weiteren Medikamenten*: M_{rm}) und für die *Sensiblen mit weiteren Medikamenten* (M_{sm}) sogar bereits ab einer Dosierung von ¼. So erhalten wir als lokale Invarianzmengen: $\Gamma(M_{ro}) = \Gamma_{ro}=[0,1]$ sowie $\Gamma_{so}=\Gamma_{rm}=[0,½]$ und $\Gamma_{sm}=[0,¼]$. Wir haben es also mit vier unterschiedlichen Situationen zu tun, die durch die vier Gruppen von Personen charakterisiert sind:

Robuste, die keine weiteren Medikamente einnehmen (M_{ro})
Robuste, die weitere Medikamente einnehmen (M_{rm})
Sensible, die keine weiteren Medikamente einnehmen (M_{so}) und
Sensible, die weitere Medikamente einnehmen (M_{sm})

Dafür treten die drei Bereiche auf, in denen die Gleichung (g) die Nebenwirkungen korrekt beschreibt, wobei wir annehmen, dass die Nebenwirkungen oberhalb der drei Bereiche schneller als linear anwachsen, wenn wir auch nicht genau wissen, in welchem Umfang das jeweils der Fall ist.

Hier zeigt sich dann die enge gegenseitige Abhängigkeit von *Modellmenge* und *Invarianzbereich*. Typischerweise erhalten wir für größere Modellmengen kleinere Invarianzbereiche und umgekehrt. Deshalb haben wir es eigentlich nicht nur mit einer Theorie (g), sondern mit drei Theorien (g_{ro}, $g_{ro+rm+so}$, $g_{ro+rm+so+sm}$) zu tun, in denen unsere Funktion f jeweils einen anderen Definitionsbereich hat. Die drei Theorien besagen dann, dass es eine Konstante c gibt, für die gilt:

(g_{ro})	Für alle $m \in M_{ro}$ gilt:
	$E(m) = cA(m)$ im Bereich $A(m) \in [0, 1]$,
($g_{ro+rm+so}$)	Für alle $m \in M_{ro} \cup M_{rm} \cup M_{so}$ gilt:
	$E(m) = cA(m)$ im Bereich $A(m) \in [0, ½]$,
($g_{ro+rm+so+sm}$)	Für alle $m \in M_{ro} \cup M_{rm} \cup M_{so} \cup M_{sm}$ gilt:
	$E(m) = cA(m)$ im Bereich $A(m) \in [0, ¼]$

Man sieht hieran sehr schön, wie bei größerer Modellmenge (d.h. bei größerer globaler Invarianz der Theorie) ihre lokale Invarianz abnimmt. Die Frage ist dann, welche der drei Theorien die größte Erklärungsstärke aufweist. Das werden wir gleich unter dem Stichwort der Vereinheitlichung weiter diskutieren. Es gibt dazu zwei unterschiedliche Ansichten, was die relevante Vereinheitlichung ist.

Es gibt aber neben diesen zwei Unterdimensionen der *Vereinheitlichung* m.E. eine weitere Dimension, die die Erklärungsstärke eines Datums E durch eine Theorie T bestimmt. Das ist die spezielle *Information*, die uns T über das Auftreten von E gibt, bzw. der Informationsgehalt von T relativ zu E, worauf ich gleich zu sprechen kommen werde (vgl. Bartelborth 2002). Beide Dimensionen lassen sich als plausible Aspekte der Erklärungskraft auffassen und werden auch immer wieder genannt, aber die Frage bleibt bestehen, was genau darunter zu verstehen ist. Außerdem übersehen oder vernachlässigen alle Ansätze jeweils ganz bestimmte Aspekte der Erklärungsstärke.

Vereinheitlichung. Was soll man nun unter *Vereinheitlichung* verstehen? Hitchcock/Woodward (2003, 184ff.) nennen einige Aspekte von Erklärungsstärke, von denen ich die wichtigsten aufgreifen möchte. Für sie zählt im Bereich der Vereinheitlichung vor allem die *lokale Invarianz*. Eine Theorie G mit größerem Anwendungsbereich Γ ist für sie deshalb erklärungsstärker als eine Theorie G* mit kleinerem $\Gamma^* \subset \Gamma$, weil G *mehr kontrafaktische Fragen* darüber beantworten kann, was passiert wäre, wenn A einen anderen Wert aufgewiesen hätte.

Das erscheint auf den ersten Blick noch plausibel, wird jedoch sofort problematisch, wenn wir berücksichtigen, dass eine *größere funktionale Invarianz* mit einer *kleineren globalen Invarianz* einhergehen kann, d.h., es könnte dann M(G*) echt enthalten sein in M(G), ähnlich wie es in unserem Beispiel oben der Fall ist. Das beunruhigt Woodward und Hitchcock nicht, da sie schlicht behaupten, dass eine Invarianz bzgl. der Objekte x keine Relevanz für die Erklärungsstärke besitzt. Für sie wäre also unsere Theorie g_{ro} eindeutig die erklärungsstärkste Theorie.

Es sollte sie jedoch beunruhigen, denn auch sie erkennen an, dass eine größere Invarianz bzgl. der Randbedingungen s die Erklärungsstärke vergrößert. Die steht aber ebenso in einem Spannungsverhältnis zur funktionalen Invarianz wie die Objektinvarianz, wie die Abhängigkeit von der Einnahme weiterer Medikamente in unserem Beispiel belegt. Außerdem wurde schon in den klassischen Vereinheitlichungsansätzen dafür argumentiert, dass auch die Objektinvarianz sowohl für Kausalbehauptungen und insbesondere für Erklärungsbehauptungen eine große Bedeutung besitzt. Deshalb wurde sie bisher in Vereinheitlichungskonzeptionen des Erklärens sogar meistens in den Mittelpunkt gestellt. Woodward und Hitchcock haben Recht, dass damit die ebenfalls wichtige funktionale Invarianz übersehen wurde, aber diese kann die globale nicht ersetzen, sondern ergänzt sie nur. Das wird in unserem Beispiel durch die enge Verzahnung der beiden Typen von Vereinheitlichung deutlich. Damit haben wir bereits im Bereich der Vereinheitlichung zumindest zwei Unterdimensionen der Vereinheitlichung, die in gegensätzliche Richtungen ziehen (außerdem können natürlich ebenso die Objektinvarianz und die Situationsinvarianz in Konflikt geraten).

Wir erhalten somit nur eine partielle Ordnung für die Erklärungsstärke. Doch es kommen noch weitere Aspekte der Erklärungsstärke hinzu.

Woodward und Hitchcock weisen darauf hin, dass Γ (jedenfalls für quantitative Größen A und E) möglichst nicht in Mengen isolierter Punkte zerfallen darf, sondern eher zusammenhängend sein sollte. Wir können zumindest behaupten, dass für eine bessere Erklärung Γ wenigstens offene Intervalle enthalten sollte, so dass benachbarte Systeme normalerweise auf ähnliche Weise durch G erklärt werden können.

Als erstes Resultat erhalten wir damit, dass die Theorie G* dann die besseren Erklärungen liefert als G, wenn ceteris paribus $\Gamma(G) \subset \Gamma(G^*)$ ist oder wenn $M(G) \subset M(G^*)$ ist und außerdem, wenn $\Gamma(G^*)$ zusammenhängender ist als $\Gamma(G)$. Für quantitative Theorien erhalten wir damit als Erklärungsschema für die Erklärung, warum ein bestimmtes Objekt o in einer bestimmten Situation s zu t' gerade die Eigenschaft E im Ausmaß v aufweist:

Erklärungsschema für quantitative Größen

Gesetz:	$\forall z \in M(G)$: $E(z) = f(A(z))$ im Bereich $A(z) \in \Gamma(G)$
Randbedingung:	$m = \langle o,s,t,t' \rangle \in M(G)$ & $A(m) = u \in \Gamma(G)$
Explanandum:	Daher ist: $E(m) = f(u) = v$

Doch es gibt noch einen weiteren Aspekt von Vereinheitlichung zu beachten. Die Vereinheitlichung sollte nicht auf *triviale Weise* erfolgen, sondern durch ein *einheitliches Muster*, das in vielen Fällen instantiiert ist. Das ist ein altes Problem der Vereinheitlichungsansätze, dass die Vereinheitlichung z.B. nicht dadurch zustande kommen darf, dass zwei Theorien mit ganz unterschiedlichen Mustern per Konjunktion zusammengefügt werden. Natürlich können mehrere Muster in einer Erklärung zusammenwirken und etwa einen komplexeren kausalen Mechanismus beschreiben, doch durch die Konjunktion wird dabei keine zusätzliche Vereinheitlichung erzielt. Diese Forderung ist intuitiv verständlich, aber nicht leicht zu präzisieren. Eine ältere Idee ist dazu, dass es zu einem vereinheitlichenden nomischen Muster keine zwei Muster geben darf, deren Konjunktion denselben (empirischen) Gehalt aufweist. Eine formale Präzisierung dieser Idee findet sich in Bartelborth (2002, 1996) im Rahmen der strukturalistischen Theorienauffassung

unter dem Stichwort der *organischen Einheitlichkeit* einer Theorie. Dabei geht es darum, wie eng die Modelle einer Theorie untereinander vernetzt sind oder ob die Modellmenge in zwei oder mehr separate Klassen zerlegt werden kann.

Neben der Vereinheitlichung durch ein nomisches Muster G müssen wir aber auch berücksichtigen, was uns G direkt zum Auftreten von E zu sagen hat. Das ist eine Dimension von Erklärungsstärke, die wiederum in einem Spannungsverhältnis zu den bisherigen Dimensionen steht.

Möglichst hohe Wahrscheinlichkeit für E. Von dem erklärenden nomischen Muster G verlangen wir, dass es eine gute Vereinheitlichung bietet, aber zunächst sollte es auch möglichst *gehaltvolle Informationen* dazu anführen, warum gerade E aufgetreten ist (und nicht etwa F) und warum E in dieser Situation zu erwarten war. Besonders Lipton (1991) hat das *kontrastive Element* von Erklärungen betont. Oft können wir sogar nur erklären, warum *E statt F* aufgetreten ist und nicht, warum E aufgetreten ist, denn wir verfügen vielleicht nur über Gründe, die E gegenüber F bevorzugen, die aber nicht viel mehr besagen. So bin ich vielleicht ins Allgäu statt nach Südtirol gefahren, weil das Allgäu nicht so weit entfernt ist, aber ich kann vielleicht nicht erklären, warum ich überhaupt verreist bin, bei meiner Abneigung gegen das Reisen.

In der Debatte um das induktiv-statistische Modell der Erklärung hat sich gezeigt, dass es für Erklärungen oft nicht ausreicht, wenn das Explanans schlicht zu einer hohen Wahrscheinlichkeit für das Explanandum führt. Es fehlt dabei eine Forderung der *Relevanz des Explanans* für das Explanandum. Daher hat sich die Bedingung der Erhöhung der Wahrscheinlichkeit als plausible Forderung durchgesetzt (das statistische-Relevanz-Modell), die allerdings als Relevanzforderung selbst zu kurz greift. Wir erinnern uns: So erhöhen zwar gelbe Finger die Wahrscheinlichkeit für das Auftreten von Lungenkrebs (durch ihre Korrelation mit dem Rauchen), haben dafür aber keinen Erklärungswert. Deshalb musste auch die SR-Konzeption durch die Forderung nach einer Kausalbeziehung ergänzt werden. Trotzdem hat der Siegeszug der SR-Bedingung dazu geführt, dass die Forderung der hohen Wahrscheinlichkeit des IS-Modells ganz aufgegeben wurde (vgl. Strevens 2000). Das ist ein offensichtlicher Fehler. Die Forderung nach einer möglichst guten Information darüber, warum dieses spezielle Ereignis E

(und nicht andere wie F) aufgetreten ist, findet sich am ehesten in der *Forderung nach möglichst hoher Wahrscheinlichkeit von E* wieder. Wenn P(E|G) größer wird, dann wird P(F|G) damit kleiner, wenn F eine echte Alternative zu E darstellt. Somit ist eine Erklärung von E durch G ceteris paribus umso besser, je höher P(E|G) ausfällt. Für die Relevanzbedingung sind wir ohnehin auf eine Forderung nach einem Kausalzusammenhang bzw. der Instantiierung des nomischen Musters angewiesen.

In unserer Gleichung (G) sind die Wahrscheinlichkeiten nicht explizit aufgeführt, aber intuitiv in der Zufallsvariable U angesiedelt, die wir allerdings in den letzten Darstellungen von (G) aus Gründen der Vereinfachung immer weggelassen haben. Wir haben einfach so getan, als ob (G) eine deterministische Theorie wäre. Doch in vielen Fällen haben wir es mit probabilistischen Theorien zu tun, für die U für jedes m nun eine Wahrscheinlichkeitsverteilung P(U=x) aufweist, die den Abweichungen vom Erwartungswert 0 jeweils eine bestimmte Wahrscheinlichkeit zuweist. Das heißt, U liefert eine reelle Zahl x für jedes einzelne Experiment mit einer Wahrscheinlichkeit P(U=x), die oft als normalverteilt angenommen wird und deren Werte typischerweise als von A unabhängig angesehen werden. Unsere Forderung nach hoher Wahrscheinlichkeit von E besagt dann, dass die Standardabweichung dieser Normalverteilung möglichst klein sein sollte.

Ist U dabei eine Zufallsvariable mit kontinuierlichem Wertebereich, so soll P(U=x) gerade die Dichtefunktion $f_U(x)$ der Verteilung repräsentieren oder wir müssten genau genommen mit kleinen Intervallen I arbeiten und die Wahrscheinlichkeit P(U∈I) betrachten. Wir würden dann genaugenommen nicht erklären, warum E einen ganz bestimmten Wert annimmt, sondern, warum E einen Wert aus dem Intervall I annimmt. Diese kleinen technischen Komplikationen werde ich hier aber weitgehend ausklammern, da sie für uns nicht von großer Bedeutung sind. U dient jedenfalls zunächst dazu, die für E kausal relevanten Faktoren zu repräsentieren, die wir mit A noch nicht erfasst haben, es kann aber ebenso dazu dienen, Messfehler anzugeben oder genuin indeterministische Effekte zu beschreiben.

Statt mit unserer Gleichung (G) zu arbeiten, lassen sich solche indeterministischen Theorien natürlich auch so beschreiben, dass wir direkt E und A als Zufallsvariable betrachten und die Funktion P(E=y|

set(A=x)) hernehmen, die uns direkt die Auswirkungen unterschiedlicher Ausprägungen von A auf E als probabilistisch darstellt. So liefert unsere Theorie bestimmte Wahrscheinlichkeiten für bestimmte Ergebnisse und unsere Forderung besagt, dass unsere Erklärung umso besser ist, je höher diese Wahrscheinlichkeit ausfällt.

Betrachten wir zur Forderung hoher Wahrscheinlichkeiten für das Explanandum ein Beispiel. Nehmen wir zwei Theorien T_1 und T_2. Die erste besagt, dass die Krankheit K mit 5% Wahrscheinlichkeit zum Tode führt, während die zweite für K 90% Todeswahrscheinlichkeit annimmt. Wenn nun Fritz ohne anderen erkennbaren Grund gestorben ist, dann ist die Auskunft, dass er K hatte, zwar in jedem Fall ein relevanter Faktor, der zu einer Erklärung führt, aber im Falle von T_1 bleibt die Frage viel offener, warum es gerade Fritz erwischt hat, wenn doch nur 5 von 100 daran sterben. Die Erklärung mit Hilfe von T_2 ist intuitiv weitaus überzeugender. Strevens (2000) erläutert das an Beispielen aus der statistischen Mechanik, in denen speziell die hohe Wahrscheinlichkeit einen wesentlichen Erklärungsfaktor darstellt. Das heißt, ceteris paribus ist eine Erklärung für E umso besser, umso größer P(E|G) ist (bzw. für kontinuierliche Zufallsvariablen: je größer ihre Dichtefunktion an dieser Stelle ist).

Das lässt sich auch anhand von Überlegungen zur *Stärke der Kausalität* bzw. Bedeutung des in (G) genannten Kausalfaktors erläutern, wobei wir auf die in Kapitel 7 Kausalkonzeption der minimalen Theorie zurückgreifen können (vgl. Bartelborth 2008). Dabei kommt dabei heraus, dass ein deterministisch wirkender Faktor A, der allerdings auf gewisse Kofaktoren angewiesen ist, um zu wirken, umso größeren Einfluss beim Hervorbringen von E zeigt, je verbreiteter seine Kofaktoren sind und umso seltener er selbst auftritt. Das liefert wichtige Hinweise darauf, welche Faktoren wir im Normalfall für besonders erklärungsrelevant erachten.

Die Erklärungsstärke wird dabei nicht nur durch pragmatische Aspekte bestimmt (was manche Autoren wie Lipton 1991 annahmen), sondern ebenso von Verteilungshäufigkeiten in unserer Welt. Ist eine Scheune abgebrannt (E), werden wir im Normalfall dafür die brennende Zigarette (A) zur Erklärung nennen und nicht die (allgegenwärtige) Anwesenheit von Sauerstoff, weil die nur als Kofaktor unsere Wahrscheinlichkeits-

differenz erhöht. Andere notwendige Faktoren sind hingegen seltener und daher ist ihre Nennung in einer Erklärung informativer. Sind ihre Kofaktoren außerdem weitverbreitet, liefern sie gute Erklärungen. Die Erklärungsstärke ist hier an die Stärke des kausalen Zusammenhangs gekoppelt, den wir in Kapitel 7.3.3 weiter explizieren werden.

Vergleiche der Erklärungskraft. Eine einfache Erklärung eines Ereignisses oder einer Tatsache E besteht somit aus zwei Elementen: einem generellen und einem singulären. Das generelle Element ist die Angabe eines nomischen Musters G (A⇒E), wonach generell Ereignisse vom Typ A Ereignisse vom Typ E (bzw. Ereignisse eines Typs, zu dem E gehört) hervorbringen. Zum singulären Element gehört, dass dieses Muster auch tatsächlich in unserem konkreten Fall instantiiert ist, bzw. im Falle kausaler Muster, dass eine Instanz von A vorliegt, die die tatsächliche Ursache einer Instanz von E darstellt. Damit wir überhaupt davon sprechen können, ein *nomisches Muster* sei instantiiert, muss die Minimalbedingung von Hitchcock und Woodward erfüllt sein, wonach es mindestens noch eine Testintervention an A gibt, so dass E nicht aufgetreten wäre (bzw. einen anderen Wert angenommen hätte), wenn A nicht aufgetreten wäre (bzw. einen anderen Wert angenommen hätte). Doch das ist noch eine recht schwache Forderung an Erklärungen und sagt uns wenig darüber, was bessere von schlechteren unterscheidet.

Die Erklärungsstärke selbst ist ein *multidimensionales Konzept*, das für den Vergleich der Stärke von Erklärungen (bzw. für einen Vergleich der erklärenden Theorien) zunächst nur eine Halbordnung liefert. Die Dimensionen für eine einfache Theorie T (mit Muster G) im Hinblick auf eine Erklärung von E haben wir nun beisammen. Sie nehmen Bezug auf die beiden Hauptaspekte des Erklärens. Zunächst muss die Theorie möglichst *gehaltvolle Informationen* über unsere zu erklärende Instanz von E liefern, die sich so zusammenfassen lassen, dass für dichotome Größen P(E|A) – P(E) möglichst groß sein sollte. Die Erklärungsstärke durch T ist also ceteris paribus umso größer, umso größer P(E|A) wird.

Das heißt, dass das Muster (G) eine relativ zu unserer Welt möglichst *starke Ursache* angeben sollte, die deshalb stark ist, weil die erforderlichen Kofaktoren normalerweise vorliegen, während die Angabe von A besonders informativ ist, weil A selbst als keineswegs selbstverständliche Hintergrundbedingung angesehen werden kann, sondern eher unge-

wöhnliche Umstände darstellt. Für quantitative Größen erwarten wir, dass sie den Wert von E möglichst genau spezifizieren, dass die Verteilung P(U) an der betreffenden Stelle also möglichst konzentriert ist (eine kleine Streuung aufweist).

Dazu kommt als Zweites der Aspekt möglichst guter *Vereinheitlichung* durch (G). Der findet sich zunächst in der *funktionalen Invarianz* als Grundlage dafür, dass wir es überhaupt mit einem nomischen Muster zu tun haben. Die Erklärungsstärke ist ceteris paribus umso größer, je umfangreicher Γ(G) ist. Aber informativ wird eine Gleichung G erst, wenn wir es auch mit einer gewissen *globalen Invarianz* zu tun haben. Das gehört seinerseits zu unserem Verständnis der kausalen Wirkungen von dispositionalen Eigenschaften. Hierhin gehört ebenfalls die Forderung, dass möglichst ganze *Phänomene* – also Typen von Situationen und Objekten – insgesamt erklärt werden.

> Die ***Erklärungskraft von (G)*** ist also größer als die von *(G')* für das Vorliegen einer Instanz von E, wenn
> (1) M(G') $\subset$ M(G) gilt, wobei zugleich Γ(G') $\subseteq$ Γ(G) gegeben ist und
> (2) P(E|A)/P(E|A') $\geq$ 1 ist.
>
> (Für Zufallsvariable E mit kontinuierlichem Wertebereich müssen wir den Quotienten der Wahrscheinlichkeitsdichten betrachten $f_G(a)/f_{G'}(a)$ anstelle des Quotienten der Wahrscheinlichkeiten.)

Entsprechende Vergleiche ergeben sich, wenn wir eine der anderen Dimensionen der Erklärungsstärke in den Vordergrund stellen. Insbesondere ist hier die Inklusionsbedingung für die Modelle so zu verstehen, dass sie möglichst für ganze *Klassen* von Modellen gilt, die intuitiv ein *Phänomen* repräsentieren. Die drei so explizierten Aspekte lassen sich womöglich in konkreten Einzelfällen gegeneinander verrechnen, aber es finden sich bisher keine allgemeinen Regeln dafür, wie das geschehen soll. Daher bleibt nur die allgemeine Redeweise, dass eine Theorie T dann gegenüber einer Theorie T* vorzuziehen ist, wenn sie entweder eine größere Vereinheitlichung (in allen Aspekten) oder eine gehaltvollere Ableitung der Daten E zu bieten hat. Dabei ist klar, dass E dabei auch eine Konjunktion von mehreren Datenaussagen darstellen kann.

Komplexe wissenschaftliche Theorien haben eine spezielle Struktur, um die beiden Hauptdimensionen der Erklärungsstärke gemeinsam

zu verwirklichen. Sie bestehen aus allgemeineren Komponenten, in denen sie ihre große *Vereinheitlichungskraft* zeigen und spezielleren Komponenten, in denen sie für kleinere Mengen intendierter Anwendungen *gehaltvollere Muster* zur Verfügung stellen. Die newtonsche Mechanik bietet mit f=m·a zunächst ein sehr allgemeines und nicht sehr gehaltvolles Muster, das dann um spezielle Kraftgesetze etwa für die Haftreibung oder für die Gravitationskraft oder für Federkräfte für spezielle Anwendungen ergänzt wird und erst dadurch einen größeren Gehalt erhält. Im Rahmen der strukturalistischen Theorienauffassung lassen sich solche Strukturen weiter präzisieren (vgl. Bartelborth 2002).

Eine etwas andere Darstellung von Erklärungsstärke findet sich bei Thagard (2007). Dort wird untersucht, welche Anhaltspunkte es in der Entwicklungsgeschichte einer Theorie für uns gibt, dass die Theorie vermutlich wahr oder wenigstens approximativ wahr ist. Thagard setzt sich dabei insbesondere mit der pessimistischen Metainduktion von Laudan (1981) auseinander, nach der sich im Laufe der Wissenschaftsgeschichte schließlich doch die meisten Theorien als falsch herausgestellt haben, selbst wenn sie zunächst eine gute Erklärungsleistung zeigten. Neben dem »broadening«, das Theorien zeigen, bei dem sie zunehmend mehr Phänomene erklären können (was in etwa der Forderung nach möglichst hoher Vereinheitlichung entspricht), setzt Thagard vor allem auf ein »deepening«, bei dem eine Theorie zu immer tieferen Erklärungen führt bzw. neue Theorien, diese vertiefenden Erklärungen liefern. Sie bieten nach Thagard die wichtigeren Hinweise auf die Wahrheit einer Theorie.

Tatsächlich *vertiefte Theorien* haben sich seiner Meinung nach in der Wissenschaftsgeschichte noch nie als ganz falsch erwiesen. Mit einer solchen Vertiefung bezieht sich Thagard speziell auf mechanistische Erklärungen, für die der erklärende Mechanismus selbst schließlich durch noch grundlegendere Mechanismen erklärt wird. Solche Mechanismen sind Beschreibungen von Systemen, die aufzeigen, wie die Komponenten des Systems ein bestimmtes Gesamtverhalten produzieren. Die Vertiefungen beziehen sich dann etwa auf weitere Teil- oder Unterkomponenten der Komponenten, für die wir noch grundlegendere Zusammenhänge aufzeigen können. Das beschreibt sicher auf intuitive Weise, was in einigen solchen Beispielen passiert (Thagard gibt dazu zahlreiche Bei-

spiele aus der Wissenschaftsgeschichte an), aber wir müssen natürlich nachfragen, was genau mit einem Mechanismus gemeint ist.

In Bartelborth (2007) findet sich dazu schon eine Diskussion, die zu dem Ergebnis kommt, dass die Beschreibungen des Zusammenspiels der Komponenten im Wesentlichen durch nomische Muster zu geschehen hat, wenn wir damit genuine Erklärungen produzieren möchten. Die Vertiefung hängt daher auch mit gehaltvolleren nomischen Mustern zusammen. Allerdings lassen sich sicher nicht alle Aspekte dieser Vertiefungen quantitativ erfassen. Wir müssen solche Vorstellungen von *vertieften Erklärungen* daher auch informell als ergänzende Idee zur Verbesserung von Erklärungen mit hinzunehmen. Die Konzeption der vertiefenden Erklärung liefert gerade für die speziellen Wissenschaften eine sehr plausible und intuitive Vorstellung von Erklärungskraft und einer weitergehenden Vernetzung und damit Erklärungskohärenz unseres Überzeugungssystems.

Insgesamt wurde hier zunächst die (metaphysische) Kernidee dafür genannt, was eine Erklärung ausmacht. Sie nennt nomische Muster, die typischerweise kausale Dispositionseigenschaften beschreiben, die tatsächlich wesentlich daran beteiligt waren, ein Explanandum Ereignis E herbeizuführen. Dazu benennt sie die aktuellen Randbedingungen, die zugegen waren, damit der Effekt E so eintreten konnte. Dazu wurden die wichtigsten Parameter für die Beurteilung von Erklärungsstärken angegeben. Sie beziehen sich auf den empirischen Gehalt der erklärenden Theorie, die uns mehr oder weniger über das Explanandum mitteilt und dadurch, dass sie mehr ausschließt, auch deutlicher macht, warum wir gerade E und nicht etwas anderes erwarten sollten. An diesen Stellen findet sich auch ein impliziter Hinweis auf den konstrastiven Charakter von Erklärungen. Außerdem zeigt sich die Erklärungsstärke auch darin, wie gut die Theorie es schafft, möglichst viele Phänomene einzubeziehen und zu erklären. Eine stark vereinheitlichende Theorie beschreibt einen für unsere Welt zentraleren und tiefergehenden Mechanismus als eine weniger vereinheitlichende Theorie. Daher betrachten wir die Vereinheitlichung als einen wesentlichen Parameter der Erklärungsstärke.

Allerdings sind unterschiedliche Aspekte der Vereinheitlichung zu unterscheiden und stehen sogar in einem Spannungsverhältnis zueinander. Außerdem besteht ebenfalls ein Spannungsverhältnis zu dem Gehalt der

Theorie. Wie gut eine Theorie also tatsächlich in einem konkreten Fall ein Phänomen oder ein einzelnes Ereignis erklärt, ist nur durch einen schwierigen Abwägungsprozess für den Einzelfall anhand der genannten Parameter durchzuführen. Der Vergleich von Erklärungsstärken ist sicher kein leichtes Geschäft, für das ganz einfache Regeln angebbar wären, die immer zu einer definitiven Einschätzung führen würden. Es verbleibt ein gewisser Beurteilungsspielraum. Das ist natürlich trotzdem weit entfernt vom »anything goes«-Geschrei mancher Relativisten, jedoch scheint hier Kuhn Recht zu behalten, dass die epistemische Beurteilung von Theorien nicht immer zu einem definitiven Ergebnis kommen muss. Sie können in dem Sinne methodologisch inkommensurabel sein, dass die eine Theorie eine bessere Vereinheitlichung bietet, während die andere eine höhere Wahrscheinlichkeit für das Auftreten des Explanandumereignisses E anbietet und damit die Konkurrenten besser zurückweist. Für rein probabilistische Ansätze wie den Bayesianismus wird das nicht so deutlich, weil sie z.B. den Aspekt der Erklärungsstärke nicht berücksichtigen und einfach davon ausgehen, dass nur die höhere Wahrscheinlichkeit der Theorie den Ausschlag bei der Theorienwahl gibt. Doch das wird der Praxis der Theorienwahl in der Wissenschaft nicht gerecht, worauf wir im Zusammenhang mit einer Diskussion des Bayesianismus im nächsten Kapitel zurückkommen werden.

4.3.5 Kausale Mechanismen

Wenn wir komplexere Erklärungen untersuchen, wird das Problem natürlich noch deutlicher, dass es keine einfache Verrechnung von Erklärungsqualitäten gibt. Oft sprechen wir davon (etwa in der Medizin, der Biochemie oder den Sozialwissenschaften vgl. Bartelborth 2007), dass wir etwas anhand eines *kausalen Mechanismus* erklären. So erklären wir etwa, warum im Normalfall ein größeres Angebot eines bestimmten Konsumgutes zu einem niedrigeren Preis dieses Gutes führt, indem wir den genaueren Mechanismus beschreiben, der zu diesem Effekt führt. Hier können wir also sogar nomische Muster selbst weitergehend erklären, bzw. eine tiefergehende Erklärung der Phänomene liefern, die sie beschreiben.

Erklärung durch einen kausalen Mechanismus
Dabei wird eine bestimmte Eigenschaft E oder ein bestimmtes Verhalten E eines Systems S (hier bestehend aus einer Gruppe von Konsumenten und Produzenten) aus dem Verhalten bestimmter Teile des Systems S anhand kausaler Gesetze (oder zumindest nomischer Muster) abgeleitet.

In unserem Fall sind es die Dispositionen der Konsumenten, möglichst wenig bezahlen zu wollen und die der Produzenten trotz eines Überangebots, ihre Waren verkaufen zu wollen, und dazu zur Not mit dem Preis herunterzugehen, um dadurch ihre Konkurrenz auszustechen, die hier zu dem genannten Zusammenhang von Angebot und Nachfrage führen. In Bartelborth (2007) habe ich dafür plädiert, dass wir für wissenschaftliche Erklärungen wiederum darauf angewiesen sind, das Verhalten und Zusammenspiel der Teile des Systems durch nomische Muster zu beschreiben. Allerdings kann es dabei natürlich Lücken geben und nicht alle mechanistischen Erklärungen werden immer gleich vollständig sein.

Jedenfalls setzt sich eine solche mechanistische Erklärung dann wieder aus vielen kleineren Ableitungen durch nomische Muster zusammen und eine Gesamtbewertung muss die dann alle gemäß den oben genannten Kriterien berücksichtigen und etwa mit entsprechenden Konkurrenzerklärungen vergleichen. Hier wird man typischerweise bestimmte Ebenen unterscheiden. So sind meistens die Mikroerklärungen gegenüber den Makroerklärungen tieferliegend und dadurch auch oft die besseren Erklärungen eventuell sogar mit der größeren Vereinheitlichungskraft. Außerhalb der Naturwissenschaften – vor allem in den Sozialwissenschaften – sind wir z.T. sogar angewiesen auf bestimmte Mikrofundierungen (s. Bartelborth 2007, Kap. 5) unserer Erklärungen. Es dürfte offensichtlich sein, dass ein solches komplexes Zusammenspiel von einzelnen Erklärungsschritten noch einmal schwieriger zu bewerten ist als die einfachen Erklärungen, die wir im vorigen Abschnitt beschrieben haben. Aber auch mit Hilfe der Mechanismenerklärungen werden wir wieder Faktenabduktionen vornehmen (wie etwa die auf Atome oder andere kleine Teilchen) und ebenso auf bestimmte Kausalgesetze schließen.

4.3.6 Offene Fragen und andere Erklärungsformen

Die Debatte um die richtige Erklärungstheorie ist so umfassend, dass ich hier nur Teile davon aufgreifen kann und darüber hinaus auf Bartelborth (2007,2008) verweisen muss. Zumindest möchte ich aber noch eine Reihe offener Fragen kurz erwähnen. Bisher habe ich mich primär mit Ereigniserklärungen beschäftigt.

In der Wissenschaft werden aber auch bestimmte Gesetze aus noch grundlegenderen Gesetzen abgeleitet und so erklärt. Dabei geht es nicht darum – wie in den Ereigniserklärungen –, für das Explanans eine geeignete Ursache anzugeben, sondern es soll eher der kausale Mechanismus beschreiben werden, der zu den Gesetzen der höheren Ebene führt. Ein Beispiel für eine Gesetzeserklärung könnte die Schilderung sein, wie es zum Gesetz von Angebot und Nachfrage kommt, wonach z.B. eine reduzierte Nachfrage zu niedrigeren Preisen führt. Dabei erklären wir den zugrundeliegenden Mechanismus, der die Zusammenhänge auf der Ebene der beteiligten Personen beschreibt. Wenn einige Anbieter auf ihren Waren sitzenbleiben, erhoffen sie sich von einer Preissenkung doch noch einen Käufer zu finden und dann müssen alle anderen Anbieter mitziehen, weil sonst die rationalen Kunden zu den billigeren Anbietern abwandern würden.

An dem Beispiel erkennen wir zugleich ein weiteres Problem: Es handelt sich um eine *Gründe-Erklärung*, bei der eine Handlungsweise dadurch erklärt wird, dass wir bei den beteiligten Personen bestimmte Gründe für ihre Handlungen annehmen und davon ausgehen, dass sie als rationale Personen entsprechend ihren Gründen handeln. Gründe sind aber zunächst einmal keine Ursachen und passen so nicht gut in das obige Erklärungsschema. Tatsächlich bin ich der Meinung, dass Gründe nur dann erklären, wenn sie auch die Gründe sind derentwegen wir tatsächlich gehandelt haben. Dann sind sie zugleich Ursachen und wir können sie letztlich unter das neue Schema subsumieren. Doch es kann trefflich darüber gestritten werden, ob wir in den Sozialwissenschaften nicht auch noch andere Erklärungsschemata benötigen.

Auch in der Biologie stoßen wir auf möglicherweise andere Erklärungsformen wie die *funktionalen Erklärungen*. So erklären wir etwa, warum Menschen ein Herz haben, damit, dass das die Funktion hat,

das Blut im Körper zirkulieren zu lassen und dass ohne ein Herz Menschen nicht überleben könnten. Doch hier wird eine Wirkung des Herzens zur Erklärung seiner Existenz herangezogen. Das scheint in der falschen Richtung zu laufen gegenüber einer Kausalerklärung. Um solche teleologischen Erklärungen akzeptieren zu können, müssen wir sie zunächst in eine kausale Form übersetzen. Dazu gibt es unterschiedliche Vorschläge. Einer funktioniert in etwa so: (1) Das Herz liegt im Menschen vor. (2) Es hat eine bestimmte Wirkung W und (3) Das Haben dieser Wirkung W ist die Ursache dafür, dass es aufgetreten ist oder dass es weiterbesteht. Diese drei Bedingungen müssen demnach erfüllt sein, damit wir sagen können, W sei die Funktion des Herzens und das Ganze sei eine Erklärung, wieso wir ein Herz haben.

Die problematische Bedingung ist dabei die Bedingung (3). Wie kann das Vorliegen von (2) dazu führen, dass ein Organ (oder in den Sozialwissenschaften eine Norm oder eine Institution) auftritt? In der Biologie liefert uns die *Evolutionstheorie* den entscheidenden Mechanismus dazu. Es treten durch Mutation bestimmte Variationen auf, wobei dann durch die natürliche Auslese ganz bestimmte davon (etwa die mit Herz) ausgewählt werden. Dazu wird etwa bei Esfeld (2011, Kap. VII.1 bes. 115) wird genauer erläutert, wieso diese Erklärungen ebenfalls als eine Form dispositionaler Kausalerklärungen betrachtet werden können.

Außerdem stellen wir fest, dass nicht alle Ursachen von E das Ereignis E erklären können. Erklärungen sind typischerweise Antworten auf ganz bestimmte Warum-Fragen des Typs: Warum trat E auf? Der Big-Bang gehört zu den Ursachen aller Ereignisse in unserer Welt, doch die Antwort, weil E vom Big-Bang verursacht wurde, stellt für die meisten Warum-Fragen keine brauchbare Antwort dar. Oder: Warum hat Fritz einen Nagel im Kopf? Weil der Nagel sich einen winzigen Sekundenbruchteil zuvor in seinem Kopf befunden hat und kein Objekt sich so schnell bewegen kann, dass der Nagel dann nicht mehr in seinem Kopf gewesen wäre.

Das Problem ist dabei wohl, dass wir in den meisten Fällen nach ganz bestimmten Ursachen fragen, die sich etwa auf einen speziellen Unterschied abzielen. Einige Erklärungstheoretiker nehmen deshalb an, dass die Warum-Fragen genau genommen *konstrastive* Fragen sind. Wir

fragen: Warum ist der Nagel im Kopf von Fritz, *statt* das Fritz' Kopf ganz unversehrt ist? Dann sind nur noch die Ursachen erklärend, die kausal für den genannten Unterschied verantwortlich sind bzw. uns darüber aufklären können. Etwa dass Fritz sehr leichtsinnig mit der Nagelpistole hantiert hat, könnte ein solcher Unterschied sein.

Einige Philosophen sind außerdem auf die Idee gekommen, dass wir beim Erklären vielleicht ganz auf die Gesetze oder nomischen Muster verzichten können. Womöglich genügt es fürs Erklären von E, einfach nur eine *Ursache* von E zu benennen. Im Alltag verfahren wir manchmal so: Warum ist die Scheibe zersprungen? Weil sie von einem Stein getroffen wurde. Doch bereits in der frühen Debatte um das Hempel-Schema haben die Proponenten darauf hingewiesen, dass wir damit kein Verstehen hervorbringen, wenn uns die generellen Zusammenhänge nicht schon bekannt sind. Wenn wir nicht wüssten, dass Glas recht zerbrechlich ist, wenn wir darauf einen Druck ausüben und dass auftreffende Steine genau das bewirken, dann würde uns die obige Erklärung kaum zufriedenstellen. Dass seine Frau ihn heute gebeten hat, seine Mutter im Alltagsheim zu besuchen, mag eine Ursache für den Autounfall von Franz sein, aber solange wir nicht wissen, auf welchen Wegen sich die Bitte so ausgewirkt hat und welche (psychologischen) Mechanismen dabei am Werk waren, liegt noch keine Erklärung vor.

Manche Erklärungen scheinen allerdings überhaupt keine Kausalerklärungen zu sein. Beim relativistischen Zwillingsparadox können wir erklären, warum der eine Zwilling jünger geblieben ist. Das liegt an der speziellen Struktur der speziell-relativistischen Raumzeit. Aber kann man die als eine *Ursache* dafür ansehen? Elliot Sober argumentiert dafür, dass auch Gleichgewichtserklärungen keine Kausalerklärungen darstellen, weil sie uns nicht den genauen kausalen Weg zum Gleichgewicht schildern, sondern nur den Endzustand benennen. Natürlich sind in all diesen Fällen sehr wohl grundlegende kausale Strukturen am Werk. Wir müssen daher versuchen, diese Zusammenhänge und ihren Beitrag zu den Erklärungen deutlich herauszuarbeiten, in der Hoffnung, dass sie sich dann doch als Kausalerklärungen erweisen.

Auch in der Mathematik können wir etwas erklären. Zum Beispiel, warum eine stetige Funktion auf einem Intervall mit negativen und positiven Funktionswerten auch eine Nullstelle besitzen muss. Das ist

natürlich keine Kausalerklärung. Solche Erklärungen können schon eher durch das zweite neue Schema von Erklärung eingefangen werden. Doch es scheint mir auch zu viel verlangt, dass solche nicht-empirischen Erklärungen von einer Theorie der empirischen Erklärungen gleich mit behandelt werden müssen. Außerdem haben wir natürlich aufzuklären, was wir genauer unter Ursache-Wirkungs-Beziehungen zu verstehen haben. Das wird in Kapitel 7 stattfinden. Und wir sollten erläutern, wieso idealisierte Modelle in der Wissenschaft, die Personen als ideal rational oder Objekte als ausdehnungslose Massenpunkte beschreiben trotzdem gute Erklärungen sein können. Das hängt nach Strevens damit zusammen, dass wir hier von allen Details abstrahieren können und sogar sollten, die keine Unterschiedsmacher für das Auftreten von E sind. In all diesen Punkten bleibt sicher noch viel Raum für weitere Forschungen.

4.4 Erklärungskohärenz und probabilistische Kohärenz

Es sind zudem noch (weitere) holistische Aspekte des Schlusses auf die beste Erklärung zu berücksichtigen. Gerade beim Akzeptieren wissenschaftlicher Theorien geht es darum, dass unser gesamtes Überzeugungssystem hinterher plausibler ist als vorher. Das bedeutet insbesondere, dass die neuen Theorien, die wir akzeptieren, gut zu unserem weiteren Hintergrundwissen passen müssen (insbesondere zu unseren anderen Theorien über die Welt) und durch ihren Erklärungsbeitrag unser Überzeugungssystem insgesamt besser zu einem System zusammenfügen – also die *Gesamtkohärenz* des Systems befördern.

Damit ist gemeint, dass unsere Überzeugungen möglichst eng untereinander vernetzt sind und sich somit *gegenseitig stützen*. Sie stehen nicht isoliert nebeneinander, sondern bilden eine möglichst zusammenhängende Geschichte darüber, wie unsere Welt funktioniert. Dadurch ist jede Überzeugung innerhalb der Geschichte durch andere Überzeugungen begründbar, während das ganze System vor allem dadurch begründet ist, dass es weiteren Input in Form von Daten auf kohärente Weise einbauen kann. Das heißt entweder, dass es diesen Daten einen Platz

in der Gesamtgeschichte so zuweisen kann, dass sie dort im Prinzip erklärt werden können, oder dass es erklären kann, wieso diese Daten Messfehler bzw. Irrtümer einer anderen Art sind, die wir nicht in unser System einzubauen haben, sondern schlicht zurückweisen können.

Dabei gibt es positive inferentielle Beziehungen wie deduktive Beziehungen, positive probabilistische Zusammenhänge und vor allem Erklärungsbeziehungen, die kohärenzstiftend wirken, und negative Bezehungen wie deduktive und probabilistische Inkonsistenzen oder Inkohärenzen und vor allem Erklärungsanomalien, die die Kohärenz eines Überzeugungssystems vermindern. Wir müssen sie alle zumindest informell in unsere Bewertung einbeziehen, wenn wir die Gesamtkohärenz eines Überzeugungssystems bestimmen. Genaugenommen müssten wir auch noch bestimmte Metaüberzeugungen einbeziehen, etwa über die Zuverlässigkeit bestimmter Informationsquellen u.Ä., aber das lassen wir hier zunächst einmal beiseite. Wie können wir dann dabei vorgehen?

4.4.1 Kohärenz, Erklärungsanomalien und Wahrheit

Mit der Zeit werden wir unser Wissen über die Welt vervollständigen, untereinander vernetzen und vertiefen. Anomalien im Sinne der Kohärenz wären dann z.B. *Wunder*, also Ereignisse von einer Art, von der wir uns nicht erklären können, wie sie in unserem System einen Platz finden könnten, aber wir sehen auch keine Möglichkeit, sie etwa als fehlerhafte Wahrnehmungen zurückzuweisen.

Denken wir uns z.B., wir würden direkt daneben stehen, wie jemand im Schneidersitz über dem Boden schwebt. Dann würden wir zunächst zugeben müssen, dass wir das nicht mehr als zumindest im Prinzip erklärbar in unser Modell der Welt einbauen können, wenn uns jedenfalls kein Trick dabei auffällt. Wir können es im Normalfall jedoch ebenso wenig als Wahrnehmungsfehler abtun, wenn wir bei guter Sicht direkt daneben stehen. Allerdings sollten uns Zauberkunststücke vor Augen führen, dass wir manchmal bestimmte Ereignisse letztlich doch einbauen können, nur dass uns nicht so schnell einfällt, wie das geschehen kann. Gelingt es aber nicht, das Schweben als Trick zu entlarven, bleibt eine *Erklärungsanomalie* in unserem System zurück, die wir als echte

Inkohärenz deuten müssen. Solche Inkohärenzen sind typischerweise ein Ansporn, unser System umzubauen bzw. zu verbessern.

Als Niels Bohr seine Theorie des Atoms in Analogie zu einem Planetensystem entwarf, konnte er damit über einen längeren Zeitraum immer mehr Daten über die Abstrahlungsspektren von Atomen erklären. Allerdings gab es von Anfang an eine wesentliche Erklärungsanomalie. Er konnte nicht erklären, warum die um den Atomkern kreisenden Elektronen nicht in kürzester Zeit ihre gesamte Bewegungsenergie abstrahlten, wie sie es gemäß den Maxwellschen Gleichungen tun sollten, und dann in den Kern stürzen (vgl. Bartelborth 1989). Bohr vermied eine direkte Inkonsistenz seiner Theorie zur Elektrodynamik, indem er den *Anwendungsbereich* der Maxwellschen Elektrodynamik auf *nichtgebundene* Elektronen einschränkte.

Trotzdem blieb es natürlich eine *Inkohärenz* in seinem System, dass die grundlegenden Theorien für elektromagnetische Felder und elektrisch geladene Teilchen zwar für freie Elektronen galten, aber nicht auf die im Atom gebundenen Elektronen ausdehnbar waren, zumal die Beschränkung des Anwendungsbereichs nicht weiter erklärt werden konnte, sondern eher einen Ad-hoc-Charakter zu haben schien. Damit wurde die vereinheitlichende Kraft der Elektrodynamik stark eingeschränkt, aber noch schlimmer war, dass Objekte, die offenbar zu ein und derselben natürlichen Art gehörten (freie und gebundene Elektronen), unterschiedlich behandelt wurden. Das war eine deutliche Inkohärenz, die uns nach neuen Theorien suchen ließ und letztlich mit zur Entwicklung der Quantentheorien beigetragen hat.

Entsprechendes findet sich in der Geschichte der Erklärung der Planetenbewegung. Einige Umlaufbahnen konnten mit Hilfe der newtonschen Gravitationstheorie erklärt werden, andere aber nicht. Einige dieser Rätsel ließen sich durch die Entdeckung neuer Planeten und ihrer Einflüsse auf die Bewegung der anderen Planeten lösen, aber insbesondere für die Perihelbewegung des Merkurs war das nicht der Fall. Den Wissenschaftlern ließ diese Inkohärenz keine Ruhe, bis sie letztlich in der allgemeinen Relativitätstheorie eine neue Theorie fanden, mit deren Hilfe die Erklärungen wieder vereinheitlicht werden konnten. Möglichst alle Ereignisse erklären zu können, ist daher eine wichtige Triebfeder

gerade der wissenschaftlichen Arbeit – jedenfalls in den empirischen Wissenschaften.

Kohärenz und Wahrheit. Erst wenn es uns gelingt, ein kohärentes Modell eines ganzen Phänomenbereichs zu entwerfen, haben wir die Hoffnung, dass es sich nun um eine in weiten Teilen korrekte Darstellung dieses Gebiets handeln könnte. Susan Haak (1993) verglich die Suche nach Kohärenz mit dem Lösen eines Kreuzworträtsels. Ob wir in den einzelnen Zeilen und Spalten jeweils die richtigen Lösungen gefunden haben, erkennen wir vor allem daran, ob sich eine Gesamtlösung ergibt, die auch an den Schnittpunkten die erforderlichen Übereinstimmungen aufweist.

Man könnte das Entstehen wissenschaftlichen Wissens vielleicht noch besser mit einem ganz speziellen Puzzle vergleichen. Bestimmte kleine Teile des Puzzles (die Daten) werden uns durch unsere Beobachtungen geliefert. Andere größere Teile (die Theorien) müssen wir selbst entwickeln. Die größeren Teile dienen dazu, die kleineren Teile zu einem Gesamtbild zusammenzufügen. Ein gutes Passen der großen zu den kleinen Teilen wird repräsentiert gerade die guten Erklärungsbeziehungen. Ergibt sich ein verständliches Gesamtbild, so ist das für uns ein wichtiges Indiz, dass wir das richtige Bild entwickelt haben. Als Besonderheit haben wir noch das Phänomen, dass wir manchmal kleine Teile zurückweisen dürfen, weil sie einfach nicht in unser Gesamtbild passen.

Natürlich verbleiben in dieser Analogie viele Disanalogien. Die Beobachtungen werden nicht einfach frei Haus geliefert. Um sie in Worte zu fassen, müssen wir schon bestimmte *Begriffe* benutzen. Die können mehr oder weniger geeignet sein und entwickeln sich zusammen mit unseren Theorien. Außerdem können wir zwar kleine Teile als Irrtümer zurückweisen, aber nur unter ganz bestimmten Bedingungen. Wir benötigen Erklärungen dafür, warum wir sie für irreführend halten. Diese schönen Analogien werden außerdem einen Skeptiker nicht besänftigen, aber sie können für uns ein wenig erhellen, wie wir Wissen generieren und warum für uns die Kohärenz und speziell die Erklärungskohärenz dabei so bedeutsam ist. Auf die Zusammenhänge zur Wahrheit werden wir noch einmal zurückkommen.

4.4.2 Erste probabilistische Explikationen von Kohärenz

Verschiedene Autoren wie Keith Lehrer (1974), Laurence BonJour (1989) oder Paul Thagard (2000) haben unterschiedliche klassische Konzeptionen von Kohärenz (speziell Erklärungskohärenz) entwickelt und z.T. sogar in Programme gegossen (vgl. auch Schoch 2000), die sie anschließend erfolgreich auf viele Bereiche in einigen Fallstudien aus der Wissenschaft oder dem juristischen Bereich angewandt haben (s. vor allem Thagard 2000). Sogar die Bayesianer entwickeln inzwischen rein probabilistische Maße für Kohärenz (vgl. z.B. Fitelson 2003, Bovens & Hartmann 2006, Meijs 2005, Glass 2007, Schippers 2014). Einen guten Überblick über immerhin 18 derartige Kohärenzmaße und einen Test, welche davon 11 Beispiele von mehr oder weniger kohärenten Aussagenmengen am besten beschreiben können, bietet uns z.B. Jakob Koscholke (2015).

Insbesondere können wir auf naheliegende Weise sogar holistische Maße für Kohärenz bilden (vgl. Meijs 2005). Wir nehmen z.B. eines der zahlreichen bayesianischen Maße der Bestätigung wie etwa B(p,q) = P(p,q) – P(p) für zwei Aussagen p und q. Für eine Menge X von Aussagen können wir dann für je zwei Teilmengen Y und Z der Aussagen aus X den Zusammenhang zwischen den Konjunktionen dieser Aussagen betrachten: B(⋀Z, ⋀Y), wobei »⋀Z« die Konkunktion aller Aussagen aus Z bezeichnet. Mit B(⋀Z, ⋀Y) soll also ausgedrückt werden, in welchem Umfang die Aussagen der Menge Z durch die der Menge Y gestützt werden. Um das Ausmaß des Zusammenhangs der Aussagen in der ganzen Menge X bestimmen zu können, bilden wir nun z.B. den Durchschnitt über alle solchen Paare von Teilmengen von X. Für jedes einzelne bayesianische Bestätigungsmaß B erhalten wir damit ein neues holistisches Kohärenzmaß, in dem genauer festgehalten wird, wie stark die Aussagen in X sich gegenseitig stützen (s. dazu Kap. 4.4.6).

Ein Problem für die probabilistischen Kohärenzmaße ist dabei allerdings, dass die Erklärungsdebatte gezeigt hat, dass keine rein probabilistische Konzeption von Erklärung adäquat ist. Man denke z.B. an das obige Diktatorbeispiel. Die probabilistischen Kohärenzmaße sind daher zunächst keine Maße für die *Erklärungs*kohärenz, sondern es handelt sich eher um komplexe Maße der probabilistischen Theorienbestätigung.

Man könnte sie eventuell um weitere Anforderungen ergänzen, die etwa die Zusammenhänge besonders hoch gewichten, die auch Erklärungsbeziehungen ausdrücken.

Außerdem können probabilistische Kohärenzmaße zu internen Konflikten führen und mir ist bisher noch kein klarer Lösungsweg dafür bekannt. So kann es passieren, dass wir die Wahl zwischen zwei Hypothesen H1 und H2 haben, wobei H1 die größere Wahrscheinlichkeit aufweist: P(H1) > P(H2). Doch unser Überzeugungssystem wird vielleicht kohärenter, wenn wir H2 statt H1 akzeptieren oder wenn wir H2 zumindest eine höhere Wahrscheinlichkeit als H1 geben würden. Wem sollen wir dann folgen, den Kohärenzüberlegungen, die letztlich auf den Wahrscheinlichkeitseinschätzungen beruhen, oder einfach den ursprünglichen Wahrscheinlichkeitseinschätzungen selbst? Im ersten Fall müsste wir ein neues Update-Verfahren entwickeln, um die Wahrscheinlichkeiten den Kohärenzüberlegungen anzupassen oder die bayesianischen Verfahren zumindest ergänzen. Im zweiten Fall scheinen die Kohärenzüberlegungen eigentlich überflüssig zu sein und wir orientieren uns letztlich doch nur an den Wahrscheinlichkeiten.

Bartelborth (2005) betont dazu, dass wir *zwei* Ziele in der Wissenschaft verfolgen (vgl. Kap. 2.1) und deshalb der Bayesianismus zu ergänzen ist, wenn es uns um die Theorienwahl geht. Dafür zählt eben nicht nur die möglichst hohe Wahrscheinlichkeit einer Theorie (um im Sinne des ersten Ziels keine falschen Theorien zu akzeptieren), sondern auch der *Informationsgehalt* einer Theorie (im Sinne des zweiten Ziels) ist zu berücksichtigen. Diese Idee greifen z.T. auch Bayesianer wie Brössel (2014) auf und entwerfen für die Theorienwahl Maße von Informativität unserer Theorien, die zusätzlich in die Auswahl einer Theorie einzubeziehen sind. Brössel kann dazu beweisen, dass einige der probabilistischen Maße von Kohärenz (die ich gleich kurz vorstellen werde) gerade die von ihm untersuchten Konzeptionen vom Gehalt einer Theorie berücksichtigen und nicht mehr nur deren Wahrscheinlichkeit.

Allerdings wird für Brössel der Informationsgehalt einer Theorie durch zwei rein probabilistische Maße bestimmt, die m.E. nicht genau das repräsentieren, was im zweiten Ziel beabsichtigt war. Als Maße dienen Brössel die Werte $P(\neg T)$ oder $P(\neg T|\neg E)$, die beide etwas darüber sagen, dass der Gehalt einer Theorie in einem Spannungsverhältnis

zur Wahrscheinlichkeit der Theorie steht, aber noch wenig darüber aussagen, welche Art von Information wir uns in der Wissenschaft genau wünschen. Brössel koppelt die Information nicht an bestimmte Vorstellungen von *Wahrheitsnähe* oder an spezifische Vorstellungen darüber, dass unsere Theorien vor allem bestimmte Phänomene *erklären* sollten. Erklärungsstarke Theorien haben zwar prima facie eine geringere Wahrscheinlichkeit, aber nicht jede Theorie mit geringer Wahrscheinlichkeit liefert uns sogleich die Erklärungen nach denen wir suchen. Wir sollten daher noch weiter erforschen, nach welchen Informationen wir suchen, ehe wir uns ganz bestimmten Explikationen zuwenden können.

Glass (2007) sieht die Defizite der rein bayesianischen Ansätze in Bezug auf die Explikation von Erklärungen sehr deutlich, hofft aber trotzdem weiterhin, ein probabilistisches Maß zumindest für die *Erklärungsstärke* entwickeln zu können. Solche Aspekte probabilistischer Kohärenz, die etwa durch die Stärke der Korrelation zweier Aussagen bestimmt werden, spielen in normalen Ansätzen der Erklärungskohärenz zwar auch eine Rolle, aber möglicherweise nicht die wichtigste. Im Vordergrund stehen die *Erklärungsbeziehungen* zwischen unseren Überzeugungen, die nicht einfach auf Korrelationen reduzierbar sind. Statt rein probabilistischer Zusammenhänge sind hier vielmehr die kausalen Zusammenhänge spielentscheidend. Daher kommt es auch nicht nur auf die Wahrscheinlichkeitserhöhung an, die die Bayesianer vor allem im Blick haben, sondern ebenso auf den Gehalt der Theorien. Olsson (2005) beweist sogar, dass die probabilistischen Maße nicht mit einer Wahrscheinlichkeitserhöhung verträglich sind, weshalb in dieser Hinsicht Kohärenzansätze und rein bayesianische Überlegungen nicht immer in einfacher Weise zusammenpassen.

Um das probabilistische Zusammenpassen bzw. die *probabilistische Kohärenz* zu messen, sind u.a. die folgenden Maße für das Zusammenpassen von zwei Aussagen A und B (und P(A), P(B) > 0) vorgeschlagen worden:

$C_1(A,B) = P(A \,\&\, B) \,/\, P(A) \cdot P(B)$ (Shogenji 1999)
$C_2(A,B) = P(A \,\&\, B) \,/\, P(A \vee B)$ (Glass 2002 und Olsson 2002)

Dabei fragen wir uns, wie wahrscheinlich es ist, dass beide Aussagen zusammen wahr sind und vergleichen das mit Situationen, in denen sie

unabhängig voneinander wären. Außerdem ist $C_1(A,B) = P(A|B)/P(A) = P(B|A)/P(B)$. $C_1(A,B)$ gibt also an, wie stark A und B sich gegenseitig stützen im Sinne des Ratio-Maßes, und das ist eine naheliegende Idee, unsere Vorstellungen von Kohärenz zu explizieren. Das Shogenji-Maß ist trotz einiger problematischer Eigenschaften sicher weiterhin ein Musterbeispiel für ein (einfaches) probabilistisches Kohärenzmaß. Das Glass-Olsson-Maß bestimmt das Ausmaß, in dem sich die beiden Aussagen überlappen und wird daher auch als ein Überlappungsmaß bezeichnet. Beide Maße sind natürlich leicht auf Mengen von Aussagen $X = \{A_1,\ldots,A_n\}$ erweiterbar (s. Kap. 4.4.6).

Fitelson (2003) stützt sich auf eine ältere Konzeption, um die *Unterstützungsfunktion* F einer Aussage B für eine Aussage A zu bestimmen und nimmt dann den Durchschnitt über all diese Stützungsbeziehungen in unserem Überzeugungssystem als Maß für seine Kohärenz. Dabei ist die Funktion F für den Fall $P(A)<1$ und $P(B)>0$ in intuitiver Weise wie folgt definiert:

$$F(A,B) = \frac{P(A|B) - P(A|\neg B)}{P(A|B) + P(A|\neg B}$$

Der Zähler gibt eine Form von Bestätigung von A durch B dar, während der Nenner dabei nur eine Form von Normierung vornimmt. Die Idee, dann Durchschnitte von solchen Stützungsmaßen zu nehmen, hat dann zu einem neuen Ansatz bei Douven und Meijs (2007) geführt. Eine Debatte dieser ersten Maße findet sich etwa schon in Bovens & Hartmann (2006). (Zur Kritik am Fitelson-Maß s.a. Siebel 2004). Alle diese Maße setzen auf unterschiedliche Art eine Idee davon um, wann zwei Aussagen (oder auch mehrere) aus der Sicht einer Wahrscheinlichkeitsverteilung P mehr oder weniger gut zusammenpassen bzw. eine gewisse Korrelation aufweisen. Das verlangt allerdings, dass wir bereit sind, allen Aussagen eine Wahrscheinlichkeit zuzuweisen, was speziell die Probabilisten kennzeichnet. Der klassische Statistiker wird uns da schon nicht mehr folgen, aber er kann zumindest noch die Likelihoods akzeptieren, auf die ich mich in meiner Konzeption von Erklärungskohärenz stützen werde.

4.4.3 Die klassische Konzeption der Erklärungskohärenz

Die Grundideen der eher traditionellen (nicht-probabilistischen) Ansätze zur Erklärungskohärenz sind inzwischen ebenfalls klar erkennbar (vgl. dazu BonJour 1985, Bartelborth 1996, Thagard 2000). Danach gibt es *positive Beziehungen* zwischen Aussagen, die die Kohärenz befördern, wie deduktive Beziehungen, Erklärungsbeziehungen und auch probabilistische Stützungsbeziehungen, bei denen eine Aussage andere wahrscheinlicher erscheinen lässt. Nur für diesen letzten Aspekt könnte man gegebenenfalls an das Maß von Fitelson (s.o.) oder ähnliche Ansätze denken (vgl. Bovens & Hartmann 2006). Außerdem gibt es *negative Verhältnisse* zwischen Aussagen, die die Kohärenz mindern bzw. zu Inkohärenzen führen. Das sind vor allem logische Inkonsistenzen und außerdem probabilistische Inkonsistenzen (das sind Fälle, in denen unsere Theorien etwa bestimmten Ereignissen unterschiedliche Wahrscheinlichkeiten zuordnen) und vor allem Erklärungsanomalien sowie isolierte Subsysteme. Daneben können bestimmte Aussagen sich natürlich ebenso neutral zueinander verhalten.

Die Kohärenzkonzeption harmoniert so schließlich mit unserer Vorstellung von (wissenschaftlichem) Wissen. Wissen ist schließlich ebenfalls ein relativ holistisches Konzept. Nehmen wir etwa an, dass all unsere akzeptierten Aussagen (also alle Überzeugungen unseres Überzeugungssystems) Wissen darstellen. Dann sind all diese Überzeugungen gut begründet und die Begründungen selbst sind wiederum Wissen (also ebenfalls gut begründet) und weisen zudem keine relevanten Unterminierer auf und haben keine starken Gegengründe gegen sich, die natürlich Inkohärenzen darstellen würden. Ein solches Überzeugungssystem (das natürlich nur eine idealisierte Zielvorstellung darstellt) weist somit eine relativ hohe Kohärenz auf, jedenfalls dann, wenn wir auch noch unsere zweite Forderung nach informativen Überzeugungen mit Vorhersage- und Erklärungskraft umsetzen. Sonst bestünde wiederum die Gefahr, dass wir uns mit trivialem Wissen zufriedengeben könnten.

Ein sehr kohärentes Überzeugungssystem bietet uns andererseits erste Hinweise darauf, dass es sich bei den Überzeugungen um Wissen handeln dürfte. Die Kohärenzkonzeption wird allerdings üblicherweise (und auch hier) als internalistisches Begründungsverfahren angesehen,

so dass die externalistischen Forderungen im Wissensbegriff nach Wahrheit oder der Abwesenheit unbekannter Unterminierer nicht direkt erfüllt werden, aber die guten Begründungen in einem kohärenten Überzeugungssystem sind für uns zumindest die besten Hinweise darauf, dass die externalistischen Forderungen tatsächlich erfüllt sind. Für abduktive Schlüsse sind typische Unterminierer Konkurrenztheorien zu unseren Theorien mit ähnlich guter Erklärungsleistung oder Daten, die nicht zu unseren Theorien passen, obwohl sie eigentlich in deren intendierten Anwendungsbereich fallen sollten.

Frühe Ansätze zur Kohärenz hielten Aussagensysteme für besonders kohärent, in denen jede Aussage durch die anderen deduktiv abgeleitet werden kann, doch das kann kaum gemeint sein (vgl. Bartelborth 1996b). Diese Systeme sind eher *redundant* als kohärent. Nehmen wir etwa eine Menge $X = \{A_1,...,A_n\}$ von beliebig inkohärenten, aber konsistenten Aussagen. Die soll plötzlich kohärent werden, wenn ich sie um die eine Aussage $A_1\&...\&A_n$ zur Menge $X^* = \{A_1,..., A_n, A_1\& ... \&A_n\}$ ergänze, obwohl eigentlich nichts substantiell Neues hinzugekommen ist. Das so beschriebene Stern-Verfahren zeigt, wie leicht wir Mengen »deduktiv-kohärent« gestalten können und wie wenig wir uns auf deduktive Beziehungen verlassen dürfen, wenn es uns eigentlich um *Erklärungskohärenz* geht.

Ganz entscheidend für die Kohärenz sind statt der deduktiven Zusammenhänge vielmehr *Erklärungsbeziehungen*, die in unserer Menge X und auch in der ergänzten Menge X* überhaupt nicht vorkommen müssen. Das Stern-Verfahren ist ebenfalls ein Problem für einige probabilistische Kohärenzmaße, die auf die gegenseitige Ableitbarkeit als Ideal setzen. Selbst das recht interessante Maß von Fitelson (2003) ist so gestaltet, dass für eine Menge X von unzusammenhängenden Aussagen der Übergang zu X* die Kohärenz i.A. erheblich erhöht, obwohl keine substantiellen neuen Einsichten hinzugekommen sind. Das liegt einfach daran, dass dann für alle Aussagen A_i $P(A_i|A_1\& ... \&A_n)=1$ ist und daher durch Hinzunahme von $A_1\& ... \&A_n$ entsprechende positive Werte hinzukommen, die dann den Durchschnitt im Normalfall anheben. Man beachte hier, dass das Maß für inkohärente Aussagen sonst im Bereich zwischen 0 und -1 liegt.

Die wichtigsten Träger der Kohärenz sind daher nicht die *deduktiven Zusammenhänge*, sondern die *Erklärungsbeziehungen*, wobei die Stärke

der Erklärung jeweils Auskunft über das Maß an Kohärenz gibt. Außerdem sind dann unsere erklärenden *Theorien* entscheidend für die Frage nach Kohärenz und Inkohärenz. Zunächst stehen etwa viele Beobachtungsaussagen zusammenhanglos nebeneinander. Trotzdem erzeugen die folgenden Aussagen desselben Tages mit Zeitindex keineswegs von sich aus eine Inkohärenz:

11.10 Uhr: Ich sehe vor mir das Brandenburger Tor.
11.11 Uhr: Ich sehe vor mir den Eiffelturm in Paris.
11.12 Uhr: Ich sehe vor mir die Freiheitsstatue in New York.

Rein logisch gesehen können wir durchaus derartige Beobachtungen machen. Sie schließen sich nicht in irgendeiner Weise logisch aus, sondern nur verkehrstechnisch (vgl. Bartelborth 1996). Sie wirken also nicht inkohärent untereinander, ohne weiteres Hintergrundwissen ins Spiel zu bringen. Wir wissen, dass es uns bei den heutigen Verkehrsmitteln nicht gelingen kann, so schnell hintereinander den Ort zu wechseln. Es ist erst unser *Hintergrundwissen* und hier speziell unsere Theorien über das Funktionieren unserer Welt, die uns die Inkohärenzen aufzeigen. Ohne unsere Theorien könnten wir noch nicht einmal sagen, dass eine der Aussagen die anderen *wahrscheinlicher* oder *unwahrscheinlicher* macht. Legen wir die zeitlichen Abstände noch deutlich kürzer, könnten wir sogar schon eine physikalische Unmöglichkeit für entsprechende Wahrnehmungen proklamieren, aber wiederum nur, wenn wir die entsprechenden physikalischen Theorien annehmen. Jedenfalls sind wir auf Theorien angewiesen, um überhaupt Zusammenhänge zwischen den Daten zu generieren.

Das gilt auch für die positiven Zusammenhänge. Dass bestimmte Abläufe eine zusammenhängende und damit kohärente Geschichte ergeben, wird erst durch unsere *Theorien* bestimmt. Wir haben Theorien oder Modelle dafür, wie sich Objekte normalerweise bewegen und wie sich Menschen typischerweise verhalten. Hierzu verfügen wir zunächst über Alltagstheorien und außerdem natürlich über wissenschaftliche Theorien. Diese spezielle Rolle von Theorien wird von Bayesianern leider nicht entsprechend geschätzt. Der Bayesianismus erweist sich sogar eher als relativ theorienfeindlich, denn gerade erklärungsstarke Theorien stellen inhaltlich anspruchsvolle Behauptungen über die Welt auf, die

dazu führen, dass wir den Theorien meist kleinere Wahrscheinlichkeiten zuweisen müssen (vgl. Kap. 5.5.19).

Das war schon Poppers Punkt, dass es eine Spannung zwischen seiner Forderung nach möglichst riskanten Theorien und den empiristischen Ansätzen gibt, die üblicherweise verlangten, dass die Theorien besonders wahrscheinlich sein sollten, bevor wir sie akzeptieren dürfen. Darauf werden wir wieder zurückkommen. Neben den Theorien mit ihrer Erklärungskraft zählen für die Erklärungskohärenz aber natürlich auch deduktive und probabilistische Beziehungen.

Auf der negativen Seite stehen zunächst die *logischen Inkonsistenzen*, die die schwerwiegendste Inkohärenz darstellen, die ein Überzeugungssystem aufweisen kann. Manche Kohärenztheoretiker möchten solche Systeme sogar ganz als vollkommen inkohärent verbieten, doch lokale Inkonsistenzen an bestimmten Stellen (auch zwischen bestimmten Theorien) müssen deshalb noch nicht die Begründungen in anderen Bereichen unterminieren.

An zweiter Stelle finden wir die *probabilistischen Inkonsistenzen*. Wir halten z.B. zwei Theorien T und T* in unserem Überzeugungssystem X für probabilistisch inkonsistent, die für ein Ereignis A unterschiedliche Wahrscheinlichkeiten vergeben: $P(A|T) \neq P(A|T^*)$. Je größer die Differenz zwischen den beiden Wahrscheinlichkeiten ist, umso inkohärenter sind T und T* zueinander. Ein typisches Maß, das wir für diesen Abstand von T und T* heranziehen können, ist der sogenannte Kullback-Leibler-Abstand (s. a. Kap. 5.6.2), der sich nur auf die Likelihoods stützen muss. Weiterhin stören vor allem die *Erklärungsanomalien* die Kohärenz unseres Überzeugungssystems X. Erklärungsanomalien sind Ereignisse oder Phänomene, die eigentlich zum regulären Anwendungsbereich einer Theorie T aus X gehören, aber sich jedem Erklärungsversuch mit Hilfe von T widersetzen.

Schließlich gibt es auch noch *holistischere Phänomene*, die zu Inkohärenzen führen können, auf die uns ebenfalls schon Laurence BonJour (1985) aufmerksam gemacht hat. Nehmen wir dazu ein Überzeugungssystem X, das in zwei Bereiche Y und Z zerfällt, die zwar in sich jeweils gut vernetzt sind, aber zwischen denen es kaum positive Verbindungen gibt. Dann würden wir im Normalfall ebenfalls sagen, dass hier weniger

Kohärenz im Gesamtsystem vorliegt, als in einem gleichmäßiger vernetzten Überzeugungssystem mit derselben durchschnittlichen Vernetzung.

So ein Überzeugungssystem mit einem *isolierten Subsystem* könnte etwa aus einem wissenschaftlichen Aussagenteil und einem astrologischen Teil bestehen. Selbst wenn der astrologische Anteil so gestaltet wäre, dass er dem wissenschaftlichen Teil nicht direkt widerspricht, wäre es schon seltsam, über zwei so verschiedene Weltbilder zu verfügen, dass es kaum Vernetzungen zwischen ihnen gäbe. Das Gesamtbild hätte dadurch eine Inkohärenz aufzuweisen, die auch intuitiv erkennbar wäre. Im Prinzip hätten wir es dann mit zwei völlig unterschiedlichen Darstellungen ein und derselben Welt zu tun, die untereinander so inkommensurabel wären, dass sie sich noch nicht einmal explizit widersprechen. Dieser Extremfall scheint kaum vorstellbar zu sein, aber zeigt doch, wieso wir schon mit relativ isolierten Subsystemen als Kohärenztheoretiker nicht wirklich zufrieden sein können.

Eine einfache klassische Explikation von Kohärenz. Die Kohärenz eines endlichen Überzeugungssystems $X = \{A_1,...,A_n\}$ berechnet sich dann insgesamt, indem wir zunächst alle positiven Zusammenhänge quantifizieren, wobei wir die *relativen* Erklärungsstärken gegeneinander abwägen müssen. Haben wir keine Anhaltspunkte dafür, dass bestimmte Erklärungen besser sind als andere, können wir einfach alle Erklärungen mit +1 bewerten. Dann summieren wir diese Werte und müssen die negativen Werte von unserer Summe abziehen. Die negativen Verbindungen wie etwa die logischen Inkonsistenzen wiegen besonders schwer. Hier müssen wir eine intuitive Kalibrierung vornehmen, so dass die Inkonsistenzen z.B. mit -4 zu Buche schlagen. Solche Kalibrierungen sind anhand von Fallstudien möglich, in denen wir uns relativ sicher darüber sind, was als Ergebnis herauskommen sollte.

Ähnliche Einschätzungen muss der Bayesianer ebenfalls vornehmen, wenn er etwa Vorher-Wahrscheinlichkeiten für ganze Theorien vergeben muss, und schließlich sind seine subjektiven Wahrscheinlichkeiten bzw. Glaubensgrade auch nur theoretische Größen (s.u.). Der Bayesianer muss sogar eine komplexere holistische Abwägung gegenüber der eher lokalen Einschätzung der Kohärenztheoretiker vornehmen, denn der letztere muss nur bestimmen, wie gut jeweils ganz bestimmte Erklärungen sind und nicht gleich komplette Theorien bewerten.

Statt der Gesamtsumme als Maß für die Gesamtkohärenz sollten wir allerdings lieber den Durchschnittswert berechnen und Überzeugungssysteme anhand ihres *durchschnittlichen* Zusammenhangs bewerten, weil es uns nicht darum gehen sollte, möglichst viele Daten anzusammeln, die nur schwach eingebunden sind, sondern eher um ein aussagekräftiges bzw. informationsreiches Modell der Welt geht.

Wenn es uns gelingt, die Kohärenzbeziehungen auf *paarweise Beziehungen* zu reduzieren, wie es etwa Thagard (2000) annimmt, dann können wir die positiven wie negativen Verbindungen zwischen zwei Aussagen A_i und A_k jeweils durch einen Faktor e_{ik} (etwa als reelle Zahl) beschreiben und erhalten als Kohärenzmaß für $X = \{A_1,...,A_n\}$ in etwa ein Maß wie das folgende:

Ein einfaches Kohärenzmaß: $\mathrm{Koh}(X) = 1/n^2 \sum_{ik} e_{ik}$

Das wäre allerdings ein komplexes Maß mit gewissen intuitiven Einschätzungen, die wir vornehmen müssten, um die e_{ik} festzulegen. Für eine System von 100 Überzeugungen wären schon 10 000 Werte zu bestimmen. Die Isolation von Subsystemen ist dabei noch nicht einmal erfasst, da sie unseren Maßstab deutlich verkomplizieren würde. Hier sind also wesentlich Vereinfachungen erforderlich, damit wir diese Konzeption in der Praxis anwenden können.

Als Erstes können wir etwa annehmen, dass die gegenseitige Stützung oder Schwächung symmetrisch ist und daher gilt: $e_{ik} = e_{ki}$. Außerdem werden wir im Folgenden sehen, dass wir die e_{ik} in Anwendungsbeispielen intuitiv gut bestimmen können, dass es aber zumindest theoretisch auch einige Explikationsvorschläge dafür gibt, die zu genaueren Werten führen. Wichtiger als das ist aber zunächst ein anderer Vereinfachungsschritt, den wir als erstes einführen wollen.

Komparative Kohärenz. Aussagekraft hat das Kohärenzmaß vor allem als *Vergleichsmaßstab für konkurrierende Überzeugungssysteme*. Daher möchte ich das Vorgehen zunächst dadurch weiter vereinfachen, dass wir nur *zwei Szenarien* S und S* miteinander vergleichen, die jeweils Varianten eines Überzeugungssystems X darstellen. Dann erhalten wir zumindest noch eine interessante *komparative Konzeption von*

Begründung, die in vielen realen Situationen in der Wissenschaft hilfreich ist. Das ist so gedacht, dass sich die Szenarien vor allem in einer Theorie T plus einigen dazugehörigen Hilfsannahmen A unterscheiden und im Prinzip in X (in allen anderen Theorien und den Beobachtungsaussagen) übereinstimmen:

Vergleiche zwei Szenarien: $S = X \cup \{T, A\}$ und $S^* = X \cup \{T^*, A^*\}$

Wir vergleichen also nur noch, wie gut sich zwei konkurrierende Theorien (oder auch eine Theorie und die Abwesenheit dieser Theorie im anderen Szenario) in unser restliches Überzeugungssystem einpassen und inwieweit sie zu dessen Kohärenz beitragen. Dazu müssen wir nur die relativen Kohärenzwerte von S und S* bestimmen. Für den Vergleich sind aber viele der Werte e_{ik} in S und S* identisch und müssen somit nicht extra geschätzt werden. Längere Listen mit mehreren konkurrierenden Theorien lassen sich durch entsprechende Paarvergleiche schließlich auf die besten beiden Theorien einschränken, wenn wir davon ausgehen, dass unsere Theoriebewertungen untereinander transitiv sind, was wir natürlich anstreben müssen. Tatsächlich werden wir immer wieder sehen, dass alle Ansätze ähnlich wie die eliminative Induktion auf solche Vergleichslisten angewiesen sind und wir uns dann auf solche einfachen Kohärenzvergleiche beschränken können. Das reduziert die Anzahl der zu bestimmenden e_{ik} so stark, dass wir das Verfahren dann relativ leicht in der Praxis anwenden können.

Der Vorteil für die Bewertung ist also, dass wir nur noch zwei Arten von Verbindungen prüfen müssen, nämlich zum einen die Verbindungen unserer neuen Theorien T und T* zu den anderen Theorien aus X und zum anderen, inwieweit T und A bzw. T* und A* bestimmte Daten aus X erklären können. Dadurch müssen wir keine absoluten Gesamtkohärenzwerte für S und S* mehr ermitteln, sondern können uns ganz auf lokale Vergleiche beschränken, die die wichtigsten Auswirkungen von T und T* auf die Kohärenz unseres Überzeugungssystems bestimmen.

4.4.4 Anwendungen der klassischen Kohärenzkonzeption

Um in einfachen Beispielen eine erste Bewertung vornehmen zu können, in denen wir keine echten Wahrscheinlichkeiten angeben können,

können wir etwa für erfolgreiche Erklärungen einen Punkt vergeben, für einfache Erklärungsanomalien –1 für schwerwiegende –2 und für Inkonsistenzen z.B. –4 und die Beziehungen zu anderen Theorien zunächst außen vor lassen. Damit können wir unser Beispiel aus der Ursachenforschung der Cholera nun rekonstruieren. Dazu betrachten wir nicht die einzelnen erklärten Beobachtungen, sondern ganze Phänomene, die aber recht fein individuiert werden. Als Phänomen, das beide Theorien erklären können, wählen wir unter der Nummer 0 die Ausbreitung der Krankheit vor allem in Städten, in denen sowohl Luft- als auch Wasserqualität nicht sehr gut waren. Ansonsten wählen wir die Nummerierung, die wir oben gewählt hatten.

Phänomene	Miasma-Theorie	Infektionstheorie
0. Epidemie	1	1
1. Zeitintervall	1	1
2. Handelswege	-1	1
3. Kontakt (Hafen)	-1	1
4. Ursprung 1854	-1	1
5. Häufung Pumpe	-1	1
6. keine Erreger	0	-1
7. spezielle Brunnen	-1	1
8. Wasserqualität	-1	1
9. Wasserreinigung	-2	1
Summe	-6	8

Tab. 4.1: Erklärungskohärenz von Miasma- und Infektionstheorie

In unserem Beispiel gibt es einen klaren Sieger, wie wir das bereits intuitiv erkannt hatten. Da würde es auch nicht viel ändern, noch weitere Debatten über die genaue Erklärungsstärke für die einzelnen Erklärungen anzustrengen. Paul Thagard hat in einer Reihe von Aufsätzen etliche Beispiele aus der Wissenschaft und von Gerichtsfällen mit Hilfe seiner Variante der Kohärenztheorie rekonstruiert und anhand seines etwas komplexeren Programms ausgewertet. Schauen wir uns nun einen Fall mit einem etwas knapperen Ausgang an.

Greifen wir dazu noch einmal unser Beispiel der Magengeschwüre auf: Anfang der 1980er Jahre war die gängige Auffassung, dass Magengeschwüre durch Übersäuerung des Magens hervorgerufen werden, die

ihrerseits vor allem durch Stress entsteht. Eine weitaus ausführlichere Darstellung der Geschichte findet sich bei Thagard (1999, Kap. 3-6), die ich hier nur stark verkürzt wiedergeben möchte, um die Grundidee der Kohärenzbetrachtung zu erläutern. Im Jahre 1983 berichteten dann die australischen Ärzte Robin Warren und Barry Marshal, dass sie eine neue Bakterienart im Magen der Patienten mit Gastritis gefunden hatten, die erst deutlich später Helicobacter Pylori (HP) genannt wurde. Sie entwickelten daraufhin die Bakterienhypothese, wonach die Gastritis und daraus resultierende Magengeschwüre auf dieses Bakterium zurückzuführen seien. Diese Hypothese wurde u.a. deshalb von der damaligen Fachwelt als völlig lächerlich abgelehnt, weil man annahm, dass im sauren Milieu des Magens keine Bakterien überleben könnten. Man vermutete vielmehr, dass die Bakterienfunde höchstens als Verunreinigungen bei der Entnahme der Proben entstanden sein könnten. Aber bereits Anfang der 1990er Jahre hatte sich die Ansicht gewandelt. 1994 gab es sogar einen offiziellen Konsens, dass die HP-Infektion eine wichtige Rolle bei der Entstehung von Magengeschwüren spielt und daher Antibiotika für die Behandlung zu empfehlen seien. 2005 erhielten die beiden Ärzte überdies den Nobelpreis für Medizin.

Dazu können wir zunächst die wichtigsten Daten etwa um 1983 zusammenstellen. Damals besagte die Stresstheorie, dass die Magengeschwüre aus Übersäuerung wegen Stressbelastungen entstehen, während die Bakterienfunde auf Verunreinigungen zurückzuführen sind. Es sind dabei einige Hilfsannahmen erforderlich. Zum damaligen Hintergrundwissen gehört aber auch, dass (A) Bakterien im Magen nicht überleben können. Die Bakterientheorie hat damit die größten Schwierigkeiten. Dafür kann sie andere Befunde erklären. Auch bei der Beschreibung der Auswirkungen der Bakterien kann die erhöhte Magensäure eine wichtige Rolle spielen. Zu den Daten gehört etwa, dass (1) einige Personen Magengeschwüre entwickeln und (2) dass Antacida die Beschwerden lindern. Allerdings hat Warren (3) die Bakterien gerade bei den Patienten mit Gastritis und Magengeschwüren gefunden. Damit ergab sich um 1983 etwa das folgende Bild:

Phänomene	Stresstheorie	Bakterientheorie
(A) Magen zu sauer für Bakterien	0	-2
(1) Magengeschwüre	1	1
(2) Antacida heilen	1	0,5
(3) Bakterienfunde	0,5	1
Summe	2,5	0,5

Tab. 4.2: Erklärungskohärenz von Stress- und Bakterientheorie 1983

Es ist also durchaus nachvollziehbar, dass um 1983 die Stresstheorie noch vorne liegt. Dann änderte sich die Situation in verschiedenen Hinsichten. Es wurden immer wieder entsprechende Bakterien im Magen von Gastritis-Patienten gefunden, so dass die »Hilfserklärung«, es handele sich um bloße Verunreinigungen, schließlich nicht mehr aufrechtzuerhalten war. Damit wurde zugleich das Hintergrundwissen verändert. Es musste nun akzeptiert werden, dass es tatsächlich bestimmte Bakterien schaffen, im Magen zu überleben. Außerdem sind zwei neue wichtige Daten hinzugekommen. Zum einen (4) haben die Konstrukteure der Bakterienhypothese im Selbstversuch gezeigt, dass die Einnahme der entsprechenden Bakterien die Symptome der Gastritis auslöst. Außerdem hat sich erwiesen, dass (5) die Behandlung mit Antibiotika nachhaltigere Heilung mit sich bringt, als die mit Antacida. Damit erhalten wir dann eine neue Tabelle für die Situation 1995:

Phänomene	Stresstheorie	Bakterientheorie
(¬A) Magen enthält Bakterien	0	0
(1) Magengeschwüre	1	1
(2) Antacida heilen	1	0,5
(3) Bakterienfunde	0	1
(4) Bakterien lösen Gastritis aus	-1	1
(5) Antibiotika heilen	-2	1
Summe	-1	4,5

Tab. 4.3: Erklärungskohärenz von Stress- und Bakterientheorie 1995

Um 1995 hatte sich die Situation also tatsächlich grundlegend geändert, was die entsprechende Modifikation in der Bewertung der Bakterienhypothese erklärt. Auch wenn man die genauen Erklärungsstärken wieder diskutieren kann, sollte doch deutlich geworden sein, wieso

eine deutliche epistemische Änderung erfolgt ist. Unsere Abduktionen sind also eingebettet in ein größeres Überzeugungssystem und diese Einbettung entscheidet mit darüber, welches die bessere Erklärung ist.

4.4.5 Probabilistische Maße der Erklärungsstärke.

Es gibt inzwischen eine ganze Reihe von Ansätzen und konkreten Vorschlägen, wie ein genaueres probabilistisches Maß für Erklärungsstärke auszusehen hätte. Dazu möchte ich exemplarisch einige Ideen aufgreifen und dann für ein eher einfaches älteres Maß plädieren. Die Wahrscheinlichkeitsmaße P werden dabei immer als regulär angenommen, d.h. sie nehmen für empirische Aussagen nicht die Werte 0 oder 1 an, und wir gehen im Folgenden jedenfalls immer davon aus, dass die Nenner nicht 0 sind.

Schupbach und Sprenger (2011) haben anhand einer Liste von sinnvollen Anforderungen an eine Funktion $E_P(e,h)$ für die Stärke, mit der h das Explanandum-Ereignis e erklärt, einen konkreten Vorschlag für solch eine Funktion erarbeitet (vgl. mit dem Fitelson-Maß):

$$E_p(e,h) = \varepsilon(e,h) = \frac{P(h|e) - P(h|\neg e)}{P(h|e) + P(h|\neg e)}$$

Dabei spielte u.a. die Idee eine wichtige Rolle, dass die Erklärungsstärke umso größer ist, je mehr P(e|h) den Wert von P(e) überschreitet. Crupi und Tentori (2012) konnten jedoch zeigen, dass das Maß ε(e,h) auch eine ihrer Meinung nach unschöne Eigenschaften hat. Schupbach und Sprenger (2011) verlangen explizit für eine Hypothese h, die e erklärt, aber nichts zu einem irrelevanten Datum e* zu sagen hat, dass $E_P(e\&e^*,h) < E_P(e,h)$ sein sollte. Crupi und Tentori verlangen dann auch eine *explanatorische Gerechtigkeit*, wonach bei einem Fehlschlagen der Erklärung diese Ungleichung beibehalten wird. Auch fehlgeschlagene Erklärungen sollten nicht durch irrelevante Explanandum-Bestandteile verbessert werden können. Doch das trifft auf den Vorschlag von Schupbach und Sprenger zu. Um das auszuschließen, verlangen sie eine entsprechende Bedingung der explanatorischen Gerechtigkeit für unsere Maße von Erklärungsstärke.

Sie formulieren daher etwas veränderte Anforderungen an solche Maße. Eine wesentliche Forderung beider Ansätze ist z.B. die Forderung (E1), und die weiteren Forderungen des Ansatzes von Crupi und Tentori sind die folgenden:

Für alle Explananda e, Hypothesen h1 und h2 soll relativ zu P gelten:
(**E1**) *Positive Relevanz*: $E_P(e,h1) \gtreqless E_P(e,h2)$ gdw. $P(e|h1) \gtreqless P(e|h2)$
(**E0**) *Struktur*: $E_P(e,h)$ ist eine Funktion von P(e&h), P(e) und P(h)
(**E3**) *Symmetrie*: $\forall$ e, e*: $E_P(e,h) \gtreqless E_P(e^*,h)$ gdw. $E_P(\neg e,h) \lesseqgtr E_P(\neg e^*,h)$

Zusammen mit der Forderung nach explanatorischer Gerechtigkeit (E2) erhalten Crupi und Tentori dann ein Repräsentationstheorem für das folgende Maß:

$$\varepsilon^*(e,h) = \begin{cases} \dfrac{P(e|h) - P(e)}{1 - P(e)} & \text{falls } P(e|h) \geq P(e) \\ \dfrac{P(e|h) - P(e)}{P(e)} & \text{falls } P(e|h) < P(e) \end{cases}$$

Durch die Normierung liegt die Erklärungsstärke im Bereich [-1,1] und ist je nachdem positiv, wenn die Annahme, dass h der Fall ist, zu einer erhöhten Wahrscheinlichkeit von e führt und sonst negativ. Das ist sicher ein interessanter Vorschlag mit vielen guten Eigenschaften (die uns Crupi und Tentori vor Augen führen), den wir nun wählen könnten, um damit einige der e_{ik} in unserer Kohärenzkonzeption zu bestimmen. Er setzt allerdings einige starke Annahmen voraus. Wir müssen den Probabilismus annehmen, wonach wir für jede Aussage eine epistemische Wahrscheinlichkeit vergeben können (vgl. Kap. 5). Insbesondere erhält somit auch e selbst eine Wahrscheinlichkeit P(e). Im Rahmen dieses Probabilismus nimmt aber auch P(e|h) subjektive Werte an, die durch unsere Informationen über e beeinflusst sind und daher nicht mehr viel damit zu tun haben, was uns h über e sagt (vgl. Kap. 5.3.11).

Obwohl unsere Hypothese h vielleicht besagt, dass es sich um eine faire Münze handelt, kann dann die Wahrscheinlichkeit P(e|h) für eine Behauptung e = »es fällt Kopf« gerade 0,9 betragen, weil e selbst schon eine hohe epistemische Wahrscheinlichkeit aufweist. Deshalb müssen dann

auch die Differenzen P(e|h)–P(e) betrachtet werden. Diese epistemischen Wahrscheinlichkeiten werden aber nur von den Bayesianern akzeptiert und z.B. nicht von einem klassischen Statistiker. Daher möchte ich mich hier auf die *objektiven* Likelihoods P(e|h) beschränken, die nur zum Ausdruck bringen, inwieweit h nun e erwarten lässt und nicht, inwieweit e durch unser gesamtes Hintergrundwissen zu erwarten ist. Besagt h gerade, dass es sich um eine faire Münze handelt, wäre dann P(e|h) = 0,5 eindeutig bestimmt. Nur diese objektiven Likelihoods werden von anderen Ansätzen außerhalb des Bayesianismus gleichfalls akzeptiert. Diese Unterscheidung wird im Kapitel 5 noch genauer zur Sprache kommen (speziell in Kap. 5.3.11). Jedenfalls kommen wir zu einfacheren Maßen für die Erklärungsstärke, wenn wir uns auf diese objektiven Werte beschränken.

Nehmen wir darüber hinaus wieder an, dass wir nur zwei Szenarien S und S* miteinander vergleichen wollen, wobei S = X ∪ {T} und S* = X ∪ {T*} sein soll. Können dann beide Theorien T und T* gleich viele unserer Daten E = $\{E_1,...,E_m\}$ erklären, dann müssen wir nur noch vergleichen, wie stark jeweils die Erklärungen von E durch T und T* ausfallen. Dazu genügt es, die betreffenden (objektiven) Likelihoods miteinander zu vergleichen etwa in Form des Likelihoodquotienten P(E|T)/P(E|T*). Dazu geht man gerne noch zum Logarithmus über, weil dann die positiven Werte für T und die negativen Werte für T* sprechen, während die 0 gerade die neutrale Mitte darstellt. Das ist ähnlich wie bei dem normierten Maß von Crupi und Tentori.

Ein einfacher Vergleichsmaßstab für die Erklärungsstärke:
LQ(T,T*:E) = log [P(E|T) / P(E|T*)],

wobei wir der Einfachheit halber davon ausgehen, dass im Folgenden alle Werte P(E|T) und P(E|T*) echt größer Null sind, denn für die Fälle von verschwindenden Likelihoods müssten wir sonst extra Vorsorge treffen (wie das z.B. Hawthorne ausführlich in 2011a und 2011c unternimmt). Ist dann LQ(T,T*:E) > 0 so erklärt T die Daten besser, ist LQ(T,T*:E) < 0, so liegt T* um den entsprechenden Betrag vorn und bei LQ(T,T*:E) = 0 haben wir es mit einem Unentschieden zu tun, und wir können keine der beiden Theorien bevorzugen.

Das passt gut zu entsprechenden Beispielen: Es sei etwa T* = »Die Münze ist Fair« und T = »Kopf hat eine Wahrscheinlichkeit von 0,9«. Und unser Datum sei, dass 10-mal Kopf geworfen wurde. Dann erhalten wir ungefähr: $P(E|T^*) = 1/1024$ und $P(E|T) = 0{,}35$ und damit $P(E|T)/P(E|T^*) = 358$. Bei der Erörterung des sogenannten Bayes-Faktors (in Kap. 6.4.2) werden wir sehen, dass das bedeutet, dass T hier sehr viel besser zu E passt – und damit E auch viel besser erklärt –, als es T* kann. Das entspricht außerdem unserer intuitiven Einschätzung dieser Situation. Daher wurde auch der Likelihoodquotient immer wieder als gutes Maß für die Erklärungsstärke betrachtet. Man könnte sagen, für einen Theorienvergleich – und wenn wir noch von dem speziellen Einfluss des weiteren Hintergrundwissens auf P(E) absehen – passt er auch zum Maß von Crupi und Tentori, aber das möchte ich nicht weiter verfolgen.

Das Maß LQ hat ebenfalls weitere schöne Eigenschaften: Sind die Daten etwa statistisch unabhängig relativ zu beiden Theorien (gilt also $P(E_i \,\&\, E_k|T) = P(E_i|T) \times P(E_k|T)$ und Entsprechendes für T* für alle i und k), dann erhalten wir sogar eine *Additivität* für die Vergleichsdaten:

$$LQ(T,T^*{:}E) = LQ(T,T^*{:}E_1) + \ldots + LQ(T,T^*{:}E_m)$$

Es könnten natürlich auch schwierigere Fälle auftreten, in denen T weniger gut zu den anderen Theorien passt als T*, aber dafür etwas bessere Erklärungen der Daten liefert. Außerdem werden wir im Kapitel 5 noch andere Möglichkeiten kennenlernen, wie man im Rahmen des Likelihoodismus – der stark auf den Likelihoodvergleich setzt – gleich mehrere Hypothesen gegeneinander antreten lassen kann.

Schwieriger wird es natürlich, wenn wir nicht nur bestimmen möchten, wie gut die Theorien zu den Daten passen, sondern auch noch, wie gut sie zu unseren anderen Theorien passen, die wir schon akzeptieren. Zunächst sind wir dafür wiederum auf intuitive Einschätzungen und eine Analyse logischer Zusammenhänge angewiesen. Haben die Theorien einen gemeinsamen Anwendungsbereich können wir als ein mögliches Maß für den Abstand von Theorien den sogenannten Kullback-Leibler-Abstand D wählen, der in der Informationstheorie eine wichtige Rolle spielt und uns auch in Kapitel 5 wieder begegnen wird. (Und auch Hawthorne stützt sich in 2011a und 2011c auf ihn, um den Abstand von

Hypothesen im Rahmen der induktiven Logik zu bestimmen, nur dass er ihn anders nennt). Damit können wir für jede der beiden Theorien T und T* bestimmen, wie stark sie sich von anderen *akzeptierten Hypothesen* H aus unserem Hintergrundwissen X unterscheidet, d.h. wie inkohärent sie zu diesen Hypothesen ist. Das Maß wird nur Null für ideal kohärente Theorien und wächst mit zunehmenden Abweichungen an, stellt allerdings keine Metrik dar, da es weder symmetrisch ist noch die Dreiecksungleichung erfüllt. Wir hätten demnach für beide Theorien den folgenden Wert jeweils negativ in Anschlag zu bringen, für alle Daten e_i, für die sowohl die Theorie wie auch die Hypothese eine positive Likelihood liefern:

$$D(H,T) := \Sigma_i \, P(e_i|H) \cdot LQ(H,T{:}e_i)$$

Damit könnten wir etwa die Werte $LQ(T,T^*{:}e_1\&\ldots\&_e n) - D(H,T)$ und $LQ(T^*,T{:}e_1\&\ldots\&_e n) - D(H,T^*)$ als erste Kohärenzabschätzungen für unsere beiden Szenarien miteinander vergleichen.

Das Szenario S ist prima facie kohärenter als das Szenario S* zu den Daten E und anderen Hypothesen H gdw.
$LQ(T,T^*{:}e_1\&\ldots\&_e n) - D(H,T) > LQ(T^*,T{:}e_1\&\ldots\&_e n) - D(H,T^*)$

Dabei können wir natürlich auch noch weitere akzeptierte Hypothesen neben H berücksichtigen. Doch letztlich lassen sich nicht alle Aspekte der Erklärungsstärke und Kohärenz durch einfache Größen erfassen und wir sind am Ende gezwungen, eine z.T. intuitive Abschätzung einiger Aspekte vorzunehmen, was die Gesamtkohärenz unseres Überzeugungssystems mehr erhöht, und ob es überhaupt noch einen klaren Favoriten gibt, den wir auswählen können, oder ob wir in dem Fall nicht lieber Agnostiker bleiben und keine der beiden Theorien akzeptieren sollten.

Probleme der Kohärenzeinschätzung. Außerdem sollten wir an weiteren Beispielen aus der Wissenschaftspraxis untersuchen, ob diese Art der Kohärenzeinschätzung tatsächlich immer zu plausiblen Ergebnissen führt. Vermutlich müssen wir noch einen Anpassungsparameter a (eine positive reelle Zahl) einführen, der die beiden Größen unterschiedlich

zu gewichten gestattet und dann etwa die entsprechenden Werte für $LQ(T,T^*:e_1 \& ... \& {}_en) - a \cdot D(H,T)$ miteinander vergleichen. Um derartige Verrechnungen kommt kein Ansatz induktiven Schließens herum. Das muss aber nicht immer problematisch sein, wie viele Beispiele zeigen. Zumindest kommt auch der Bayesianer nicht um intuitive Einschätzungen herum, denn er muss zu Beginn seines Verfahrens immer quantifizieren, wie gut die vorgelegten Hypothesen H durch das bisherige gesamte Hintergrundwissen gestützt werden bzw. wie plausibel H in unserem bisherigen Hintergrundwissen genau ist.

Ein weiterer problematischer Aspekt sollte schon an dieser Stelle erwähnt werden. Fitelson hat in (2007) ein Beispiel gegen einfache Likelihoodvergleiche angegeben, das uns vielleicht Sorgen bereiten könnte: Wir ziehen eine beliebige Karte aus einem gut gemischten Kartenspiel und haben dazu zwei Hypothesen: (H1) *Die Karte ist ein Pikass* und (H2) *Die Karte ist schwarz* und unser Datum (E) besagt: *Die Karte ist ein Pik*. Dann gilt: P(E|H1)=1 und P(E|H2)=1/2. Demnach würde H1 durch E stärker gestützt als H2, aber andererseits folgt H2 sogar logisch aus E. Das passt nicht zusammen. Das Problem ist hier, dass es sich in dem Beispiel nicht um Erklärungsbeziehungen, sondern vielmehr um logische Zusammenhänge handelt. Die Likelihoodquotienten sollten im Rahmen der Kohärenzkonzeption jedoch nur dazu dienen, einen bestimmten Aspekt der Erklärungsstärke zu bestimmen und sind in diesem Ansatz nicht allgemein und unabhängig vom Vorliegen einer Erklärungsbeziehung als bedeutsam anzusehen.

Der Bayesianer wird vielleicht gegen das ganze Verfahren protestieren und sagen, er hätte doch ein Verfahren, in dem wir alle Theorien nach einem Maßstab bemessen können, nämlich gemäß ihrer Nachher-Wahrscheinlichkeit und das wäre alles, was wir benötigen oder anführen können. Doch wenn wir das ernst nähmen, würde für die Daten $E_1,...,E_n$ immer die Theorie $t = E_1 \& ... \& E_m$ dabei als Sieger hervorgehen. Sie hat im Lichte unserer Daten $E = \{E_1,...,E_m\}$ die Wahrscheinlichkeit 1 und folgt sogar deduktiv aus den Daten (vgl. dazu Kap 5.8.4). Mehr zählt für den Bayesianer strenggenommen nicht. Jedenfalls könnten auch keine Abwertungen durch andere Theorien erfolgen, denn die Daten $E = \{E_1,...,E_m\}$ haben wir schließlich in jedem Fall in X zu integrieren. Die neue »Theorie« t sollte danach sogar besser sein als

echte Theorien mit Erklärungskraft, denn die enthalten immer weitere hypothetische Elemente und sollten daher eine Wahrscheinlichkeit kleiner als 1 erhalten.

Die »Theorie« t *erklärt* allerdings keines unserer Daten und trägt kein Stück zu einem Verständnis oder zu begründeten Vorhersagen bei. Sie ist erkenntnistheoretisch völlig wertlos. Das zeigt wieder, dass wir eigentlich zwei epistemische Ziele haben (keine falschen Meinungen, aber möglichst viele informative Meinungen) und die Idee der hohen Wahrscheinlichkeit nur die erste davon bedient. Daher hilft uns das Verfahren der Bayesianer hier auch nicht wirklich weiter.

Ein Bayesianer hat mir gegenüber dazu einmal den Vorschlag gemacht, es sollten nur solche Theorien anhand ihrer Nachher-Wahrscheinlichkeiten miteinander verglichen werden, die dieselbe Erklärungskraft haben. Einen ersten Schritt in dieser Richtung fanden wir schon bei Brössel (2014). Das kommt jedenfalls dem hier gemachten Vorschlag sehr entgegen, setzt aber schon voraus, dass wir die Erklärungskraft überhaupt berücksichtigen und ein Maß der Erklärungsstärke kennen, was den Bayesianismus deutlich verändern würde.

Vorrang hätte nach diesem Vorschlag die Erklärungsstärke und die Wahrscheinlichkeiten würden erst in zweiter Linie zur Theorienwahl herangezogen. Außerdem bliebe natürlich noch das Problem der gegenseitigen Verrechnung, wenn die eine Theorie mehr erklärt und die andere dafür etwas wahrscheinlicher ist. Es muss ja nicht gleich so ein extremer Fall vorliegen wie bei unserer Theorie t. Um eine Bestimmung der Erklärungsstärken und eine Verrechnung kommen wir also auch im Falle des Bayesianismus nicht herum, denn natürlich wollen wir nicht in jedem Fall eine sehr erklärungsstarke Theorie akzeptieren, wenn diese dafür vollkommen spekulativ bleibt und nur eine sehr geringe Wahrscheinlichkeit aufweist.

An dieser Stelle hat somit Kuhn definitiv einen Sieg davon getragen, wenn er von *methodologischer Inkommensurabilität* spricht. Damit ist neben der Vagheit der Maßstäbe vor allem gemeint, dass wir unterschiedliche Dimensionen der Beurteilung für unsere Theorien haben, die in unterschiedliche Richtungen ziehen können, für die es aber nicht nur eine Möglichkeit der gegenseitigen Verrechnung gibt. So entstehen

Spielräume in der Theorienwahl, die nicht durch zwingende Argumente entschieden werden können. Damit liegt er m.E. richtig, obwohl er noch nicht klar die entscheidenden Dimensionen der Beurteilung zu benennen weiß. Wir haben das bereits in Kapitel 2.1 kennengelernt, wo ich die zwei Ziele der Erkenntnistheorie vorgestellt habe. Jedes für sich ist leicht zu erreichen und eine eindeutige Gewichtung, welches uns im Einzelfall wichtiger ist, liegt hier nicht vor. Konkreter werden die möglichen Konflikte in den Debatten um die Erklärungsstärke und schließlich um die Kohärenz. Die Frage ist aber, ob diese Fälle von schwierigen gegenseitigen Verrechnungen der Vor- und Nachteile von Theorien in der Praxis wirklich so oft auftreten.

De facto müssen wir selbst im Alltag in anderen Kontexten immer wieder derartige Abwägungen verschiedener Kriterien vollziehen. Wenn wir z.B. eine neue Wohnung aussuchen, haben wir in der Regel eine Reihe von Kriterien, die alle in Konflikt miteinander geraten können und das oft genug auch tun wie etwa: der Preis der Wohnung, die Größe, Lage, Verkehrsanbindung, Helligkeit etc. In einigen Fällen wird es uns schwerfallen, eine Abwägung vorzunehmen, weil es gleichgewichtige Vor- und Nachteile auf beiden Seiten gibt, aber in vielen Fällen sind wir uns trotzdem ganz sicher, dass die eine Wohnung besser ist als eine andere, weil sie in einem Punkt klar vorne liegt, selbst wenn sie in anderen Punkten etwas schwächelt.

So ähnlich sieht es auch für wissenschaftliche Theorien aus. Wir haben die relevanten Dimensionen der Beurteilung genauer beschrieben und haben schon erste Daumenregeln für eine Abwägung angeben. Trotzdem behält Kuhn im Prinzip Recht, dass es keinen vernünftig begründbaren einfachen Bewertungsprozess für Theorien gibt, der quasi algorithmisch auf eine bestimmte Theorie hinweist. Es verbleiben manchmal Spielräume der Unterbestimmtheit, und es ist nicht garantiert, dass eine detailliertere Analyse uns jedes Mal weiterhilft.

Eine Frage der Erklärungskohärenz war die, wie unsere Theorien überhaupt zueinander in Konflikt geraten können. Ein Beispiel finden wir in dem Verhältnis zwischen der darwinschen Evolutionstheorie und den klassischen Vorstellungen über die Energieerzeugung auf der Sonne aus der Physik. Lord Kelvin war einer der prominentesten Kritiker Darwins, dessen Überlegungen auch Darwin verunsicherten. Er hatte

nämlich innerphysikalische Gründe dafür, dass die Sonne noch nicht sehr lange brennen konnte (er kannte noch keine Kernfusion), und dann hätte der Evolution nicht genügend Zeit zur Verfügung gestanden, um zu so komplexen Lebewesen zu gelangen, wie wir sie auf der Erde finden. So können Erkenntnisse aus der (Atom-) Physik von entscheidender Bedeutung für die Beurteilung biologischer Theorien sein. Diese Inkohärenz zwischen Theorien kann sogar eine ansonsten erklärungsstarke Theorie erkennbar schwächen.

4.4.6 Moderne probabilistische Kohärenzmaße

In neuerer Zeit wurden vor allem die holistischen, probabilistische Kohärenzmaße diskutiert, die sogar die Problematik isolierter Subsysteme behandeln können. Sie sind allerdings bereits auf den Probabilismus angewiesen, den wir ausführlicher erst im nächsten Kapitel behandeln werden. Sie sind aber zumindest von theoretischem Interesse und können unsere Vorstellungen von Kohärenz verdeutlichen. Daher möchte ich noch einmal explizit auf eines davon eingehen, das in Koscholke (2015) besonders gut abgeschnitten hat.

Für eine endliche Aussagenmenge $X = \{A_1,...,A_n\}$ definieren wir die Gesamtkonjunktion $\bigwedge X := A_1\&...\&A_n$. Dabei handele es sich um kontingente Aussagen einer aussagenlogischen Sprache $\mathscr{L}$.

Systematische Kohärenz: Sei X eine endliche Aussagenmenge und P eine reguläre Wahrscheinlichkeitsfunktion auf der Sprache $\mathscr{L}$, dann ist X *systematisch kohärent* (»any-any coherent«) im Sinne von Meijs (2005) und Douven & Meijs (2007) gdw. für je zwei beliebige Teilmengen S, $S^* \subsetneq X$ mit $S \cap S^* = \emptyset$ gelte: $P(\bigwedge S|\bigwedge S^*)) < P(\bigwedge S)$.

Das ist die stärkste Form von Kohärenz, die Douven und Meijs beschreiben. Hier werden offensichtlich auch holistische Effekte mit berücksichtigt. Aussagen können einander vielleicht einzeln gegenseitig stützen, aber trotzdem im größeren Rahmen zusammengenommen inkohärent sein. Das wird in dem bisherigen Rahmen, in dem nur Paare von Aussagen betrachtet werden, noch kein Thema. Selbst *isolierte Subsysteme* können nun im Prinzip ermittelt werden. Ist etwa S ein solches

Subsystem, das nicht wirklich zu unserem übrigen Hintergrundwissen passt, dann wird es bei unserem Test auf (holistisch-) systematische Kohärenz vermutlich durchfallen, da es dann zumindest Teile unseres Hintergrundwissens R geben wird, die S nicht im oben geforderten Sinne stützen, sondern wahrscheinlich sogar schwächen werden.

Damit muss X natürlich noch nicht in Gänze verworfen werden. Man sieht hier aber auch, dass es sich bei der systematischen Kohärenz um eine sehr starke Anforderung handelt. Es darf noch nicht einmal mehr der Fall sein, dass es in X Teile gibt, die irrelevant füreinander sind. Alle Teilmengen müssen positiv relevant für alle anderen sein. Das könnte man natürlich abschwächen. Auf jeden Fall ist die nächste Aufgabe nun ein quantitatives Maß für die Kohärenz zu entwickeln, um damit auch auf Grade von Kohärenz Zugriff zu haben.

Douven und Meijs haben dazu eine naheliegende Idee vorgestellt, die auf ihrer Idee von systematischer Kohärenz beruht. Wir benötigen zunächst ein (probabilistisches) Maß B(s,r) für die *Bestätigungsstärke*, d.h. dafür, wie stark die Aussage r die Aussage s bestätigt. Im Rahmen des Bayesianismus sind dafür allerdings viele unterschiedliche Maße vorgeschlagen worden und es gibt keinen klaren Sieger, wie wir im nächsten Kapitel noch sehen werden. Ein einfaches Beispiel finden wir im *Differenzmaß* $D(s,r) = P(s|r)–P(s|\neg r)$. Es gibt an, wieviel plausibler s ist, wenn wir annehmen, dass r wahr ist, als wenn wir annehmen, dass r falsch ist. Es kann damit intuitiv als (probabilistisches) Maß dafür dienen, wie stark s durch r gestützt wird. Wir können an dieser Stelle aber auch andere Maße und sogar nicht-probabilistische einsetzen.

Das Bestätigungsmaß B können wir in einfacher Weise auf Mengen von Aussagen ausdehnen: $B(S,R) := B(\bigwedge S, \bigwedge R)$. Nun müssen wir noch alle Paare von geeigneten Teilmengen von X betrachten: $[X] := \{(S,R)|\ S, R \subset X;\ S, R \text{ nichtleer}; S \cap R = \emptyset\} = \{(S_1,R_1),\ldots, (S_m,R_m)\}$ mit $m = \sum_{i=1}^{n} \binom{n}{i}(2^{n-i} - 1)$. Dann können wir zu einem Bestätigungsmaß B und einer Wahrscheinlichkeitsfunktion P ein entsprechendes Kohärenzmaß »Koh« entwerfen, indem wir den Durchschnitt über die Bestätigungswerte über alle Paare (S,R) aus [X] bilden:

Systematische Kohärenz von X: $\mathrm{Koh_B(X)} = 1/m \sum_{i=1}^{m} B(S_i, R_i)$

Zum Vergleich noch einmal die beiden Kohärenzmaße von Shogenji und Olsson, die wir oben schon erwähnt haben, nun in einer allgemeineren Form mit n Aussagen:

$C_{S,P}(X) = P(\bigwedge X) \,/\, \prod_{i=1}^{n} P(A_i)$ (Shogenji 1999)
$C_{O,P}(X) = P(\bigwedge X) \,/\, P(\bigvee X)$ (Glass 2002 und Olsson 2002)

Leider weisen alle Maße gewisse Defizite auf, die von den Vertretern der anderen Maße jeweils anhand von Gegenbeispielen aufgezeigt wurden. Dann bleibt vor allem noch die Frage offen, welches Bestätigungsmaß B wir nun wählen sollten. Im Falles des Differenzmaßes für B verstößt Koh sogar gegen das Prinzip, das Douven und Meijs (2007, 418) selbst nennen, wonach die Kohärenz nicht höher werden darf, wenn wir logische Konsequenzen in die Menge aufnehmen. Sie schlagen dann vor, das Maß auf solche Mengen von Aussagen zu beschränken, die paarweise logisch unabhängig voneinander sind.

Welches Maß B ist aber das beste für unsere Zwecke? Schippers (2014) plädiert anhand einiger Adäquatheitsbedingungen für zwei Bestätigungsmaße. Im Bereich der *inkrementellen* Bestätigungen plädiert er für die normierte Variante λ des Differenzmaßes D, das wir oben schon eingeführt haben, das für Aussagen s und r wie folgt aussieht:

$$\lambda^*(s,r) = \begin{cases} \dfrac{P(s|r) - P(s|\neg r)}{1 - P(s|\neg r)} & \text{falls } P(s|r) \geq P(s|\neg r) \\ \dfrac{P(s|r) - P(s|\neg r)}{P(s|\neg r)} & \text{falls } P(s|r) < P(s|\neg r) \end{cases}$$

Bei den absoluten Bestätigungsmaßen ist nach Schippers einfach die Nachher-Wahrscheinlichkeit P(s|r) als Maß für die Bestätigung von s durch r zu wählen.

Das grundlegende Problem dieser probabilistischen Ansätze bleibt natürlich die Bezugnahme auf die epistemische Wahrscheinlichkeitsfunktion P. Wo kommt die her? Das wird uns im nächsten Kapitel noch

ausführlicher beschäftigen. Das Problem ist dabei: Wenn wir keine objektiven Beziehungen nennen können, die zu diesen Wahrscheinlichkeiten führen, dann bleiben die auf eine rein subjektive Einschätzung angewiesen. Statt das uns die Kohärenzüberlegungen solche Einschätzungen liefern, sind wir von solchen Einschätzungen schon abhängig, wenn wir feststellen wollen, ob Kohärenz vorliegt. Insbesondere wissen wir noch nicht, wie die Wahrscheinlichkeit einer Konjunktion von Aussagen zu bestimmen ist, selbst wenn wir bereits Wahrscheinlichkeiten für die einzelnen Konjunkte vergeben haben. Wir sind an dieser Stelle darauf angewiesen – ebenso wie der klassische Kohärenztheoretiker – einzuschätzen, wie gut die Aussagen zusammenpassen.

Denken wir dazu an ein einfaches Beispiel: In einem Kriminalfall haben wir es nur mit zwei Aussagen zu tun: a = »Der Mörder hat eine Glatze« und b = »Der Mörder hat einen Bart«. Wir kohärent ist nun die Menge X = {a, b}? Das hängt ganz von der gewählten Wahrscheinlichkeitsfunktion ab: Nehmen wir P und P' als zwei solcher Funktionen, und es sei:

Ein Beispiel: $P(a) = P(b) = 0{,}7 = P'(a) = P'(b)$ und außerdem sei $P(a\&b) = 0{,}6$ während $P'(a\&b) = 0{,}4$ sei.
Dann ist: $C_{S,P}(X) = 0{,}6/0{,}49 > 1$ und $C_{S,P'}(X) = 0{,}4/0{,}49 < 1$

Wir arbeiten hier der Einfachheit halber mit dem Shogenji-Maß, das sicher am weitesten verbreitet ist, aber man erhält entsprechende Ergebnisse ebenfalls für die anderen Maße. Im ersten Fall ist die Menge X demnach kohärent und im zweiten Fall nicht mehr. Alles hängt von der Wahl der Wahrscheinlichkeitsfunktion ab. Doch gerade da wird es spannend. Wir müssen einschätzen, ob Mörder mit Glatze auch vermutlich einen Bart haben. Ist das nur so ein subjektive Gefühl, steht unsere ganze Kohärenzeinschätzung auf tönernen Füßen. Die Menge X ist dann etwa kohärent, weil a »gefühlt« b stützt. Doch das wollten wir eigentlich auf eine rationale Basis stellen und durch objektive Zusammenhänge untermauern.

Die klassischen Kohärenzkonzeptionen stützen sich deshalb nach Möglichkeit nur auf die objektiven Likelihoods P(e|h), die eine Hypothese h direkt für ein Datum e vergibt, um diese subjektiven Anteile möglichst

gering zu halten. Wenn h etwa besagt, dass die Wahrscheinlichkeit für Kopf bei einer Münze 0,6 ist, so folgt dann direkt: P(2-mal Kopf|h) = 0,36. Dabei ist keine *gefühlte* Stützung einzusetzen. Allerdings müssen wir zugeben, dass man nicht immer solche objektiven Likelihoods erhält. Unsere wissenschaftlichen Hypothesen sind nicht immer so einfach und so stark, dass sie die Likelihoods vorgeben würden. Darauf kommen wir im nächsten Kapitel zurück.

Wheeler und Scheines (2013) gehen noch einen anderen Weg und betonen, dass oft kausale Zusammenhänge darüber entscheiden, ob scheinbar kohärente Aussagen auch den Bestätigungsgrad der beteiligten Aussagen erhöhen. Insbesondere stellen sie dabei die Kohärenz zwischen den Daten und bestimmten Hypothesen wieder in den Vordergrund.

Denken wir uns etwa den Fall zweier Zeugenaussagen f und u (etwa von Franz und Ute), die übereinstimmend (kohärent) aussagen, dass sie Adolf als den Täter identifizieren. Diese Kohärenz der Aussagen erhöht ihre bestätigende Kraft für unsere Hypothesen h, dass Adolf der Täter war. Doch ist nur solange der Fall, wie wir annehmen, dass die Zeugenaussagen unabhängig voneinander entstanden sind, etwa aufgrund der Beobachtung von Adolf bei der Ausführung der Tat. Beruht u aber darauf, dass Franz der Ute von seiner Beobachtung erzählt hat, so wird uns die zusätzliche Aussage von Ute nicht in unserer Ansicht h bestärken. Also nur, wenn u und f von h verursacht wurden (und nicht von f verursacht wurde), liefert die Kohärenz von u und f einen besonderen Hinweis auf h. Wheeler und Scheines analysieren die kausalen Beziehungen anhand kausaler bayesscher Netze und kommen so zu ersten Einsichten, wie die zugrundeliegende Kausalstruktur den Zusammenhang zwischen oberflächlicher Kohärenz und unseren Bestätigungsbeziehungen vermittelt.

Ich würde das eher so beschreiben, dass wir uns bei Zeugenaussagen immer fragen müssen, was die *beste Erklärung* für diese speziellen Aussagen ist. Sind sie gleichlautend, weil sich die Zeugen unterhalten oder sogar abgesprochen haben, oder sind sie gleichlautend, weil die Zeugen dasselbe gesehen haben? Dementsprechend müssen wir dann die Bestätigungsbeziehungen beurteilen und dementsprechend müssen wir dann auch unsere Wahrscheinlichkeitsfunktionen P aufstellen. Daher bleibt es eine offene Frage, ob der Probabilismus uns tatsächlich weiter-

hilft, wenn es darum geht, die Kohärenz von Überzeugungssystemen zu bestimmen. Er gibt aber zumindest interessante theoretische Modelle dafür an, welche systematischen und holistischen Zusammenhänge dabei relevant sein können. Die praktische Anwendbarkeit scheint dagegen zunächst eher gegeben zu sein anhand einer einfachen klassischen Kohärenzkonzeption, was auch die vielen Fallstudien aus diesem Bereich (etwa von Paul Thagard) belegen.

4.5 Van Fraassens Kritik an der Abduktion

Bas van Fraassen hat (1984) eine spezielle Art von Kritik an dem Schluss auf die beste Erklärung vorgetragen, die in verschiedenen Varianten immer wieder zu hören ist. Sie stützt sich in ihrer speziellen Form auf bestimmte Dutch-Book-Argumente im Rahmen des Bayesianismus, die wir genauer erst im nächsten Kapitel behandeln werden. Sie setzt dazu auch auf das sogenannte Reflexionsprinzip, das ich ebenfalls im nächsten Kapitel als unbegründet ablehnen werde. Aber es gibt einen verständlichen Kern dieses Einwands, den wir auch ohne größeren technischen Aufwand diskutieren können. Van Fraassen gibt ein Dutch-Book-Argument (s. Kap. 5.3.4) dafür an, dass es nicht rational sein könne, dem abduktiven Schlussschema zu folgen, weil man dann Wetten gegen uns etablieren könne, bei denen wir nur verlieren könnten.

Das setzt allerdings schon voraus, dass wir unser Überzeugungssystem probabilistisch gestalten, was wir im nächsten Kapitel kritisch beleuchten werden. Doch die Frage an den Abduktivisten können wir einfacher formulieren: Woher weißt Du, dass die am besten erklärende Theorie auch die wahrscheinlichste ist? Und wenn sie das nicht ist, ist es dann nicht irrational an sie zu glauben?

Es gibt zu dieser Frage inzwischen eine längere Debatte in der Zeitschrift *Analysis*, die mit dem Artikel »What Price Coherence?« (1994) von Peter Klein und Ted Warfield begann. Auf den Einwand, dass die kohärenten Theorien oder Geschichten nicht unbedingt die wahrscheinlichsten sein müssen, sind wir oben schon kurz zu sprechen gekommen, aber da er immer wieder in verschiedenen Varianten gegen den Schluss auf die beste Erklärung vorgetragen wird, möchte ich mich doch noch

expliziter mit ihm auseinandersetzen und gleich den Skeptiker mit ansprechen.

Was spricht überhaupt dafür, eine Theorie zu akzeptieren, die viele Phänomene erklären kann? Die Beispiele sollten als Erstes belegen, dass es sich dabei um ein intuitiv plausibles Verfahren handelt. Wir können natürlich nicht erwarten, dass wir für unsere grundlegendsten Induktionsverfahren selbst wieder zwingende Begründungen finden. Den Zahn sollte uns Hume mit seiner Kritik der Induktion schon gründlich gezogen haben. Die Beispiele belegen aber sehr wohl, dass wir in der Wissenschaft dieses Verfahren anwenden und damit sowohl im Alltag aber vor allem in der Wissenschaft bisher gut gefahren sind. Dann sollten wir uns auch weiter darauf stützen. Die zugrundeliegenden Verfahren der eliminativen Induktion und der hypothetisch-deduktiven Theorienbestätigung waren selbst schon plausibel und das abduktive Schließen war gerade so konstruiert worden, dass es daran anknüpfte, aber deren Schwächen vermied. Somit ist klar, dass eine Theorie, die viele Daten erklären kann, durch diese Daten in jedem Fall ein Stück weit gestützt wird. Das sollten wir anerkennen.

Des weiteren tragen erklärenden Theorien erheblich zur Kohärenz unseres Überzeugungssystems dar, was m.E. ebenfalls ein Indiz für ihre Wahrheit darstellt. Dafür habe ich u.a. in Bartelborth (1996) argumentiert. Meine obigen Analogien mit einem erfolgreich gelösten Rätsel oder Puzzle sollten uns das vor Augen führen. Wiederum dürfen wir nicht zu viel erwarten, sondern können nur darauf verweisen, dass das Verfahren in sehr vielen Anwendungsbeispielen untersucht wurde (u.a.v. Paul Thagard) und sich dabei als sehr erfolgreiche wissenschaftliche Praxis erwiesen hat. Es bleibt eigentlich immer nur noch der Vergleich zu anderen Verfahren zu ziehen. Insbesondere bleibt noch der Verglich mit den Probabilisten zu untersuchen. Sie fragen uns nämlich: Warum sollten wir eine am besten erklärende Theorie akzeptieren, wenn sie nicht die wahrscheinlichste ist? Ist sie aber die wahrscheinlichste, so werden wir sie mit unseren Verfahren am ehesten als solche erweisen und benötigen nicht mehr den Schluss auf die beste Erklärung.

Tatsächlich könnte es passieren, dass eine Theorie T1 die beste Erklärung für eine Menge von Daten E_1,...,E_n liefert, aber eine Theorie T2 sich als wahrscheinlicher im Lichte der Daten erweist. Einen einfachen

Kandidaten für unsere Theorie T2 kann ich sogleich anbieten: Wähle einfach T2 ≡ E_1& ... &E_n. Die hat die Wahrscheinlichkeit 1 im Lichte unserer Daten. Also warum sollten wir lange weitersuchen? Die Antwort habe ich allerdings schon in Kapitel 2.1. gegeben. Wir haben zwei epistemische Ziele zu erfüllen und der Bayesianer hat ganz in der Tradition der Empiristen stehend nur das eine im Sinn (vgl. Bartelborth 2005).

Es kann also tatsächlich passieren, dass wir eine Theorie zu unseren Daten finden, die wahrscheinlicher ist als T1 und die die Daten sogar abzuleiten gestattet, wir sie aber trotzdem nicht gegenüber T1 vorziehen würden. Unsere Theorie T2 kann etwa nichts erklären und nichts vorhersagen. Sie ist eigentlich komplett epistemisch wertlos. Sie geht nicht über die Daten hinaus. Spannende Fälle wären also erst solche, in denen eine etwa gleich gut erklärende Theorie T2 auch noch die wahrscheinlichere wäre. Allerdings würde dann auch diese Theorie eher zur Gesamtkohärenz beitragen, sie hätte etwa einen geringeren probabilistischen Abstand zu unseren anderen Theorien oder den Daten (oder woher soll ihre höhere Wahrscheinlichkeit ansonsten stammen?) und würde damit auch nach dem Verfahren ausgewählt, die Theorien zu akzeptieren, die zu der höchsten Gesamtkohärenz führen.

Wenn wir also nicht wieder in einen Skeptizismus verfallen wollen, gibt es gute Gründe am Schluss auf die beste Erklärung festzuhalten und dann ist auch nicht zu sehen, wieso wir dann nicht die wahrscheinlichste Theorie unter denen wählen sollten, die in etwa die gleiche Erklärungskraft haben. Allerdings können wir wieder auf Kuhns Problem zurückkommen, dass wir keine einfachen Verrechnungsverfahren für den Fall finden, dass eine Theorie, die etwas besseren Erklärungen liefert, während die andere etwas wahrscheinlicher ist. Doch ich habe schon zugegeben, dass gewisse Spielräume in der Theorienwahl verbleiben, die wir nicht auf einfache Weise und zugleich epistemisch zwingend eliminieren können. Die nächste Frage wird im nächsten Kapitel dann aber sein, wie wir überhaupt zu Wahrscheinlichkeiten für Theorien kommen. Das wird ausführlich zu diskutieren sein und sich keineswegs als unproblematisch erweisen. Jedenfalls kann uns die Forderung von van Fraassen, doch immer die wahrscheinlichste Theorie zu wählen kaum überzeugen, da man dann immer die Konjunktionstheorie wählen

müsste, und ist mit Sicherheit um andere Aspekte zu ergänzen. Dann gibt aber der Schluss auf die beste Erklärung einen sehr guten Kandidaten dafür ab.

Das Problem kennen wir schon aus dem Alltag. Nehmen wir zunächst die folgende Geschichte G an: Ute sei mit einem Messer ermordet worden und auf dem Tatwerkzeug seien die Fingerabdrücke von Franz, der Ute gehasst hat und schon mehrfach gedroht hat, sie umzubringen. Er war zur Tatzeit tatsächlich am Tatort, und es gibt keine anderen Verdächtigen. Es waren zudem keine anderen Personen zur Tatzeit in der Nähe des Tatortes, und Ute zeigte auch keine Selbstmordtendenzen. Wir können der Geschichte gerne noch weitere Indizien gegen Franz hinzufügen. Dann wird der Kohärentist schließlich auf die Annahme A schließen, dass Franz die Ute ermordet hat, weil diese Annahme dazu beiträgt, die meisten der in G genannten einzelnen Fakten nun zu erklären. Außerdem scheint uns der Schluss intuitiv zwingend zu sein.

Die Bayesianer werden dazu trotzdem einwenden, dass normalerweise gilt: P(G&A) < P(G). Durch das Hinzunehmen der Behauptung A wird die gesamte Geschichte weniger wahrscheinlich. Das ist natürlich formal richtig, aber wir sind doch gerade daran interessiert, die ganze Wahrheit zu erfahren und dabei auch so umfassende Theorien über die Situation aufzustellen, dass wir die Geschichte verstehen können. Ohne die Annahme A bleibt die ganze Geschichte genaugenommen unverständlich. Wir fragen uns, was da passiert sein mag. Ohne A ergeben die Fakten »keinen Sinn«. Hier zeigt sich wieder, dass das Ziel, einige besonders sichere Wahrheiten über die Welt zu ermitteln, nicht richtig wiedergibt, was wir uns von unseren Theorien über die Welt erwarten. Natürlich möchten wir insbesondere wissen, wer in unserem Fall der Täter ist. Und das ist ganz typisch für unsere Modellierung unserer Umwelt. Wir gehen über unsere direkten Daten hinaus, um zu verständlichen Geschichten zu gelangen, die uns dann erst eine gewisse Handlungskompetenz vermitteln und es gestatten, zielgerichtet in das Geschehen einzugreifen. Das beschreibt der Kohärentismus viel besser als der Bayesianismus, der aber natürlich andere Meriten hat, die wir im nächsten Kapitel kennenlernen werden.

Ein anderer Kritikpunkt, den van Fraassen ebenfalls stark macht, ist der Hinweis auf die Abhängigkeit des abduktiven Schließens von unserer Hypothesenliste. Unser Verfahren der Theorienwahl wird natürlich nicht

zu dem gewünschten Ergebnis (einer zumindest approximativ wahren und erklärungsstarken Theorie) führen, wenn diese nicht in unserer Liste enthalten ist. Da wir in der Liste alle bekannten berücksichtigen müssen und ebenso alle, die uns einfallen, kann es sich dann bei der gesuchten Theorie nur um eine handeln, die uns völlig entgangen ist. Dann kann aber auch keines der anderen Induktionsverfahren uns in dieser Situation weiterhelfen. Das ist eben ein Risiko des induktiven Schließens, das wir nicht völlig vermeiden können. Die meisten der anderen Verfahren sind letztlich gleichfalls auf das Vorgehen der eliminativen Induktion angewiesen – insbesondere der Bayesianismus – und sind damit ebenfalls davon abhängig, dass wir eine Liste aufstellen, die zumindest bereits eine approximativ wahre Theorie enthält. Es ist schließlich auch schon eine wichtige erkenntnistheoretische Aufgabe, die beste Theorie in einer vorgegebenen Liste zu ermitteln.

4.6 Kreative Formen der Abduktion

Abduktive Schlüsse nehmen viele Formen an und unterscheiden sich vor allem darin, worauf wir jeweils schließen (vgl. Schurz 2008). Zu einer Erklärung gehören jeweils die einzelnen Randbedingungen und die Hilfsannahmen sowie die nomischen Muster. All diese Bestandteile der Erklärung können dann Ziele des Schlusses auf die beste Erklärung sein. Im Falle des nassen Gastes, der von draußen hereinkommt, schließen wir auf den Regen. Hier sind uns einige Rahmenbedingungen und vor allem die relevanten Gesetze längst bekannt, und nur das einzelne Faktum, dass es draußen regnet, wird von uns noch abduktiv erschlossen. Aber wir können abduktiv sogar auf noch *gänzlich unbekannte Objekte* schließen. Im Falle des Rinderwahnsinns waren uns bestimmte Mechanismen der Ansteckung durchaus vertraut, nur das infizierende Agens war neu und dazu noch unbekannt und praktisch unbeobachtbar.

Wir können aber ebenso die *Gesetze* erschließen, die hinter bestimmten Phänomenen stehen. Betrachten wir dazu noch einmal das Beispiel der Entdeckung der Cholera aus Kapitel 4.1: Als der Londoner Arzt John Snow im 17. Jahrhundert die Ursachen der Cholera erforschte, waren Infektionskrankheiten noch weitgehend unbekannt. Daher musste

er neue Entitäten postulieren und zugleich neue Hypothesen darüber aufstellen, welche Gesetzmäßigkeiten sich hinter den Erscheinungen der Cholera verbargen. Seine Beschreibung der Mechanismen unterschied sich daher stark von der der Konkurrenztheorie der Miasma-Theorie, die das Ganze als eine Art von Vergiftung auf schlechte Luft zurückführte. Snow musste auf eine komplett neue Theorie schließen, die viele Phänomene erstmals erklären konnte. Auch Newton führte zur Erklärung der Bewegung von Planeten, Pendeln, Kanonenkugeln, des Mondes und eine Reihe anderer Phänomene seine Gravitationstheorie als völlig neue Theorie ein. Er erschloss zugleich eine neue Entität (nämlich die Gravitationskraft) und ein neues Gesetz dafür.

Diese doppelt kreativen Abduktionen sind offensichtlich besonders riskant. Wie gut sind unsere Gründe für die Annahme dieser neuen Entitäten bzw. ihrer Eigenschaften jeweils? Nach Schurz (2008) sollten wir nur solchen kreativen Abduktionen zustimmen, in denen eine *kausale Vereinheitlichung* vorgenommen wird, d.h., in denen unterschiedliche Phänomene auf eine neue grundlegende Eigenschaft zurückgeführt werden.

Das scheint mir allerdings eine etwas zu eingeschränkte Sichtweise der kreativen Abduktion zu sein, die jedoch besonders gut zu seinen Beispielen aus der Chemie passt. Hitchcock und Woodward (2003) halten dagegen diese Art der Vereinheitlichung für nicht so bedeutsam für die Erklärungsstärke und setzen stattdessen ganz auf die *funktionale Invarianz* als lokale Vereinheitlichung im Einzelfall. Dabei geht es darum, dass eine bestimmte funktionale Abhängigkeit zweier Größen auch unter Änderungen dieser Größen invariant im konkreten Anwendungsfall erhalten bleibt. Diese Art von Invarianz eines funktionalen Zusammenhangs deutet für Woodward & Hitchcock bereits auf einen kausalen Zusammenhang hin und kann als Indiz für echte Erklärungskraft im Unterschied zu Pseudoerklärungen dienen. Da m.E. beide Formen von Vereinheitlichung Indizien guter Erklärungen darstellen (s. Kap. 4.3.2), sind sie auch beide geeignet, um abduktive Schlüsse als akzeptabel auszuweisen. Gelingt es also zu einer neuen Entität nomische Muster anzugeben, die ihre Wirkungen auf invariante Weise beschreiben, so dürfen wir ebenfalls auf diese Entitäten schließen, wenn sich diese Muster empirisch bewähren.

Im Falle der Prionen schließen wir sinnvollerweise auf neue Entitäten, obwohl diese zunächst nur einen relativ einheitlichen Phänomenbereich erklären können. Faradays Einführung von Kraftlinien diente dagegen dazu, verschiedene unterschiedliche elektromagnetische Phänomene zu vereinheitlichen und uns zu erklären, wie elektrisch geladene Körper oder Magnete auf andere Körper einwirken können. Sie wurde trotzdem kaum beachtet, weil sie keine sehr große Erklärungskraft besaß, da die Stärke seiner Größen nur geometrisch durch eine höhere Dichte der Kraftlinien angezeigt wurde und die funktionale Invarianz zunächst unerkannt blieb. Erst James Clerk Maxwell gelang es, mit seinen Gleichungen eine gehaltvolle Theorie der Elektrodynamik zu formulieren und entsprechende invariante Generalisierungen anzugeben, die gute Gründe für die Existenz eines *elektromagnetischen Feldes* darstellten. Sie beschreiben dessen Erzeugung und Interaktionen exakt und in nachprüfbarer Weise. Die Rezeption verzögerte sich zwar noch, weil Maxwell und andere damit zunächst ein seltsames mechanisches Äthermodell verknüpften, aber nachdem Heinrich Hertz sie davon befreit hatte, stand der Annahme eines Feldes nichts mehr im Wege. Was sind demnach unsere Kriterien für die Einführung neuer Größen? Zum einen müssen die resultierenden Erklärungen eine möglichst große Erklärungsstärke aufweisen, indem sie gehaltvolle nomische Muster einführen, zum zweiten sollte ein deutlicher Vorsprung in der Erklärungsstärke gegenüber den Konkurrenztheorien vorliegen. Im Idealfall können wir uns das Phänomen E nur noch erklären, indem wir bestimmte Randbedingungen A und ein Gesetz G annehmen, dann spricht das für A und G.

Darunter fallen allerdings auch Dämonenerklärungen (die Schurz in ähnlicher Weise warnend anspricht), wenn tatsächlich keine Konkurrenzerklärungen denkbar sind. So wurden *Wunder* m.E. zu Recht als (schwache?) Hinweise auf das Wirken übernatürlicher Wesen angesehen, solange wir uns keine anderen Erklärungen dafür denken konnten. Sie haben heute nicht mehr diese epistemische Kraft, weil wir sie nicht mehr für echte Wunder halten, d.h., weil uns inzwischen andere Erklärungen für die entsprechenden Phänomene einfallen. Wie gut solche Begründungen für Übernatürliches sind, hängt natürlich von weiteren Annahmen über die Welt ab, denn wir hatten bereits gesehen, dass die Beurteilung von Erklärungen letztlich eine holistische Angele-

genheit ist und solche Abduktionen nur dann zulässig sind, wenn sie die Gesamtkohärenz unseres Überzeugungssystems vergrößern. Unter dem Stichwort der Klabautermanntheorien werde ich gleich noch einmal auf dieses Problem abduktiven Schließens hinweisen.

Zu den wichtigen Formen von Abduktion in der Wissenschaft gehören aber sicher die *Mikroabduktionen*. Wir erklären bestimmte Makrophänomene wie Festigkeit, Wasserlöslichkeit oder elektrische Leitfähigkeit unter Hinweis auf eine unbeobachtbare Mikrostruktur, die wir so abduktiv erschließen. Das ist schon eine recht alte Strategie kreativen Schließens, die allerdings wiederum die Gefahren dieses Verfahrens aufdeckt. So waren die Vorstellungen der antiken Atomisten sicher noch sehr weit entfernt von unseren durch die Quantenmechanik geprägten Vorstellungen von der Mikrostruktur der Materie, aber auch bei ihnen handelte es sich um typische Abduktionsschlüsse. Es ist also eine ständige Herausforderung für unser induktives Schließen, genauer zu bestimmen, in welchen Fällen Schlüsse auf unbekannte oder sogar unbeobachtbare Entitäten gut begründet sind und wann sie eher problematisch werden.

Der Empirist macht es sich einfach, indem er darauf antwortet, sie seien nie gut begründet und wir sollten ganz auf sie verzichten. Allerdings muss er dann einen großen Teil der wissenschaftlichen Theorien als rein fiktiv beschreiben. Es handelt sich für den Empiristen bei vielen theoretischen Größen in der Wissenschaft dann nur um Fiktionen, die einer Systematisierung unseres Wissens dienen, die wir aber nicht zu korrekten Erklärungen heranziehen dürfen, da wir über keine guten Gründe dafür verfügen, an die Existenz entsprechender Objekte oder Eigenschaften zu glauben. Die darauf folgende Debatte um den wissenschaftlichen Realismus werde ich gleich noch einmal kurz aufgreifen.

4.7 Indirekte Formen der Abduktion

Eine schwierige Frage ist außerdem noch, inwiefern auch *Analogieschlüsse* abduktiven Charakter haben. Thagard (2000) und Schurz (2008) gehen davon aus. Eine Idee ist dabei, dass eine bestimmte *Strukturtheorie*, die verschiedenen Phänomenen zugrunde liegt, abduktiv erschlossen

wird. Schließen wir etwa von unserer Kenntnis von Wasserwellen darauf, dass auch Schall aus Wellen besteht, dann identifizieren wir demnach zunächst eine *allgemeine Wellentheorie*. Die erklärt bestimmte Phänomene durch Überlagerung und Auslöschung von Wellen. Finden wir diese Phänomene dann im Schallbereich wieder, können wir unsere Wellentheorie dort wiederum in Analogie zum Ausgangsbereich zum Einsatz bringen.

Auch die Übertragung vom Tierexperiment zum Menschen etwa für eine bestimmte Medizin lässt sich wohl eher als ein Analogieschluss denn als eine einfache Extrapolation verstehen. Die Übergänge sind hier allerdings fließend und wir müssen irgendwo einen sinnvollen Schnitt zwischen einfachen Extrapolationen und Analogieschlüssen ziehen. Man könnte sonst schon die Übertragung bestimmter Ergebnisse von einem Menschen zu einem anderen als Analogieschluss betrachten. Für die hier angegebene Lesart von Analogien spricht, dass die natürliche Art *Lebewesen* in sich bereits sehr heterogen ist und wir natürlich viel eher für die einzelnen Arten von Lebewesen substantielle Verallgemeinerungen erwarten dürfen. Daher werden wir im Normalfall nur diese Gruppierungen unseren Extrapolationen zu Grunde legen, denn letztlich zählt wieder die Erklärungsleistung im Unterschied zur Konkurrenz, für die Frage, wie gut unsere Gründe nun sind, an bestimmte Verallgemeinerungen zu glauben.

Wir sind jedenfalls oft darauf angewiesen, über die einfachen extrapolierenden Schlüsse hinauszugehen und etwa zu Analogieschlüssen zu greifen. Die könnte man als *indirekte Abduktionen* bezeichnen, da wir nicht direkt von einem Phänomen auf ein anderes schließen, sondern zunächst auf eine dahinterstehende gemeinsame Theorie und erst anhand dieser Theorie Vorhersagen für neue Arten von Anwendungen generieren. Ähnliche indirekte Abduktionen finden wir in anderen Formen des induktiven Schließens wieder, die wir bereits kennengelernt haben. Schließe ich vom Fall des Barometers auf den aufziehenden Sturm, so ist das genau genommen ein indirekter Abduktionsschluss. Ohne einen Schluss auf die zugrundeliegende erklärende Kausalstruktur hat der Schluss praktisch kaum Gewicht und genügt zumindest nicht wissenschaftlichen Standards. Wir müssen zumindest vermuten, dass es eine gemeinsame Ursache für beides gibt, um vom Barometerstand auf

einen Sturm schließen zu dürfen, denn sonst nähern wir uns wieder dem Bereich der zufälligen Korrelationen, auf die wir keine Schlüsse stützen dürfen.

Vergleichbare indirekte abduktive Schlüsse finden wir schon in den Schlüssen von Ursachen auf zu erwartende Wirkungen. Geben wir einem Kranken ein Medikament M, von dem wir annehmen, dass es seine Krankheit heilen wird, so sollte das keine einfache Extrapolation darstellen, sondern einen indirekten Weg über die Generalisierung nehmen, dass Kranke eines bestimmten Typs durch M mit einer bestimmten Quote geheilt werden. Diese Generalisierung ist zunächst selbst abduktiv zu erschließen, bevor wir Gründe für unsere Annahme der Heilung des Kranken haben. Da uns dieser indirekte Weg praktisch immer bei abduktiven und beim induktiven Schließen begleitet, werde ich das Prädikat *indirekt* nicht jedesmal wiederholen.

Ich kann jedenfalls Schurz (2008, 202) nicht zustimmen, der für die *konservative Induktion* eine gewisse Sonderstellung gegenüber der Abduktion in Anspruch nimmt:

> That, vice versa, inductions cannot be reduced to abductions is seen as follows. Harman (1965) and Armstrong (1983, p. 78ff) have tried to reduce inductions to abductions by the following argument: the best explanation of the regularities $R(t_i)$ which we have observed at times $t_1,...,t_n$ so far is that they are instances of a *universal law* $\forall t$: R(t). However, I think that this argument is reasonable only if one already *presupposes* that our world is *inductively uniform*. In the absence of an inductive uniformity assumption, there is no reason why the 'true' laws of nature should not change in time, and why the infinitely many *Goodman-laws* $\forall t$: R*(t) (where R*(t): $(t \leq t_n \,\&\, R(t)) \vee (t > t_n \,\&\, R'(t))$, for R'(t) incompatible with R(t)) should not count as equally good candidates for explanation (cf. Howson 2000, p. 43ff). This shows that an independent justification of induction is needed – although I will not speak about this problem in this article.

Schurz hat Recht, dass wir mit einem abduktiven Schluss zugleich bestimmte (metaphysische) Annahmen über unsere Welt unterschreiben. Das müssen wir aber auch schon bei humeschen Induktionen tun,

denn sonst ergeben sie keine sinnvollen Schlüsse. Das wird geradezu nachgewiesen durch die von Schurz genannten goodmanschen Gesetze. Eine allgemeine Gleichförmigkeitsannahme kann gerade die goodmanschen Gesetze vom »grue«-Typ nicht ausschließen. Sie stellen sich als äquivalente Verallgemeinerungen zu unseren normalen Extrapolationen dar. Hier wird R* so gebildet, dass es mit R bis zu einem bestimmten Zeitpunkt t_0 übereinstimmt und danach dann mit R', dass eine andere Behauptung formuliert.

Doch warum sollte R* nicht ein Grundbegriff einer bestimmten Sprache sein, während R und R' dort nur abgeleitet wären? Dann aber wäre in dieser Sprache ∀t: R*(t) durch eine allgemeine Gleichförmigkeitsannahme gerechtfertigt. Es ist zumindest deutlich, dass wir auf einer rein syntaktischen Ebene nicht zwischen sinnvollen und unsinnigen Verallgemeinerungen unterscheiden können. Wir werden durch das »grue«-Paradox gezwungen, eine inhaltliche Einschränkung für solche Generalisierungen vorzunehmen und der Schluss auf die beste Erklärung liefert Hinweise, warum das so ist und in welcher Richtung wir suchen müssen. Wir suchen nach natürlichen Arten, die in geeigneten nomischen Mustern auftreten und nur die haben dann die Eigenschaft der Projizierbarkeit. Welche Prädikate aber natürliche Arten bezeichnen, wird es im größeren theoretischen Rahmen bzw. unseren allgemeinen Modellen der Welt deutlich. Es gibt demnach keine vortheoretischen einfachen Induktionsverfahren, die sich noch nicht auf weiteres Hintergrundwissen stützen müssen, und die wir zunächst begründen können, ehe wir dann zu komplexeren abduktiven Schlüssen übergehen dürfen. Jedes sinnvolle induktive Schließen ist bereits dreistellig und muss sich schon auf weitere Annahmen über die vorliegenden kausalen Strukturen stützen. Dieser Art von Holismus entkommen wir leider nicht und es gibt danach keine akzeptablen theoriefreien konservativen Induktionsschlüsse.

Es scheint also vielmehr eine Schwäche der konservativen Induktion zu sein, die Schurz hier im Auge hat, dass sie versucht, mit möglichst substanzlosen allgemeinen Annahmen zu arbeiten, nach der die Welt einfach möglichst viele, irgendwelche Regelmäßigkeiten aufweist. Doch so verfahren wir glücklicherweise nicht. Stattdessen habe ich dafür argumentiert, dass wir keine allgemeine Gleichförmigkeitsannahme benötigen, sondern vielmehr konkretere (substantiellere) nomische Mus-

ter erschließen müssen, weil wir sonst überall Regularitäten erkennen können, die nur momentane zufällige Zusammenhänge darstellen und dann fälschlicherweise daraus schließen.

Bei der großen Anzahl von Faktoren, die unser Leben bestimmen, werden sich immer bestimmte Regularitäten zeigen, denen wir trotzdem keine Beachtung schenken, wenn wir dahinter nicht einen systematischen kausalen Zusammenhang vermuten. So scheint es immer dann nicht zu regnen, wenn ich einen Schirm mitnehme. Darüber mache ich schon Scherze zu meiner Frau und sage voraus, dass wir wohl vom Regen verschont bleiben werden, wenn ich einen Regenschirm mitnehme. Aber das ist natürlich keine ernstgemeinte Vorhersage, wenn ich nicht einen kausalen Mechanismus dahinter vermuten würde, etwa der Art, dass ein allmächtiger Regenmacher von meinen Handlungen Kenntnis hat und sich ein Spielchen mit mir erlauben möchte. Da ich nicht zu solchen Erklärungen neige, schließe ich nicht im Ernst, dass es nicht regnen wird, wenn ich meinen Regenschirm mitnehme. Vermutlich habe ich einfach das Glück in einer relativ regenarmen Stadt zu leben, in der trotz aufziehender Regenwolken in den meisten Fällen kein Regen fällt, sondern diese über uns hinwegziehen. Dann sollte ich mir aber auch keine Sorgen machen, wenn ich den Regenschirm mal daheim gelassen habe. Es zeigt sich wieder, dass die genaue Art des induktiven Schließens stark von weiteren Hintergrundannahmen abhängt und keineswegs eine zweistellige Struktur hat, wonach wir von Daten gleich auf bestimmte Generalisierungen oder andere Daten schließen können.

Harman und Armstrong haben also Recht, dass wir nur dann induktiv schließen dürfen, wenn wir damit die Annahme verbinden, dass zumindest ein nomisches Muster (das ist etwas weniger als ein allgemeines Gesetz) hinter diesem Schluss steht, das wir zunächst abduktiv erschließen müssen, weshalb insgesamt wiederum ein indirekter abduktiver Schluss vorliegt. Sonst entstehen zu viele Regularitäten, die zu Fehlschlüssen Anlass geben würden. Fälle von »grue«-artigen Prädikaten verlangen nach speziellen Gleichförmigkeitsannahmen und nicht nach allgemeinen.

Psychologische Resultate von Thomas Gilovich (1991) zeigen allerdings, dass wir zu schnell selbst dort nach Mustern suchen, wo eigentlich keine vorliegen. So sind Sportler etwa im Basketball nicht davon abzubringen,

dass sie Phasen mit einer »glücklichen Hand« haben, in denen ihnen fast jeder Wurf gelingt und andere, in denen es in der anderen Richtung geht, obwohl eine statistische Auswertung der Daten das keineswegs bestätigt. Das zeigt für uns wieder einmal, dass wir bei Schlüssen auf die beste Erklärung große Vorsicht walten lassen müssen und immer offen für die Daten und alternative Erklärungen bleiben sollten.

Im Folgenden werde ich jedenfalls in der Richtung von Harman argumentieren, dass alle anderen induktiven Schlussformen genau genommen Unterarten der Abduktion im hier verstandenen Sinne sind. Wir dürfen nur dann aus bestimmten Daten auf neue Daten extrapolieren, wenn dieser Schluss durch zugrundeliegende nomische Muster gedeckt wird. Das Grundmerkmale des abduktiven Schließens, das typischerweise komparativ und eliminativ ist, und das auf zugrundeliegende nomische Muster mit Erklärungswert angewiesen ist, finden wir in allen Schlussformen mehr oder weniger wieder. Allerdings sind abduktive Schlüsse, besonders wenn wir die indirekten mit hinzunehmen, sehr vielgestaltig und lassen sich nicht auf ein einfaches Schema zurückführen, so gerne wir das induktive Schließen weiter vereinheitlichen würden.

4.8 Problemfälle abduktiven Schließens: »Klabautermanntheorien«

Eine ständige Herausforderung für das abduktive Schließen sind »Klabautermanntheorien«, so möchte ich sie jedenfalls nennen. Das sind Theorien, die uns letztlich als unbegründet erscheinen, für die es aber (scheinbar?) gute abduktive Gründe gibt. Der Klabautermann ist eine Figur aus dem Aberglauben der Seemänner, der Eingang in die Literatur gefunden hat. Er ist meist unsichtbar und warnt den Kapitän eines Schiffes vor Gefahren. Insbesondere geht er von Bord, wenn das Schiff sinkt. Bemerkbar macht er sich vor allem durch Polter- und Bumsgeräusche (s. Wikipedia). Dort liegt wohl seine vornehmliche »Erklärungskraft«. Doch der Schluss von derartigen Geräuschen auf die Annahme, das müsse der Klabautermann gewesen sein, scheint uns eindeutig unseriös zu sein. Dazu müssen wir aber fragen, ob uns der Vertreter der Abduktion

das erklären kann. Stellt die Einführung des Klabautermanns eine gute Erklärung für bestimmte Phänomene im Schiff dar? Das ist natürlich eine analoge Frage zu der Einführung anderer Geister in entsprechenden Kontexten und lässt sich schließlich auch auf die Einführung theoretischer Größen wie Felder oder Neutrinos oder bohmsche Quantenpotentiale übertragen. Wann dürfen wir abduktiv auf die Existenz derartiger Entitäten schließen und wann sollten wir uns lieber zurückhalten? Wie lässt sich eine Antwort darauf mit den Hilfsmitteln der Abduktion erläutern? Es gibt keine einfache Antwort auf diese Fragen, sondern sie verlangt immer nach einem komplexen holistischen Abwägungsprozess, der die Frage beantworten muss: Welches Modell unserer Welt ist im Lichte unserer Daten insgesamt kohärenter? Ist es das ohne Annahme eines Klabautermanns oder das unter der Annahme, dass ein Klabautermann einer bestimmten Art existiert?

Was muss eine Klabautermanntheorie nach Möglichkeit alles mitbringen, damit sie den Klabautermannstatus hinter sich lassen kann und den Status einer seriösen Theorie erreichen kann? Dazu müssen wir die Klabautermanntheorien gemäß unseren Bewertungskriterien beurteilen. Bieten sie etwa Theorien, die gehaltvolle (kausale) nomische Muster aufzeigen? Dazu müssten sie u.a. Zusammenhänge aufzeigen, die Auskunft darüber geben, wie wir zumindest im Prinzip durch bestimmte Interventionen Veränderungen in den Verhaltensweisen und Auswirkungen der Klabautermänner angeben. Das ist mir in diesem Falle nicht bekannt und fehlt z.T. auch in anderen Theorien wie der Astrologie.

Wir müssen weiterhin überprüfen, wie erfolgreich und gehaltvoll die Erklärungen und insbesondere die Vorhersagen der Theorie tatsächlich sind. Auch da können Klabautermanntheorien nicht gerade punkten. Außerdem müssen wir immer nach alternativen Theorien Ausschau halten, die ebenfalls die in Frage stehenden Phänomene erklären können. Auf Segelschiffen können Taue an Stellen des Schiffes schlagen, Holzplanken reiben aneinander und das Holz »arbeitet«. Das kann eine Vielzahl von Geräuschen erklären, ohne dass wir zu neuen Entitäten wie Klabautermännern greifen müssen.

Letztlich gibt es aber keine Kristallkugel, die für alle Theorien vorhersehen kann, ob sie nicht doch irgendwann Erklärungserfolge feiern werden. Es gilt also auch die fallibilistische Regel, dass wir uns unserer

Theorien nie ganz sicher sein dürfen und somit immer ein offenes Ohr für andere Überlegungen bewahren müssen. Der Wunsch, der oft an die Wissenschaftstheorie herangetragen wird, sie möge doch einfache Kriterien vorlegen, nach denen sich sofort unsinnige Theorien erkennen und aussondern lassen, ist daher zwar verständlich, aber eben nicht realisierbar. Wir müssen uns immer der Mühe unterziehen, komplexere Bewertungen vorzunehmen gemäß der Richtschnur: Liefert die Theorie gute Erklärungen für viele Phänomene, und wie gut sind diese Erklärungen im Vergleich zu denen der Konkurrenztheorien? Die in den nächsten Kapiteln zu erörternden Schlussverfahren können dazu weitere Hilfestellungen anbieten, bieten aber auch nicht die einfachen hellseherischen Verfahren an, die wir uns wünschen würden.

4.9 Wissenschaftlicher Realismus

Ein wichtiges Anwendungsfeld der Abduktion und der Kohärenzüberlegungen innerhalb der Philosophie finden wir in den Fragen des Realismus. Uns soll es vor allem um den *wissenschaftlichen Realismus* gehen. Die grundlegende Fragestellung ist hier, ob die von unseren Theorien postulierten Entitäten tatsächlich existieren oder ob wir mit unserem Akzeptieren einer Theorie nur behaupten möchten, dass sie empirisch adäquat ist, also die beobachtbaren Phänomene richtig beschreibt, wie es van Fraassen (1980) annimmt. Dazu sind verschiedene antirealistische oder instrumentalistische Deutungen der theoretischen Größen einer Theorie denkbar. Sie werden dann z.B. als bloß heuristische Hilfsmittel betrachtet, mit deren Hilfe wir uns leichter ein Bild der Welt machen können, ohne dass wir sie als korrekte Beschreibungen der Welt annehmen sollten.

So wie man sagt: (A) »Die Pflanze wendet sich der Sonne zu, weil sie mehr Licht aufnehmen möchte.« Mit (A) wollen wir nicht wirklich behaupten, dass die Pflanze einen eigenen Willen hätte. Es ist einfach eine vermenschlichende Beschreibung, die uns hilft, das Verhalten der Pflanze zu verstehen und sogar ein Stück weit vorherzusagen, ohne dass wir uns mit den physiologischen Details der Pflanze auseinandersetzen müssen. Ähnlich gehen wir oft vor, wenn wir das Verhalten von Tieren

erklären. Wenn man uns dazu befragt, geben wir meistens zu, dass wir diese »Erklärung« nicht wörtlich nehmen. Es bleibt dann allerdings die weitere Frage offen, ob es sich dabei überhaupt noch um eine Erklärung handelt. Schon Hempel forderte zu Recht, dass tatsächliche Erklärungen im Unterschied zu bloß potentiellen Erklärungen ein wahres Explanans aufweisen müssen.

Gegenüber der *instrumentalistischen Deutung* behaupten wissenschaftliche Realisten zunächst, dass wir mit dem Akzeptieren einer Theorie zugleich die Entitäten akzeptieren müssen, die von der Theorie postuliert werden. Hier gibt es als Erstes den bescheidenen reinen *Entitätenrealismus*, der nur fordert, dass die Entitäten akzeptiert werden müssen, ohne damit zugleich zu fordern, dass unsere Theorie die meisten Eigenschaften dieser Entitäten bereits zutreffend beschreibt. Daneben gibt es den *anspruchsvolleren wissenschaftlichen Realismus*, der auch diese weitergehende Forderung unterschreibt. Ein »richtiger« wissenschaftlicher Realist wird sich dabei wohl immer dem anspruchsvolleren Realismus verschreiben.

In der Debatte um den wissenschaftlichen Realismus sind an unterschiedlichen Stellen Schlüsse auf die beste Erklärung im Spiel. Es fängt bereits auf der normalen Objektebene der wissenschaftlichen Theorien an. Wenn wir die Theorien abduktiv erschließen, schließen wir bereits auf ihre Wahrheit, denn bloße Klabautermanntheorien erklären nicht. Sie liefern potentielle Erklärungen, aber keine aktualen. Das sollten uns Beispiele schnell vor Augen führen. Bin ich davon überzeugt, dass es keine Klabautermänner gibt, wird eine Klabautermann-Erklärung für die Geräusche auf einem Schiff nicht mehr ernsthaft akzeptiert. So geht es uns auch in der Wissenschaft. Stellt sich etwa heraus, dass es doch keine Neutrinos gibt, wären Erklärungen für die Energieerhaltung beim Beta-Zerfall, die sich auf Neutrinos stützen, nicht mehr akzeptabel. Wenn wir also eine Erklärung als richtig akzeptieren, akzeptieren wir damit im Normalfall die Existenz der daran beteiligten Entitäten. Würden diese Entitäten sich dann ganz anders verhalten als in den Theorien behauptet, würde das ebenso unsere Erklärungen untergraben, allerdings sind sicher gewisse kleinere Fehler in der Beschreibung erlaubt.

Gegen die Annahme, dass Erklärungen i.w. wahre Theorien voraussetzen, argumentierte Bas van Fraassen (1980) unter Hinweis auf die Erklä-

rungen durch die newtonsche Gravitationstheorie, von denen wir viele noch immer akzeptieren, obwohl wir inzwischen wissen, dass die Theorie falsch ist und eigentlich durch die entsprechende relativistische Theorie zu ersetzen wäre. Allerdings ist die newtonsche Gravitationstheorie keine Klabautermanntheorie. Sie postuliert bestimmte Gravitationskräfte zwischen je zwei Körpern und gibt an, wie die zu Beschleunigungen führen. Diese Kräfte existieren und werden von der newtonschen Theorie auch i.w. korrekt beschrieben. Nur für besondere Situationen, in denen hohe Geschwindigkeiten auftreten oder sehr große Räume betrachtet werden, benötigen wir einen relativistischen Korrekturterm. Das ist alles andere als eine Klabautermanntheorie. Wir erkennen die Erklärungen der newtonschen Theorie insbesondere deshalb noch an, weil wir sie für *approximativ korrekt* halten. Also ist das erste abduktive Argument für die Existenz bestimmter theoretischer Größen die normale abduktive Rechtfertigung dieser Theorien. Vertreter der Ansicht, dass selbst falsche Theorien erklären können, müssten zeigen, dass auch echte Klabautermanntheorien wie die Phlogistontheorie (bei denen wir also bestimmte Größen inzwischen komplett verworfen haben) immer noch Erklärungskraft aufweisen. Solche Fälle sind mir jedenfalls nicht bekannt.

Darüber hinaus gibt es aber noch ein recht allgemeines Argument, dass sich in verschiedenen Varianten ebenfalls auf einen Schluss auf die beste Erklärung stützt. Diese Erklärung ist aber eher intuitiv und keine kausale Erklärung, die Ableitungen aus nomischen Mustern vornimmt. Sie ist auch als das »*no-miracle* Argument« von Putnam (1975) bekannt. Danach wären die Erfolge unserer akzeptierten Theorien als ein *Wunder* zu betrachten, wenn sie nicht im Großen und Ganzen wahr wären. Diese Idee steckt natürlich ein Stück weit hinter dem ganzen induktiven Schließen und schon hinter der hypothetisch-deduktiven Theorienbestätigung. Es wäre wohl ein sehr großer Zufall, wenn unsere konkreten Vorhersagen anhand unserer Theorien zwar immer wieder zutreffen, obwohl die Theorien eigentlich ganz falsche Beschreibungen der Welt lieferten.

Nehmen wir an, wir hätten möglicherweise die falsche Karte auf eine Wanderung mitgenommen, dann würden wir kaum erwarten, dass sie uns trotzdem immer wieder korrekte Hinweise auf unseren Weg liefert. Je häufiger und an umso mehr Stellen unsere Karte uns zutreffende

Daten über unsere Gegend liefert, umso mehr sind wir zu Recht davon überzeugt, dass es sich um eine i.w. korrekte Beschreibung der Gegend handelt. Sicherheit können wir allerdings beim induktiven Schließen dafür nie erwarten. Aber es wäre schon ein Wunder, wenn unsere falsche Karte immer die richtigen Beschreibungen unseres jeweiligen Weges liefern würde. Erst wenn wir annähmen, dass jemand – wie Descartes böser Dämon – uns systematisch in die Irre führen wollte, sähe die Sache schon etwas anders aus, aber wir gehen im Hinblick auf unsere Naturerkenntnis nicht davon aus, dass es da jemanden gibt, der unsere Sinne beliebig täuschen kann und das auch tun möchte. Solche Ideen verfolgt höchstens der radikale Skeptiker, den wir aber bereits hinter uns gelassen haben.

Die Meta-Erklärung des allgemeinen Erfolgs unserer Theorien anhand ihrer Wahrheit verlangt allerdings nach anderen Formen von Erklärung als die überwiegend kausalen Erklärungen auf der Objektebene. Hier sind wir noch viel mehr auf eine vage und intuitive Einschätzung der Erklärungskraft angewiesen als im Falle der Erklärungen auf der Objektebene. Doch das ist in der speziellen Debatte um den wissenschaftlichen Realismus weiter auszuloten und soll hier nicht verfolgt werden.

Der moderate Skeptiker stellt höchstens in Frage, ob unsere Theorien tatsächlich immer so erfolgreich sind, wie wir das in dem Argument als Prämisse annehmen. Er setzt dem die *pessimistische Metainduktion* entgegen, wonach wir aus der Geschichte der Wissenschaften vielmehr extrapolieren sollten, dass sich alle Theorien letztlich als falsch herausstellen. Das kann nur anhand einer sorgfältigen Betrachtung der Wissenschaftsgeschichte beurteilt werden und geht oft damit einher, dass wir auch zwischen Theorien in einem frühen Stadium der Entwicklung einer Disziplin und einer in einem fortgeschrittenen Stadium unterscheiden, doch das geht über unseren jetzigen Rahmen hinaus.

4.10 Fazit

Der Schluss auf die beste Erklärung gibt für alle bisher vorgestellten induktiven Schlussverfahren den Rahmen ab, innerhalb dessen wir sie anwenden sollten. Die einfachen Extrapolationsverfahren übersehen

zunächst den Einfluss des Hintergrundwissens, aber in kritischen Situationen wird deutlich, dass wir es eigentlich berücksichtigen sollten und sonst zu offensichtlich falschen Schlüssen gelangen. Im Rahmen der falsifikationistischen Ansätze wird ersichtlich, dass es letztlich doch »nur« Erklärungsanomalien sind, die zu einer Elimination bestimmter Hypothesen führen. Diese Eliminationen sind aber nur dann tatsächlich weiterführend, wenn wir eine weitere Annahme des abduktiven Schließens unterschreiben, nämlich, dass es uns gelingt, eine endliche Liste aller relevanten Hypothesen aufzustellen, unter denen wir eine wahre Hypothese mit guten Gründen vermuten.

Auch die positive Bestätigung von Theorien, die sie aus erfolgreichen Vorhersagen gewinnen, sind vor allem anhand der Erklärungsleistung der Theorien zu beurteilen. Erst über die Erklärung von Daten gewinnt eine Theorie tatsächlich eine induktive Bestätigung, denn erst dadurch stützen wir die Vermutung, auf ein grundlegendes nomisches Muster in unserer Welt gestoßen zu sein, das wir dann auch für Vorhersagen nutzen dürfen. Allerdings liefert uns das Schema des abduktiven Schließens nur eine Reihe von Kriterien für die Beurteilung, wie stark unsere Theorie nun durch bestimmte Daten gestützt wird. Eine gegenseitige Verrechnung ist einer lokalen Analyse der Einzelfälle vorbehalten und muss natürlich nicht immer zu einem eindeutigen Ergebnis führen. Dabei ist vor allem wichtig, dass wir uns auf Vergleiche zwischen einzelnen Theorien zurückziehen können. Das ist letztlich eine wesentliche Einsicht, wonach wir mit unseren Theorien vor allem die Konkurrenz besiegen müssen und wenn es geht, epistemisch eliminieren sollten, um unseren Favoriten zu bestätigen.

Selbst außerhalb der empirischen Wissenschaften setzen wir gern auf Schlüsse auf die beste Erklärung und weitergehende Kohärenzannahmen. Dazu hat Paul Thagard (2000) zahlreiche Beispiele zusammengetragen. Die belegen wiederum, wie zentral das abduktive Schließen für unser epistemisches Projekt ist. Es beherrscht unser ganzes induktives Schließen. Im nächsten Kapitel werden wir nun einen echten Konkurrenten genauer kennenlernen, der bereits einige Grundannahmen verändert, wie etwa die, wie unser Überzeugungssystem zu repräsentieren ist. Trotzdem werden wir viele Annahmen des induktiven Schließens wieder finden, wie etwa die, dass wir auf eine Liste von Hypothesen angewiesen

sind und dass die Likelihood P(E|H) ein wesentlicher Parameter für die Stärke der Bestätigung einer Hypothese H durch ein Datum E ist.

5 Probabilistische Ansätze

Es ist eine naheliegende Idee, unsere Unwissenheit über unsere Theorien und unsere Daten sowie unsere induktiven Schlüsse mit Hilfe des Begriffs der Wahrscheinlichkeit zu beschreiben. Einige Ansätze versuchen auf diesem Wege außerdem genauer zu bestimmen, wann eine induktive Rechtfertigung vorliegt und möglichst sogar zu quantifizieren, wie stark sie ausfällt. Der prominenteste Ansatz ist hier sicher der *subjektive Bayesianismus*. Bevor wir uns diesen Ansatz genauer ansehen können, sind allerdings noch einige Vorarbeiten zu leisten.

5.1 Ein grundlegendes Beispiel: Dschungelfieber

Zur Einführung in dieses Gebiet möchte ich mit einem einfachen fiktiven Beispiel beginnen, das in ähnlicher Form immer wieder als besonders erfolgreicher und typischer Anwendungsfall für den Bayesianismus genannt wird (vgl. a. Beck-Bornholdt & Dubben 1998). Nehmen wir an, Herr Vorsichtig kommt aus dem Dschungelurlaub in den Tropen zurück. Er weiß, dass man sich dort das *Dschungelfieber* holen kann. Diese Krankheit bricht erst nach einigen Wochen aus und ist dann schwer zu behandeln. Also möchte er schon jetzt Gewissheit darüber haben, ob er sich nun infiziert hat oder nicht. Deshalb lässt er beim Arzt einen Bluttest auf Dschungelfieber durchführen. Der Arzt beruhigt ihn zunächst und meint, dass sich nur jeder *tausendste Urlauber* infizieren würde. Außerdem hat er einen sehr zuverlässigen Test auf Dschungelfieber, den er nun an Herrn Vorsichtig ausprobiert. Überraschenderweise fällt der Test T positiv aus (T^+). Wie wahrscheinlich ist es aber nun, dass Herr Vorsichtig tatsächlich Dschungelfieber hat? Dass das so ist, sei unsere Hypothese H.

Dazu müssen wir als Erstes wissen, wie zuverlässig der Test tatsächlich ist. Üblicherweise werden hierfür zwei Kennzahlen herangezogen, nämlich die *Sensitivität* und die *Spezifität* des Tests. Die seien in unserem

Fall: Sensitivität $P(T^+|H)$ = 99% und Spezifität $P(T^-|\neg H)$ = 97%. Das sind die bedingten Wahrscheinlichkeiten zunächst dafür, dass der Test anschlägt (T^+), wenn wir einmal annehmen, dass unsere Hypothese H wahr ist, Herr Vorsichtig also infiziert ist, und dann dafür, dass der Test nicht anschlägt, wenn wir annehmen, dass Herr Vorsichtig nicht infiziert ist. Die Werte sollen bedeuten, dass der Test bei 100 Infizierten im Durchschnitt 99-mal Alarm schlägt, allerdings auch bei 100 Nichtinfizierten 3-mal Fehlalarm gibt (bzw. nur in 97 Fällen Entwarnung gibt). Das sieht dann wohl nicht gut aus für Herrn Vorsichtig, denn der Test scheint eine hohe Aussagekraft zu besitzen.

Die meisten Menschen (übrigens auch die Ärzte vgl. Gigerenzer 2002) schätzen, dass die Wahrscheinlichkeit deutlich über 90% liegen muss, dass Herr Vorsichtig bei positivem Testresultat nun auch tatsächlich Dschungelfieber aufweist. Doch das ist ein Irrtum. Wir neigen dazu, die *Prävalenz* oder *Basisrate* für Dschungelfieber von 1/1000 zu übersehen, die der Bayesianer berücksichtigt, der klassische Statistiker nach Ansicht der Bayesianer aber ebenfalls übersieht. Der Bayesianer spricht an dieser Stelle vom *Basisratenfehlschluss*. Doch dazu später mehr. Die gesuchte Wahrscheinlichkeit berechnet sich anhand des bayesschen Theorems und dem sogenannten Theorem der totalen Wahrscheinlichkeit für $P(T^+)$:

Bayessches Theorem (ausführlichere Variante)

$$P(H|T^+) = P(H) \cdot \frac{P(T^+|H)}{P(T^+|H) \cdot P(H) + P(T^+|\neg H) \cdot P(\neg H)}$$

Hier kommt nun unter anderem die sogenannte *Prävalenz* zum Tragen. Das ist die Wahrscheinlichkeit dafür, dass Herr Vorsichtig überhaupt infiziert zurückgekommen ist: P(H) = 1/1000. Entsprechend ist P(¬H) = 999/1000. Außerdem ergibt sich $P(T^+|\neg H)$ = 3/100. Setzen wir diese Werte in (*) ein, erhalten wir für die Wahrscheinlichkeit, dass Herr Vorsichtig nach positivem Test nun Dschungelfieber aufweist: $P(H|T^+)$ = 0,032, also bloß ca. 3,2%.

Die Daten der Berechnung

P(H) = 1/1000 P(¬H) = 999/1000
$P(T^+|\neg H)$ = 3/100 $P(T^+|H)$ = 99/100
Resultat: $P(H|T^+)$ = 0,032

Das Ergebnis zeigt, wie sehr wir daneben gelegen haben, aber es überzeugt die meisten Personen bei dieser Berechnungsart noch nicht. Doch die folgende Abschätzung sollte es tun. Eine Veranschaulichung durch relative Häufigkeiten ist nach Gigerenzer 2002 meistens anschaulicher und erfolgreicher, wenn wir mit Wahrscheinlichkeiten rechnen müssen: Nehmen wir an, 100 000 Menschen kämen aus den Tropen zurück und ließen sich testen, dann hätten ca. 100 davon Dschungelfieber. Davon würden 99 positiv getestet. Von den übrigen knapp immer noch 100 000 gesunden Menschen würden aber immer noch ca. 3000 Personen positiv getestet. Damit würden also insgesamt 3099 Personen positiv getestet, aber nur 100 davon hätten auch Dschungelfieber. Also wäre die Quote 99/3099 = 0,032 dafür, an Dschungelfieber zu leiden, wenn man positiv getestet wird. Das ist erheblich weniger als bisher gedacht. Wir haben die Prävalenz bzw. Vorher-Wahrscheinlichkeit dafür, überhaupt an Dschungelfieber zu erkranken, zunächst übersehen.

100 000 Urlauber	davon positiv getestet
davon ca. 100 mit Dschungelfieber	99
weiterhin immer noch ca. 100 000 ohne Dschungelfieber	3000
	also 3099 positiv, aber davon nur 99 mit Dschungelfieber

Tabelle 5.1: Test auf Dschungelfieber

Doch das ganze Schlussverfahren enthält bereits eine ganze Reihe an einzelnen Annahmen oder Teilschlüssen, die wir nun genauer unter die Lupe nehmen wollen. Zunächst einmal stellt sich die Frage, warum wir überhaupt die bedingte Wahrscheinlichkeit $P(H|T^+)$ für so bedeutsam halten. Das ist die Wahrscheinlichkeit dafür, dass Herr Vorsichtig das Dschungelfieber hat, wenn wir schon davon ausgehen, dass sein Test positiv ausfällt. Das ist uns vermutlich deshalb so wichtig, weil wir sie als die neue Wahrscheinlichkeit interpretieren, dass Herrn Vorsichtig nach positivem Testausgang nun tatsächlich an Dschungelfieber leidet: $P_{nachher}(H) = P_{vorher}(H|T^+)$. Das nennt man in der allgemeinen Form mit bestimmten Daten E auch die *Konditionalisierungsregel*.

Bayessche Konditionalisierungsregel:
$P_{nachher}(H) = P_{vorher}(H|E)$

Sie ist die zentrale Regel des *Bayesianismus* und keineswegs so unumstritten, wie sie hier erscheint. Sie besagt, dass wir, sobald eine neue Beobachtung E (E für »evidence«) oder mehrere aufgetreten sind, als neue Wahrscheinlichkeit für eine Hypothese H die Wahrscheinlichkeit wählen sollten, die wir vor dem Auftreten von E als bedingte Wahrscheinlichkeit P(H|E) angenommen haben. Das ist in gewisser Weise eine Form von *zeitlicher Kohärenz* oder *Konsistenz*. Wir halten uns damit heute nur an das, was wir gestern geglaubt haben.

Aktiv könnte man das so interpretieren: Haben wir bisher behauptet, dass wir H die Wahrscheinlichkeit r zuweisen würden, sollte E wahr sein, und nun ist tatsächlich noch E aufgetreten, dann weisen wir H nun schlicht die Wahrscheinlichkeit r zu und beginnen nicht mit einer kompletten Neubewertung. Das bayessche Theorem selbst spielt in der Konditionalisierungsregel noch keine Rolle. Es wird nur eingesetzt, um die bedingte Wahrscheinlichkeit P(H|E) rechnerisch zu bestimmen, aber wir sollten im Blick behalten, dass es sich beim bayesschen Theorem nur um eine mathematisch wahre Aussage handelt. Die eigentliche Behauptung des Bayesianismus steckt in der Konditionalisierungsregel, auf die wir später noch eingehen werden. Nur so viel: Sie kam oben zum Einsatz, indem der positive Testausgang unser empirisches Datum darstellte und wir haben dann die beiden folgenden Theoreme angewandt, die sich leicht aus den Wahrscheinlichkeitsaxiomen ergeben (s.u.).

Bayessches Theorem

$$P(H|E) = P(H) \cdot \frac{P(E|H)}{P(E)}$$

Gesetz von der totalen Wahrscheinlichkeit

$P(E) = P(E|H) \cdot P(H) + P(E|\neg H) \cdot P(\neg H)$

Dabei wird der rechte Ausdruck oft zur Berechnung von P(E) im Nenner ins bayessche Theorem eingesetzt, und so haben wir auch die obere Gleichung für das bayessche Theorem in der ausführlicheren Form erhalten. Das Gesetz von der totalen Wahrscheinlichkeit besagt einfach,

dass wir, um die Wahrscheinlichkeit von E zu bestimmen, zwei Fälle betrachten müssen: Zum einen den Fall, in dem H gilt und was das dann für E bedeutet. Und zum anderen den »Rest«, also hier den Fall, in dem H nicht gilt und was dann für E bedeutet. Die Ergebnisse für diese beiden einander ausschließenden Fälle sind anschließend zu addieren, wobei wir als Gewichte berücksichtigen müssen, wie häufig (oder wie wahrscheinlich) H und ¬H auftreten.

Außerdem ist klar, dass wir es hier meistens mit sogenannten *subjektiven Wahrscheinlichkeiten* oder *Glaubensgraden* zu tun haben, die z.B. ein klassischer Statistiker vollständig ablehnt. Dass Herr Vorsichtig nun Dschungelfieber hat, trifft entweder zu oder ist falsch. Es sollte also nur die trivialen objektiven Wahrscheinlichkeiten 0 oder 1 aufweisen. Die genannten Wahrscheinlichkeiten drücken dagegen unser Unwissen aus bzw. unseren Kenntnisstand im Hinblick auf das Vorliegen von Dschungelfieber (vgl. Kap. 5.3). Ein klassischer Statistiker dürfte dagegen nur sagen: Das ganze Verfahren des Tests T ist von der Art, dass in ca. 3,2% aller Fälle mit positivem Testergebnis auch ein Fall von Dschungelfieber vorliegt. Damit haben wir noch keine direkte Aussage über Herrn Vorsichtig getroffen, sondern nur über das Verfahren und seine Zuverlässigkeit gesprochen. Um zu einer Aussage über Herrn Vorsichtig zu gelangen, müssen wir zu subjektiven oder epistemischen Wahrscheinlichkeiten übergehen und außerdem noch den sogenannten *statistischen Syllogismus* anwenden, der eigentlich ein wichtiges Grundprinzip der sogenannten *induktiven Logik* darstellt, die wiederum von den klassischen Statistikern abgelehnt wird.

Übrigens sollte Herr Vorsichtig, um nun größere Sicherheit über seinen Gesundheitszustand zu erhalten, weitere Tests in Auftrag geben. Auch die Wiederholung desselben Tests kann dabei gute Dienste leisten, wenn der mögliche Fehler im Falle von Herrn Vorsichtig nur zufälliger Natur war und keinen systematischen Charakter hatte, gemäß dem gerade bei Herrn Vorsichtig oder ähnlichen Personen der Test nicht zuverlässig funktioniert. Da wir das jedoch nicht so genau wissen, wäre es vermutlich hilfreicher, nun einen anderen Test zu verwenden.

5.2 Der statistische Syllogismus

Der *statistische Syllogismus* entspricht dem folgenden Schluss von einer Gesamtheit auf den Einzelfall (vgl. Campbell & Franklin 2004): Wenn in einer Urne 100 Kugeln enthalten sind und davon sind 99 weiß und nur eine schwarz, und ich ziehe eine Kugel, über die ich sonst nichts weiter weiß, dann ist die (subjektive/objektive/logische?) Wahrscheinlichkeit (vgl. Kap. 5.4) dafür, dass ich eine weiße Kugel ziehe 99%. Das erscheint uns ganz natürlich und liegt vielen Überlegungen auch in der klassischen Statistik zugrunde. Im Prinzip kommt keine Anwendung von Wahrscheinlichkeiten auf empirische Fragen um den Einsatz eines solchen Prinzips herum (vgl. Franklin 2001). Welchen anderen Wert sollte man an dieser Stelle auch sonst wählen?

Oder sollte man etwa behaupten, die Wahrscheinlichkeitsrechnung sei hier nicht zuständig, bzw. es gäbe nur die trivialen Wahrscheinlichkeiten 0 und 1 in diesem Fall. Man könnte sagen, dass hier ein bestimmtes epistemisches Prinzip im Hintergrund zur Anwendung kommt, das als eine schwache epistemische Variante des Satzes vom unzureichenden Grunde beschrieben werden kann, wonach wir ohne zureichende Gründe keine epistemischen Unterschiede machen sollten. Man könnte das Prinzip ungefähr so formulieren:

> **Epistemisches Gleichbehandlungsprinzip (EG)**: Wenn in unserem Überzeugungssystem zwei Aussagen A und B gleich stark begründet sind (oder sogar dieselben Gründe für A wie für B sprechen), so sollten wir A und B epistemisch gleich behandeln (bzw. bewerten).

Das Prinzip klingt recht plausibel und eher harmlos. Wir werden im Verlauf des Kapitel noch darauf zurückkommen. Die Anwendung auf unseren Fall können wir dann so beschreiben: Wir wissen, dass von 100 Kugeln im Normalfall 99 weiß und nur eine schwarz ist. Da wir keine weiteren Informationen darüber haben, welche der 100 Kugeln gezogen wird, können wir auf je zwei Kugeln, das epistemische Gleichbehandlungsprinzip anwenden und erhalten damit, dass wir allen Kugeln dieselbe epistemische Wahrscheinlichkeit zuweisen sollten, gezogen zu

werden. Oder wir können auch sagen: Jeder Grund, der dafür spricht, dass die eine schwarze Kugel gezogen wird, spricht ebenso dafür, dass eine der 99 weißen Kugeln gezogen wird und umgekehrt. Also müssen wir schließlich jeden der 100 Fälle als epistemisch gleichwertig behandeln. Wenn wir epistemische Wahrscheinlichkeiten vergeben, kann das nur bedeuten, dass wir alle Fälle als gleichwahrscheinlich behandeln nach dem Gleichbehandlungsprinzip. Dann werden wir aber in 99 von hundert Fällen eine weiße Kugel ziehen. Also sollte unsere (epistemische oder logische) Wahrscheinlichkeit dafür, eine weiße Kugel zu ziehen, gerade 99% sein.

Nehmen wir an, wir wüssten, dass die Kugeln durch einen *fairen Zufallsprozess* gezogen würden, bei dem jede Kugel dieselbe Wahrscheinlichkeit von 1/100 hätte, gezogen zu werden. Dann sollten Bayesianer und klassische Statistiker sofort zustimmen, dass sich damit das beschriebene Resultat einstellt, wonach wir in 99% aller Fälle eine weiße Kugel erwarten dürfen und nur in 1% eine schwarze. Da wir das in unseren Beispielen nichts über einen bestimmten Auswahlprozess wissen, aber insbesondere auch keine Indizien dafür haben, dass eine der Kugeln wahrscheinlicher gezogen wird als eine andere, besagt unser Gleichbehandlungsprinzip, dass wir dann zumindest als epistemische Wahrscheinlichkeit keine Kugel bevorzugen dürfen. Sonst würden wir nicht unsere epistemische Neutralität wahren, sondern würden stattdessen eine bestimmte Voreingenommenheit an den Tag legen. Wir wüssten ansonsten auch nicht, welche Kugeln wir bevorzugen sollten. Also sollten wir unsere Erwartungen an einer Gleichbehandlung orientieren. Das heißt zumindest, dass unsere subjektive oder epistemische Wahrscheinlichkeit für eine weiße Kugel im Einzelfall gerade 99% sein sollte.

Das ist eine Form oder eine Anwendung des *Indifferenzprinzips*, nach der wir alle Kugeln (bzw. alle Möglichkeiten) als gleichwahrscheinlich einstufen sollten, solange wir keine Gründe für eine gegenteilige Annahme haben. Wir werden später sehen, dass ein solches Indifferenzprinzip zwar als sehr plausibel erscheint, und wir auch tatsächlich darauf angewiesen sind, es aber keineswegs unproblematisch ist (vgl. Franklin 2001). Es bleibt zunächst noch die Frage offen, ob wir nur eine *subjektiv epistemische* Lesart finden können oder ob wir nicht sogar eine *objektiv epistemische* Lesart unterschreiben sollten, wonach die

99% eine Art von logischer Wahrscheinlichkeit darstellen. Tatsächlich wird unser Schluss auf die Aussage »Die nächste gezogene Kugel ist mit Wahrscheinlichkeit 99% eine weiße Kugel« durch objektive Tatsachen untermauert, nämlich die tatsächlichen Anzahlen der jeweiligen Kugeln in der Urne. Allgemeiner lässt sich der statistische Syllogismus wie folgt formulieren:

Statistischer Syllogismus: Ist die relative Häufigkeit von Fs in der Grundgesamtheit G gerade r und wir betrachten ein a aus G, von dem wir keinen Grund zu der Annahme haben, dass seine Auswahl speziell mit dem Haben der Eigenschaft F verknüpft ist bzw. bei dem wir nichts weiter darüber wissen, ob es F ist, dann sollte unsere (subjektive/logische) Wahrscheinlichkeit P(Fa) = r sein.

Die Form des statistischen Syllogismus:
$P_{\text{epistemisch}}(Fa|h(F/G) = r \,\&\, G(a) \,\&\, \neg W(Fa)) = r$

Dabei soll h(F/G) die relative Häufigkeit der Fs in G angeben und ¬W(Fa) soll heißen, dass wir kein weiteres Wissen über das Vorliegen von Fa haben. Dieser Schluss scheint zumindest keinen rein subjektiven Charakter zu haben, sondern ist ebenfalls in objektiver Weise gerechtfertigt. Er ist ein grundlegendes Prinzip der induktiven Logik. Die besagt, dass der statistische Syllogismus uns die objektive Wahrscheinlichkeit angibt, mit der unser Hintergrundwissen in einem solchen Fall die Vermutung Fa stützt, nämlich im Grade r. Die subjektiven Bayesianer sind darauf allerdings zunächst nicht zwingend festgelegt, sondern könnten stattdessen genauso einen beliebigen anderen subjektiven Glaubensgrad wählen. Allerdings sieht man schon in dem Beispiel, dass sie damit die Anbindung an die Empirie verlieren würden und das ihre Position im Normalfall kaum plausibler macht. Später werden wir noch sehen, dass Kritiker der rein subjektiven Lehre wie Patrick Maher dafür argumentieren, dass wir zu stärkeren und objektiveren Interpretationen unserer Wahrscheinlichkeiten ganz im Sinne der induktiven Logik greifen sollten.

Betrachten wir noch ein weiteres Beispiel: Nehmen wir an, wir wüssten über Joe nur, dass er Banker ist, und wüssten außerdem, dass 80% aller

Banker reich sind. Wie groß sollte dann unser Glaubensgrad dafür sein, dass auch Joe reich ist? Der statistische Syllogismus sieht das so:

P(Joe ist reich| Joe ist Banker & 80% der Banker sind reich) = 80%

Das epistemische Gleichbehandlungsprinzip sagt uns dazu, dass wir für zwei beliebige Banker X und Y, über die wir sonst nichts weiter wissen, die Aussagen (A) »X ist reich« und (B) »Y ist reich« epistemisch gleichbehandeln müssen. Das heißt, wenn wir ihnen epistemische Wahrscheinlichkeiten zuweisen, dann sollten die jeweils gleich sein. Das heißt weiterhin aber auch, dass wir für alle Banker Z dieselbe Wahrscheinlichkeit dafür zu wählen haben, dass sie reich sind. Insgesamt ist aber nur bei 80% der Banker Reichtum zu erwarten. Also sollten wir für jeden einzelnen nicht etwa die Wahrscheinlichkeit 1 wählen bzw. »Z ist reich« ableiten (das wird manchmal als statistischer Syllogismus bezeichnet s.u.), sondern bei der entsprechenden Quote von 80% bleiben, denn nur dann ist auch vernünftigerweise zu erwarten, dass 80% der Banker insgesamt reich sind. Vom epistemischen Gleichbehandlungsprinzip findet sich also ein naheliegender Weg zum statistischen Syllogismus.

Anwendungen des statistischen Syllogismus. Der statistische Syllogismus kommt in unserem Dschungelfieber-Beispiel gleich mehrfach und in gewissen Variationen zur Anwendung. Zunächst geben wir Herrn Vorsichtig die Wahrscheinlichkeit 1/1000 mit Dschungelfieber infiziert zu sein, als er in die Praxis kommt, auf der Grundlage der relativen Häufigkeit der Krankheit bei Tropenurlaubern. Das wäre natürlich anders, wenn er bereits wegen bestimmter Symptome von Dschungelfieber zum Arzt gegangen wäre. Es trieb ihn aber nur die allgemeine Vorsicht und sein Tropenurlaub an und sonst nichts. Dann dürfen wir nach dem statistischen Syllogismus davon ausgehen, dass die Prävalenz von 1/1000 gerade die Wahrscheinlichkeit ist, mit der wir davon ausgehen sollten, dass er, ohne weitere Symptome aufzuweisen, am Dschungelfieber erkrankt ist. Allerdings steckt im Hintergrund genau genommen bereits ein statistischer Schluss hinter unseren Daten, den wir aber hier zunächst einmal akzeptieren. Schließlich haben wir die Prävalenz vermutlich nicht erhalten, indem für alle Heimkehrer aus den Tropen endgültig ermittelt wurde, ob sie das Dschungelfieber aufwiesen (zumindest die

Zukünftigen sind definitiv noch nicht dabei), sondern wir haben diesen Anteil bereits anhand von Stichproben *geschätzt*. Nur für einen kleinen Teil der Heimkehrer aus den Tropen wurde diese Quote erhoben. Und die Stichprobe war vermutlich noch nicht einmal repräsentativ, denn schließlich gehen nur die besonders vorsichtigen Zeitgenossen zum Arzt und die verhalten sich wahrscheinlich auch in den Tropen besonders vorsichtig.

Ebenso ist dann 3,2% die richtige epistemische Wahrscheinlichkeit *nach positivem Testergebnis* für das Vorliegen von Dschungelfieber, weil wir sagen können, dass in der Grundgesamtheit der positiv Getesteten im Durchschnitt gerade 3,2% tatsächlich das Dschungelfieber aufweisen. Somit finden wir in der Anwendung der Sensitivität und der Spezifität des Tests T auf die entsprechenden Werte bei Herrn Vorsichtig wiederum Anwendungen des statistischen Syllogismus. In der Grundgesamtheit der Dschungelfiebrigen weisen eben 99% ein positives Testergebnis auf. Ohne weitere Hinweise darauf, dass Herr Vorsichtig auf diesen Test außergewöhnlich reagiert, sind das gemäß dem statistischen Syllogismus unsere besten Wahrscheinlichkeitswerte für Herrn Vorsichtig. Wir benötigen den statistischen Syllogismus immer wieder, wenn wir von Aussagen über relative Häufigkeiten oder solchen über allgemeine Wahrscheinlichkeiten auf die Wahrscheinlichkeiten im Einzelfall schließen möchten, und die allgemeinen Wahrscheinlichkeiten haben meistens nur dann eine Bedeutung für uns, wenn wir damit eine Aussage ebenfalls über den Einzelfall treffen können. Sie müssen uns etwas darüber sagen, in welchem Ausmaß wir bestimmte konkrete Sachverhalte zu *erwarten* haben. Wenn wir etwa wissen, dass die Aufgabe des Rauchens im Allgemeinen die Lungenkrebswahrscheinlichkeit um einen erheblichen Betrag senkt, so können wir daraus nur dann eine Empfehlung für den einzelnen Bürger gewinnen, wenn die Information auch eine Bedeutung für den einzelnen Bürger besitzt. Er muss rationalerweise erwarten dürfen, dass seine Aussicht einen Lungenkrebs zu bekommen durch die Aufgabe des Rauchens sinkt, sonst bietet die ganze Information keine rationale Motivation für ihn, mit dem Rauchen aufzuhören.

Ohne den statistischen Syllogismus wäre der Arzt bzw. ein Bayesianer schon für die Vorher-Wahrscheinlichkeit auf eine rein subjektive Einschätzung angewiesen. Seine Erwartungen im Einzelfall wären eher

willkürlich, weil sie sich nicht mehr an den Daten orientieren würden. Um den subjektiven Anteil zu reduzieren und solche Beispiele in sinnvoller Weise behandeln zu können, sollte sich der Bayesianer also auf den statistischen Syllogismus festlegen. Er gibt am besten wieder, was uns die bisherigen Daten über ein bestimmtes Resultat besagen und ist unabhängig von speziellen Vorlieben der epistemischen Subjekte.

Diese Art des Schließens in unserem Beispiel hat offensichtlich praktische Bedeutung, nicht nur im Einzelfall, sondern auch für die Gesundheitspolitik. Wenn wir einmal von ca. 40 000 HIV-Infizierten in Deutschland ausgehen, so ergäbe sich etwa für den gängigsten und recht genauen HIV-Test, den ELISA-Test, mit einer Sensitivität von 99,9% und einer Spezifität von 99,8%, dass bei einer flächendeckenden Reihenuntersuchung aller (etwa 80 Millionen) Deutschen ca. 160 000 Deutsche fälschlicherweise HIV-positiv getestet würden gegenüber ca. 40 000 tatsächlich Infizierten. Das belegt schon, wie fragwürdig die Ergebnisse wären. Von den dann ca. 200 000 positiv Getesteten wäre nur 1/5 tatsächlich infiziert und 4/5 würden umsonst alarmiert und in Schrecken versetzt. Das zeigt wiederum die große Bedeutung der Prävalenz, die vor allem der Bayesianer ernst nimmt, die aber in der klassischen Statistik etwa bei Signifikanztests leider keine entsprechende Berücksichtigung findet.

Der statistische Syllogismus hat für unser statistisches Schließen ebenfalls in anderen Anwendungen eine enorme Bedeutung. Er ist nicht nur die Grundlage, um von Grundgesamtheiten auf Stichproben zu schließen, sondern genauso unsere Grundlage, um von Stichproben auf Grundgesamtheiten zu schließen (vgl. Campbell & Franklin 2004). Nehmen wir z.B. an, ich ziehe aus einer großen Grundgesamtheit G mit $h_G(F) = r$ viele Stichproben von je 1000 Elementen, und ich nehme dabei im Sinne des statistischen Syllogismus an, dass jede der möglichen Stichproben dieselbe Wahrscheinlichkeit hat, gezogen zu werden, dann kann man berechnen, dass die meisten dieser Stichproben eine relative Häufigkeit der Fs nahe bei r aufweisen werden; was wir letztlich für einen Rückschluss von der Stichprobe auf die Grundgesamtheit G benötigen.

Betrachten wir etwa den Raum aller derartigen Stichproben S_{1000}, dann ist eine Mehrheit dieser Stichproben *gleichartig* bzw. *ähnlich* bezüglich der relativen Häufigkeit von Fs in der Stichprobe zu der relativen

Häufigkeit von Fs in G. Das bedeutet, dass die relative Häufigkeit der Fs in den *meisten* Stichproben nahe bei r liegt. Diese Gleichartigkeit ist aber eine symmetrische Beziehung. Für diese Stichproben ist auch G wiederum gleichartig zu unserer Stichprobe. Das bedeutet dann, dass ich ebenso von einzelnen solchen Stichproben approximativ korrekt auf $h_G(F)$ schließen kann, weil der statistische Syllogismus besagt, dass ich zumindest mit hoher (epistemischer) Wahrscheinlichkeit erwarten darf, dass eine einzelne Stichprobe, für die ich keine besonderen Gründe zu der Annahme habe, dass sie besonders viele oder wenige Fs aufweist, in Bezug auf den F-Anteil der Grundgesamtheit G ähnelt. Diese Voraussetzungen sind typischerweise für Zufallsstichproben erfüllt, so dass ich dann den statistischen Syllogismus anwenden darf.

Speziell für Punktschätzungen oder Intervallschätzungen etwa in Form von sogenannten *Konfidenzintervallen* nutzen wir genau diese Umkehrung aus (vgl. zu Schätzungen Kap. 6.6). Um hier aber etwa das Konfidenzintervall auch im *Einzelfall* entsprechend interpretieren zu können, sind wir wieder auf den statistischen Syllogismus angewiesen. Das sieht dann so aus, dass wir annehmen dürfen, alle Stichproben der Stichprobenmenge S_{1000} hätten dieselbe Chance gezogen zu werden, wenn keine Gründe dafür vorliegen, dass bei der Stichprobenauswahl eine Verzerrung stattfindet, die ganz bestimmte Stichproben bevorzugt. Haben wir dann eine spezielle Stichprobe vor uns, bei der die relative Häufigkeit von Fs gerade q ist, so kommt der statistische Syllogismus ins Spiel und besagt: Wenn 95% aller Stichproben in einem Abstand d um die relative Häufigkeit r herum zu finden sind (also $q \in [r–d,r+d]$), dann ist der wahre Wert r in 95% aller Fälle nicht weiter als d von der relativen Häufigkeit s in unserer Stichprobe entfernt; d.h. er liegt mit einer epistemischen Wahrscheinlichkeit von 95% im Intervall KI = [q–d,q+d]. Für diese Stichprobe dürfen wir also berechtigterweise annehmen, dass wir mit KI in unserer Stichprobe die relative Häufigkeit der Fs in der Grundgesamtheit korrekt schätzen können.

Betrachten wir das kurz an einem einfachen Zahlenbeispiel. Nehmen wir an, unsere Grundgesamtheit G wäre sehr groß gegenüber unserer Stichprobe $s \in S_{1000}$, so dass wir eine Binomialverteilung für unsere Stichproben annehmen dürfen. (Sonst müssten wir mit der hypergeometrischen Verteilung arbeiten, die berücksichtigt, dass sich mit jedem

gezogenen Element der Stichprobe die relative Häufigkeit von F in G ein klein wenig ändert, was nur ein wenig umständlicher ist, aber in der Sache keinen Unterschied bedeutet.) Außerdem sei in G $h_G(F)$ = r = 0,3. Dann liegen die relativen Häufigkeiten von über 95% aller Stichproben aus S_{1000} in dem Intervall [0,27;0,33]. Betrachten wir die als *gleichartig* zu G bezüglich F, so haben wir einen »95%-igen Grund« zu der Annahme, dass wir mit s eine gleichartige Stichprobe zu G erzielen werden. Das besagt jedenfalls der statistische Syllogismus. Entsprechend können wir mit etwas größerem Aufwand genauso in der anderen Richtung schließen und das *Konfidenzintervall* KI(s) etwa zu einem Stichprobenergebnis s mit $h_s(F)$ = q = 0,3 bestimmen. Da können wir etwa KI(s) = [0,27;0,33] zum Konfidenzniveau 95% festlegen und dürfen so mit Hilfe des statistischen Syllogismus behaupten, dass wir uns zu 95% sicher sein können, dass der wahre Wert $h_G(F)$ sich in KI(s) befindet. Damit hätten wir eine wichtige Einsicht über unsere Grundmenge G erzielt.

Ein besonders einfaches Verfahren, um ein entsprechendes Konfidenzintervall zunächst approximativ zu bestimmen, soll kurz geschildert werden: Für große Stichproben (>30) können wir diese als approximativ normalverteilt betrachten und dort gilt die *3-Sigma-Regel*, wonach im Durchschnitt im Abstand von einer Standardabweichung σ vom Mittelwert 67%, im Abstand von 2σ: 95,5% und im Abstand von 3σ vom Mittelwert 99,7% aller Werte zu finden sind. Dann können wir umgekehrt ein 2σ-Intervall um den Mittelwert der Stichprobe bilden und erhalten mit $\sigma^2 = 1000 \cdot p \cdot (1-p)$ (als Varianz der Binomialverteilung) schließlich σ = 14,5 und somit wieder in etwa das Intervall [270;330], was für die relativen Häufigkeiten unser Konfidenzintervall ergibt. Dabei wird die Standardabweichung in der Grundgesamtheit meist anhand der Standardabweichung in der Stichprobe geschätzt, was ebenfalls bereits Anwendungen des statistischen Syllogismus beinhaltet.

Leider zieht der klassische Statistiker hier nicht mit, da er den Einsatz von epistemischen Wahrscheinlichkeit – besonders im Sinne der subjektiven Wahrscheinlichkeiten – vermeiden möchte. Er wird über das Konfidenzintervall KI(s) etwa nur sagen, dass es nach einem Verfahren entwickelt wurde, das in 95% aller Fälle den gesuchten Wert überdeckt. Damit trifft er aber überhaupt keine Aussage mehr über unseren konkre-

ten Einzelfall. Was sagt nun KI(s) über unsere konkrete Grundgesamtheit G in unserem Fall? Genau genommen nichts mehr, wenn wir den statistischen Syllogismus nicht anwenden. Die Häufigkeit von 95%, mit der solche Konfidenzintervalle den wahren Wert überdecken, kann nur mit Hilfe des statistischen Syllogismus als eine epistemische Sicherheit im Einzelfall gedeutet werden. Um also mit dem Konfidenzintervall im konkreten Fall eine verständliche Behauptung verbinden zu können, sind wir wiederum auf den statistischen Syllogismus angewiesen.

Der klassische Statistiker könnte sich dabei einer logischen Deutung (im Sinne einer induktiven Logik) des statistischen Syllogismus anschließen, um die Gefahren durch subjektive Elemente in seinen Überlegungen zu vermeiden. Dass wir auf diese Art von logischen Schlüssen in den Anwendungen der Wahrscheinlichkeitsrechnung angewiesen sind und daher nicht alle Formen logischer Wahrscheinlichkeit zurückweisen sollten, dafür plädiert schon Franklin (2001) sehr überzeugend. Der statistische Syllogismus ist eine recht basale und dabei plausible Form, relative Häufigkeiten und Wahrscheinlichkeiten zusammenzubringen und sollte daher von Bayesianern wie von klassischen Statistikern gleichermaßen akzeptiert werden, obwohl beide von ihrem Grundprogramm her dazu keineswegs verpflichtet sind. Im Hintergrund steht das epistemische Gleichbehandlungsprinzip bzw. bestimmte Indifferenzprinzipien, auf die ich noch gesondert eingehen werde.

Der klassische Statistiker möchte den Einsatz epistemischer Wahrscheinlichkeiten unbedingt vermeiden, aber erst durch eine entsprechende Übersetzung von objektiven Wahrscheinlichkeiten oder relativen Häufigkeiten in subjektive Wahrscheinlichkeiten oder entsprechende Glaubensgrade gewinnen diese Informationen überhaupt eine Bedeutung für unser Handeln. Erst wenn wir die Informationen für uns als unsere Erwartung annehmen, werden sie dadurch handlungsrelevant. Diese Übernahme scheint zu einem echten Verständnis objektiver Wahrscheinlichkeiten geradezu dazuzugehören, auch wenn wir diesen Schritt nicht immer explizit durchführen. Ganz explizit findet er sich im sogenannten »principal principle« wieder, das wir gleich einführen und uns später auch noch genauer ansehen werden. Jedenfalls sollten objektive Wahrscheinlichkeiten einen Wegweiser für unser Leben darstellen.

Der subjektive Bayesianer arbeitet mit rein subjektiven Wahrscheinlichkeiten bzw. Glaubensgraden und muss sich daher nicht zwingend an relativen Häufigkeiten orientieren. Er könnte andere wählen und trotzdem in sich konsistent bleiben, aber die meisten modernen Bayesianer akzeptieren aus guten Gründen bestimmte *Wahrscheinlichkeitskoordinierungsprinzipien* wie den statistischen Syllogismus (vgl. Kap. 5.3.11) oder zumindest das sogenannte »principal principle«, das ein schwächeres Prinzip darstellt, das wir später (vgl. Kap. 5.4.4.) noch ausführlicher diskutieren werden. In der Praxis ist jedenfalls der statistische Syllogismus für den Bayesianer das wichtigste Prinzip, um von relativen Häufigkeiten zu objektiven Wahrscheinlichkeiten zu gelangen. Ohne dieses Prinzip hätte der Bayesianer keine Möglichkeit, Informationen über relative Häufigkeiten auszunutzen und bliebe auf den bloß subjektiven Wahrscheinlichkeiten sitzen, die den Bayesianern immer vorgeworfen werden. Der Bayesianer sollte daher unbedingt den statistischen Syllogismus als zusätzliches Wahrscheinlichkeitskoordinierungsprinzip akzeptieren und damit für seine Ausgangswahrscheinlichkeiten eine Anbindung an die Empirie erreichen. Zusammen mit der Likelihoodanbindung werden das die wichtigsten Objektivierungsschritte für den Bayesianismus sein. Allerdings ist er damit ein gutes Stück in Richtung der induktiven Logik gegangen, doch das ist sicher gut so.

Überhaupt scheint mir der wichtigste Vorteil der probabilistischen Ansätze etwa gegenüber anderen Ansätzen, die mit epistemischen Unsicherheiten umgehen (wie z.B. den spohnschen Rangfunktionen (vgl. Kap. 5.6.12) oder dem sogenannten AGM-Ansatz für rationalen Überzeugungswandel), der zu sein, dass die probabilistischen Ansätze es gestatten, so gut an unsere Daten anzuknüpfen und diese aufzugreifen. Dafür ist der statistische Syllogismus ein wesentliches Hilfsmittel. Wir müssen uns nur immer darüber im Klaren sein, dass wir uns damit bereits einen Teil der induktiven Logik einkaufen. Eine völlige Ablehnung der induktiven Logik kann dann nicht mehr auf unserem Programm stehen.

Das Hauptprinzip besagt im Unterschied zum statistischen Syllogismus (auch hier lassen sich allerdings unterschiedliche Formen finden, die z.T. wieder stärker dem statistischen Syllogismus ähneln), dass wir als subjektive Wahrscheinlichkeit für q in dem Fall, in dem eine objektive Wahrscheinlichkeit oder Chance für q vorliegt (ch(q) = r), diese auch als

subjektive Wahrscheinlichkeit übernehmen sollten. Wir sollten sie sogar dann übernehmen, wenn wir noch über anderslautende Informationen E zu p verfügen, die dem jedoch nicht zwingend entgegenstehen. Das ist in jedem Fall ein Unterschied zum statistischen Syllogismus, der den Schluss nur erlaubt, wenn keine derartigen Informationen vorliegen.

Hauptprinzip: (»principal principle« von David Lewis (s.u.))
P(q|ch(q) = r & E) = r, mit zulässigen Informationen E

In vielen Fällen werden wir aber nur vermuten können, welches die objektiven Wahrscheinlichkeiten wohl sein könnten und können diese Vermutung wiederum nur mit Hilfe des statistischen Syllogismus begründen, weshalb dieser normalerweise das hilfreichere Prinzip in der Praxis darstellen dürfte. Der statistischen Syllogismus verlangt keine so starken Annahmen über vorliegende Chancen oder Propensitäten (wie wir sie später nennen werden), die hier mit der Angabe ch(q) = r gemeint sind. Mit ch(q) = r wird typischerweise unterstellt, dass es sich nicht nur um eine logische Wahrscheinlichkeit handelt, sondern sogar um eine objektive physikalische Eigenschaft der Welt (eine Art von Tendenz) mit einer bestimmten Stärke r gerade q hervorzubringen. Kennen wir die Stärke dieser Eigenschaft, können uns irgendwelche relativen Häufigkeiten, die in diesem Zusammenhang auftreten, sogar vollkommen gleichgültig sein. Darauf werde ich weiter unten noch genauer eingehen.

Der klassische Statistiker wird sich in Anwendungen der Statistik sicher an bestimmten Stellen implizit auf den statistischen Syllogismus stützen, auch wenn er offiziell epistemische Wahrscheinlichkeiten (subjektive und objektive) ablehnt. Dieser Abschnitt sollte also ein Plädoyer dafür sein, dass er hier aufgeschlossener sein müsste, um seine Methoden überhaupt auf die Wirklichkeit anwenden zu können (vgl. a. Franklin 2001).

Leider gibt es für die Wissenschaftstheorie gleich noch ein Problem des statistischen Syllogismus zu vermelden. Probabilisten und speziell die Bayesianer versuchen auch für wissenschaftliche Theorien mit Glaubensgraden zu arbeiten, doch es ist z.B. nicht erkennbar, wie uns der statistische Syllogismus für diesen wichtigen Fall weiterhelfen

kann. Was sind die statistischen Daten, die uns eine Ausgangswahrscheinlichkeit für wissenschaftliche Theorien liefern könnten? Alle Ansätze, die hier denkbar sind, wirken recht phantastisch. Sollten wir uns etwa an der Quote orientieren, mit der ein bestimmter Wissenschaftler bisher wahre neue Theorien produziert hat? Das hätte neben anderen absurden Konsequenzen zur Folge, dass eine Theorie dann bereits als glaubwürdiger oder unglaubwürdiger gelten würde, je nachdem, wer sie vorgeschlagen hat. Intuitiv sollte sie dagegen besser anhand unseres bisherigen Hintergrundwissens als mehr oder minder plausibel eingestuft werden. Alles andere sind höchstens psychologische Effekte, die etwa bestimmten »Big-Shots« auch eine besondere epistemische Bedeutung geben würden.

Doch neue wissenschaftliche Hypothesen werden üblicherweise nicht mit der bisherigen Erfolgsquote ihrer Autoren versehen, zumal eine solche Quote normalerweise noch nicht einmal existiert, denn wer schlägt schon so viele neue Theorien vor, von denen wir dann sogar noch wissen, ob sie wahr sind oder nicht? Mir ist jedenfalls nicht bekannt, dass man ernsthaft versucht hätte, den statistischen Syllogismus zu diesem Zweck einzusetzen. Hier sind offensichtlich die Grenzen seiner Anwendbarkeit erreicht. Möchten wir also wissenschaftlichen Hypothesen eine epistemische Wahrscheinlichkeit zuweisen, sind wir auf andere Verfahren angewiesen und können diese Wahrscheinlichkeiten nicht mehr direkt an bestimmte relative Häufigkeiten anbinden.

Da die Bezeichnungen in der Literatur leider nicht eindeutig sind, gilt es noch, ein ähnliches Schlussverfahren zu erwähnen, dass auch manchmal den Namen statistischer Syllogismus erhält. Ich werde es als starken statistischen Syllogismus bezeichnen. Es ist durchaus relevant, wenn es darum geht, aus Wahrscheinlichkeitsaussagen zu unbedingten (und nicht-modalen) Aussagen zu gelangen. Wenn die relative Häufigkeit für Fs in G sehr hoch ist und wir ein Element a aus G erhalten, über das wir sonst nichts weiter wissen, schließt man manchmal direkt darauf, dass Fa gilt.

Starker statistischer Syllogismus: Wenn gilt: h(F/G) ist sehr hoch (etwa >90%) und a ist aus G und wir verfügen über keine Gründe dafür, dass ¬Fa gilt, dann schließen wir auf Fa.

Der Schluss ist eher stärker als der einfache statistische Syllogismus. Man könnte fast sagen, dass gilt: Ist $P_{epistemisch}(Fa|h(F/G))$ sehr hoch, dann akzeptieren wir Fa. Aber das passt nicht ganz, weil es keinen so einfachen Zusammenhang zwischen epistemischen Wahrscheinlichkeiten und dem Akzeptieren von Aussagen gibt. Man könnte eher davon sprechen, dass es sich um eine Art von Schwellenwertkonzeption des Akzeptierens handelt, nach der wir eine Aussage akzeptieren, sobald ihre Wahrscheinlichkeit einen bestimmten Schwellenwert k überschreitet. Doch selbst darauf wird sich der Probabilist nicht sogleich einlassen. Doch dazu später mehr. Die obigen Überlegungen zur Begründung des statistischen Syllogismus lassen sich jedenfalls nicht einfach auf dieses stärkere Schlussverfahren übertragen, aber es bildet zumindest ebenfalls gewisse grundlegende Intuitionen dazu ab, was eine hohe epistemische Wahrscheinlichkeit (bzw. die dahinter vermutete hohe objektive Wahrscheinlichkeit) für uns bedeutet. Wird sie sehr hoch, heißt das, dass wir Fa erwarten dürfen. Das wird uns im Kapitel 5.4 bei den Interpretationen von Wahrscheinlichkeit wieder begegnen und ebenfalls bei den Debatten um den Zusammenhang zwischen Probabilisten und klassischen Erkenntnistheoretikern in Kapitel 5.3.7. Der starke statistische Syllogismus stellt somit eine wichtige Verbindung zwischen Wahrscheinlichkeitsaussagen und solchen ohne Wahrscheinlichkeiten dar, allerdings ist seine epistemische Begründung nicht so klar. Diese Zusammenhänge werden uns im weiteren immer wieder beschäftigen.

5.3 Klassische und probabilistische Überzeugungssysteme

Wir hatten nun schon des Öfteren über eine subjektive bzw. eine epistemische Wahrscheinlichkeit gesprochen, die wir benötigen, um statistische Resultate interpretieren zu können und z.B. in konkrete Handlungsanweisungen umzusetzen. Klassische Erkenntnistheoretiker und Statistiker lehnen diese Konzepte allerdings ab. Für sie gibt es nur das *Akzeptieren* (oder den *Glauben* g) von bestimmten Aussagen (g(A)) bzw. das *Ablehnen* von Aussagen (g(¬A)) und außerdem noch eine *neutrale Einstellung* ihnen gegenüber (n(A) ≡ ¬g(A) & ¬g(¬A)).

Es sollte zumindest die Möglichkeit geben, dass wir uns weder auf A noch auf ¬A festlegen müssen, wenn wir für beide Behauptungen über keine guten Gründe verfügen, sie zu akzeptieren. Dafür ist die dritte Option gedacht, nach der wir gegenüber A auch Agnostiker bleiben können. Die Grundidee des klassischen Ansatzes ist jedenfalls, dass wir eine Aussage A *vernünftigerweise* akzeptieren sollten, wenn wir über hinreichend gute Gründe für sie verfügen, aber das eben nicht tun sollten, wenn wir über keine guten Gründe für A verfügen. Wenn wir auch für ¬A keine guten Gründe haben, müssen wir A und ¬A in den neutralen Bereich einordnen, um die Grundidee aufrechtzuerhalten.

Wahrscheinlichkeiten kommen dagegen für den klassischen Erkenntnistheoretiker bestenfalls bestimmten Zufallsereignissen (bzw. den Aussagen darüber) zu, aber nicht allen Aussagen unserer Sprache. Unsere wissenschaftlichen Hypothesen fallen auf jeden Fall nicht in diese Kategorie. Für sie bleibt die zweiwertige (bzw. genaugenommen dreiwertige) oder kategorische Auffassung von Überzeugungen im Rahmen der klassischen Erkenntnistheorie verbindlich.

Diejenigen Aussagen, die wir akzeptieren oder glauben, weil wir dafür gute Gründe haben, enthalten dann die Untergruppe von Aussagen, die sogar *Wissen* darstellen. Für Wissen verfügen wir über besonders gute Begründungen, für die es keine relevanten Unterminierer und übertrumpfenden Gegengründe gibt (vgl. dazu Kap. 2). Damit können wir mit Hilfe eines Glaubensprädikats g (mit g(A) bedeutet, dass das jeweilige epistemische Subjekt A glaubt) und einem Neutralitätsprädikat n(A) :≡ ¬g(A) & ¬g(¬A) ein klassisches Überzeugungssystem wie folgt darstellen:

Die klassische Konzeption unseres Überzeugungssystems (L,g)
Alle Aussagen A einer Sprache L werden in genau eine der drei Gruppen eingeteilt: (1) g(A) oder (2) g(¬A) oder (3) n(A)

Für diese Sicht von Überzeugungen, mit einer einfachen Kategorie der akzeptierten Aussagen, spreche ich auch von *kategorischen Überzeugungen*, die eben kategorisch behauptet werden (und nicht mit den Einschränkungen einer bestimmten Wahrscheinlichkeit versehen

sind), oder ich nenne sie *zweiwertige Überzeugungen*, was sich ein Stück weit eingebürgert hat, wobei der neutrale Wert (die Möglichkeit ihnen gegenüber Agnostiker zu bleiben) meist unter den Tisch fällt. Wir könnten allerdings das Schwergewicht auf den Glauben legen und die explizite Ablehnung sowie die neutrale Einstellung unter dem Nichtglauben zusammenfassen. Allerdings unterscheiden sich die formalen Zusammenhänge für diese beiden Fälle. So gilt typischerweise (wenn wir etwa logische Konsistenz für unser Glaubenssystem verlangen), dass aus n(A) auch n(¬A) folgt, während für g(¬A) gerade ¬g(A) folgt.

Die klassische Auffassung von Überzeugungen muss natürlich nicht ganz auf den Einsatz von Wahrscheinlichkeiten verzichten. Es können etwa Aussagen A akzeptiert werden, die ihrerseits bestimmten Ereignissen eine Wahrscheinlichkeit zuordnen, wie z.B.:

A: Die Wahrscheinlichkeit an Lassa-Fieber zu sterben ist 20%.

Nur die Aussage A selbst ist entweder wahr oder falsch und sollte daher nach Ansicht des klassischen Erkenntnistheoretikers nicht mit einer Wahrscheinlichkeit qualifiziert werden, denn die kann man seiner Meinung nach nicht sinnvoll interpretieren. A ist also entweder zu akzeptieren oder abzulehnen und, solange man sich ihrer nicht sicher genug ist, sollten wir schlicht unentschieden ihr gegenüber bleiben.

Probabilisten versuchen stattdessen, eine radikal neue Konzeption für unser Überzeugungssystem einzuführen. Demnach müssen wir uns von der alten Zwei- oder Dreiteilung verabschieden und stattdessen als Glaubensgrade reelle Zahlen zwischen 0 und 1 für alle Aussagen einführen, die die Plausibilität dieser Aussagen wiedergeben. Es wird dann dafür argumentiert, dass diese Glaubensgrade den Wahrscheinlichkeitsaxiomen gehorchen müssen, wenn sie *rational* sein sollen, weshalb wir schließlich von den rationalen Glaubensgraden sagen, dass es sich um *subjektive Wahrscheinlichkeiten* handelt. Probabilisten behaupten also zwei Dinge:

Probabilismus (probabilistische Überzeugungssysteme)
(1) Für alle Aussagen einer Sprache L (inklusive ihrer logischen Kombinationen) verfügen wir über einen Glaubensgrad zwischen 0 und 1.
(2) Die *vernünftigen* Glaubensgrade gehorchen den Wahrscheinlichkeitsaxiomen.

Unser *probabilistisches Überzeugungssystem* ist dann für uns ein Paar (L,P), bestehend aus einer einfachen aussagenlogischen Sprache L und einer Wahrscheinlichkeitsfunktion P darauf.

Die klassische Wahrscheinlichkeitsrechnung. Als Wahrscheinlichkeitsaxiome wählen wir hier nur die einfachen drei Forderungen:

Wahrscheinlichkeitsaxiome: Für alle Aussagen A und B einer Sprache L nennen wir eine reellwertige Funktion P mit den folgenden Eigenschaften eine Wahrscheinlichkeit:

(1) $0 \leq P(A) \leq 1$
(2) $P(T) = 1$ für jede Tautologie T
(3) $P(A \vee B) = P(A)+P(B)$, falls A und B sich gegenseitig logisch ausschließen

Die Axiome sind einfach gestaltet, und es soll zunächst weder die σ-Additivität gefordert werden, noch verlangen wir unbedingt, dass die Menge der Aussagen, auf denen P definiert wird, eine Algebra darstellt oder Ähnliches. Ich möchte den Einsatz von Wahrscheinlichkeiten nicht durch starke Anforderungen unnötig erschweren.

Doch die Bedingung der Algebra soll zumindest kurz erläutert werden. So wird etwa oft von unserer Sprache L gefordert, dass mit A und B immer auch A&B und A∨B und ¬A in L liegen. Wir sagen dann, dass die Menge L zusammen mit diesen Operationen eine *boolesche Algebra* darstellt. Auch ich gehe im Folgenden normalerweise davon aus, dass es keine Einschränkungen gegenüber solchen logischen Operationen gibt, wäre aber bereit, sie zu im Notfall zu diskutieren, falls das in speziellen Fällen hilfreich erschiene. Man könnte damit z.B. die Flut der Glaubensgrade etwas einschränken, die wir gemäß den Probabilisten kennen müssen.

Man könnte nun sagen, dass die Wahrscheinlichkeitsverteilung P unser neues Überzeugungssystem repräsentiert, und so werde ich jedenfalls manchmal auch reden. Mit dem neuen Überzeugungssystem meine ich also die Funktion P für die Glaubensgrade zusammen mit einer bestimmten Menge von Aussagen L, auf denen P definiert ist. Man könnte weiterhin sagen, die klassischen Überzeugungssysteme stellen eine Art von Grenzfall dar, in dem den Aussagen nur die Werte 0 und 1 zugeschrieben werden und vielleicht noch die 0,5 für den neutralen Zustand. Allerdings wird eine genauere Analyse der Zusammenhänge belegen, dass es nicht so leicht ist, eine Zuordnung zwischen klassischer Betrachtungsweise und der der Probabilisten herzustellen.

In der mathematischen Wahrscheinlichkeitsrechnung arbeitet man meist mit einer etwas anderen Darstellung. Statt von Aussagen spricht man von einem Grundraum von Ereignissen Ω und den entsprechenden Mengenoperationen darauf, also Durchschnittsbildung, Vereinigung und Komplementbildung. Auch manche Philosophen oder Informatiker, die sich mit der Repräsentation von epistemischer Unsicherheit und bestimmten Plausibilitätsmaßen beschäftigen, wählen in entsprechender Weise gern eine Menge W von *möglichen Welten* oder (weniger prätentiös) von *Möglichkeiten* oder manchmal auch von *Ereignissen* und wählen als Aussagen darauf eine Algebra W von Teilmengen von W. Jede Menge $A \subseteq W$ stellt dann die Aussage dar, die genau in allen A-Welten wahr ist bzw. die gerade die A-Möglichkeiten zulässt. Im Falle des Würfels besteht die Menge W gerade aus den Möglichkeiten $\{1, 2, \ldots, 6\}$ und wir arbeiten dann etwa mit Aussagen wie »Es kommt eine gerade Zahl«, die eine Menge von Möglichkeiten bzw. ein Ereignis auszeichnen und durch die Menge $\{2, 4, 6\}$ wiedergegeben werden. Diese Darstellung ist etwas allgemeiner, weil sie u.a. auch abzählbar unendliche Durchschnitte zulässt, während wir für Aussagen normalerweise nur endliche Konjunktionen gestatten. Im Übrigen lassen sich die Redeweisen von Wahrscheinlichkeiten für Sätze und die für Ereignisse (oder Aussagen als Mengen von möglichen Welten) aber ineinander übersetzen.

Hier und im Folgenden spreche ich ebenfalls meistens über *Aussagen* und nicht über *Sätze*, werde aber die Unterscheidung nicht weiter verfolgen, sondern gehe einfach davon aus, dass es eine einfache Zuordnung zwischen Aussagen und Sätzen gibt. Genaugenommen entspricht eine

Aussage einer Äquivalenzklasse von logisch äquivalenten Sätzen. Der besseren Handhabbarkeit wegen, beziehe ich mich dabei anhand von Sätzen einer Sprache L auf die Aussagen. Stattdessen arbeiten viele Philosophen gleich mit einer endlichen Menge W von möglichen Welten als Ausgangsmenge und jede Aussage wird dann als Teilmenge davon dargestellt (s.o.). Das hat selbst im endlichen Bereich den Vorteil, dass wir auch immer über Mengenbeziehungen sprechen können, ist aber m.E. etwas unanschaulicher als der Zugang über Sätze. Letztlich entsprechen sich die beiden Zugangsweisen in nachvollziehbarer Weise. Haben wir eine Menge von logisch unabhängigen Basisaussagen entsprechen die möglichen Welten typischerweise den Vollkonjunktionen aus diesen Basisaussagen. Die Wahrscheinlichkeiten werden also strenggenommen den Aussagen bzw. den Äquivalenzklassen von Sätzen zugesprochen, aber diese Trennung werde ich nicht immer strikt durchführen.

In der mathematischen Wahrscheinlichkeitsrechnung nimmt man gerne zur einfachen Additivität eine Stetigkeitsbedingung hinzu, die für viele Beweise, in denen man zu Grenzwerten übergeht, hilfreich ist, nämlich die Sigma-Additivität. Damit ist gemeint, dass auch noch für abzählbar unendliche Mengensysteme $M_{i \in \mathbb{N}}$ von paarweise disjunkten Mengen M_i eine Additionsregel gilt:

σ-Additivität
$P(\cup_i M_i) = \sum_i P(M_i)$ für paarweise disjunkte Mengen M_i

Die einfachen Regeln entsprechen einander für beide Ansätze (mit Wahrscheinlichkeiten auf Aussagen oder auf Ereignissen definiert), aber unendliche Disjunktionen werden in den meisten Sprachen ebenfalls nicht mehr zugelassen, weshalb die σ-Additivität eher ein Spezialfall der mathematischen Herangehensweise ist, die ich nur am Rande erwähnen möchte. Mengen, die solche abzählbaren Vereinigungen oder Disjunktionen ebenfalls noch enthalten, nennt man übrigens σ*-Algebren*.

In jedem Fall benötigen wir nicht nur die Glaubensgrade und Wahrscheinlichkeiten für einfache Aussagen, sondern auch die entsprechenden *bedingten* subjektiven Wahrscheinlichkeiten bzw. Glaubensgrade. Einige Autoren halten sogar die bedingten Wahrscheinlichkeiten für basaler (s. Kap. 5.3.16), andere führen sie durch eine Definition ein:

(4) **Bedingte Wahrscheinlichkeit:** $P(A|B) := P(A\&B)/P(B)$

Mit P(A|B) ist die Wahrscheinlichkeit gemeint, die A aufweist, wenn wir schon wissen, dass B der Fall ist. Für einen fairen Würfel ist die Wahrscheinlichkeit für eine 4 gerade 1/6: P(4) = 1/6. Wissen wir über einen Wurf aber schon, dass eine gerade Zahl gekommen ist, setzt das die Wahrscheinlichkeit für eine 4 auf 1/3 herauf, weil dann nur noch drei Möglichkeiten übrigbleiben: P(4|gerade Zahl) = 1/3.

Das heißt jedenfalls, dass wir nicht nur die Wahrscheinlichkeit für A, sondern auch noch die für A&B und die für A&B&C usw. kennen müssen, um damit alle benötigten Glaubensgrade bzw. Wahrscheinlichkeiten bestimmen zu können. Für eine endliche Sprache L, deren Aussagen sich durch die aussagenlogischen Kombinationen von endlich vielen atomare Aussagen $X = \{A_1,\ldots,A_n\}$ ergeben, müssen wir daher die gemeinsame Wahrscheinlichkeitsverteilung P für alle sogenannten *Vollkonjunktionen* $\pm A_1\&\ldots\&\pm A_n$ von Aussagen aus X kennen, die sich jeweils ergeben, indem wir Kombinationen der Basisaussagen bzw. ihrer Negationen konjunktiv zusammenfügen. (Dabei bedeutet »±« also, dass wir für jede der atomaren Aussagen $A_1,\ldots,A_n$ entweder die Aussage selbst oder ihre Negation in die Konjunktion einbinden.) Diese Vollkonjunktionen spannen für uns den Raum aller möglichen Weltzustände relativ zu unserer Sprache auf. Wir erhalten als Menge $W = \{w\in L|w \equiv \pm A_1\&\ldots\&\pm A_n\}$, die Menge aller *möglichen Welten* oder Zustände, die unsere Sprache L beschreiben kann. Alle weiteren Aussagen A unserer Sprache L sind dann logisch äquivalent zu einer *Disjunktion solcher Vollkonjunktionen*: $A \equiv w_1 \vee\ldots\vee w_k$ (das ist ihre sogenannte *kanonische disjunktive Normalform*). In der Darstellung als Mengen von Möglichkeiten entspricht A somit schlicht die Menge $\{w_1,\ldots,w_k\}$.

Allerdings benötigen wir nun Wahrscheinlichkeiten für immerhin 2^n Kombinationen bei n logisch unabhängigen Basisaussagen (genauer 2^n–1, da sich die Wahrscheinlichkeiten zu eins aufaddieren müssen, aber diese kleine Ersparnis fällt hier nicht ins Gewicht). Für 10 atomare Aussagen benötigen wir schon ca. 1000 Wahrscheinlichkeiten, für 20 atomare Aussagen eine Million, für 50 atomare Aussagen bereits 10^{15} Wahrscheinlichkeiten und für 64 Basisaussagen bereits ca. 10^{19} Wahrscheinlichkeiten. Das sind enorm große Zahlen, die mit wachsendem

n sehr schnell weiter anwachsen, weshalb diese Vorgehensweise sogar Informatikern Sorgen bereitet, weil dieses exponentielle Wachstum für größere Mengen an Aussagen selbst mit dem Computer nicht mehr beherrschbar ist. Es handelt sich jedenfalls um eine recht starke Idealisierung, mit der Probabilisten an dieser Stelle leben müssen.

Im Falle von wissenschaftlichen Hypothesen und unseren Daten dazu müssen als zu berücksichtigende Aussagen sogar alle *denkbaren* Hypothesen, die wir überhaupt irgendwann einmal berücksichtigen könnten, und alle möglichen Daten erfasst werden. Außerdem müssen wir alle logischen Beziehungen kennen, also *logisch allwissend* sein. Das deutet schon an, dass hier Einiges von den epistemischen Objekten erwartet wird.

Aus den Axiomen folgen schnell weitere Rechenregeln, von denen ich einige kurz angeben möchte. So gilt für alle A, B:

(T5) $P(A) = P(B)$, falls A und B logisch äquivalent sind.
(T6) $P(\neg A) = 1 - P(A)$, für alle A.
(T7) $P(A) = 0$, falls A eine Kontradiktion ist.
(T8) $P(A) \leq P(B)$, falls A logisch B impliziert.
(T9) $P(A \vee B) = P(A) + P(B) - P(A\&B)$, für alle A und B.

(T10) $P(A\&B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$, für alle A und B.
(T10a) $P(A\&B\&C) = P(A|BC) \cdot P(B|C) \cdot P(C)$ usw. (Kettenregel)

Und mit der Definition der bedingten Wahrscheinlichkeit sowie (T10) erhalten wir das einfache *Bayessche Theorem*:
(T11) $P(B|A) = P(B)\ P(A|B)\ /\ P(A)$, für alle A und B.

Definition: A und B sind *statistisch unabhängig* gdw
(T12) $P(A|B) = P(A)$ oder $P(A\&B) = P(A) \cdot P(B)$

Und der Satz von der *totalen Wahrscheinlichkeit*:
(T13) $P(A) = P(A|B)\ P(B) + P(A|\text{non-B})\ P(\text{non-B})$
(T14) $P(A) = \sum_i P(A|B_i)\ P(B_i)$, für eine Menge von Aussagen $\{B_1,\ldots,B_n\}$, die gegenseitig inkompatibel sind und für die $B_1 \vee \ldots \vee B_n$ eine Tautologie darstellt.

Die Regel (T5) ergibt sich eigentlich schon direkt aus der Definition der Wahrscheinlichkeitsfunktion, denn die wird wie oben erläutert wurde

auf Aussagen definiert, so dass hier zwei Sätzen oder Aussagen, die logisch äquivalent sind, automatisch derselbe Wert unter P zugewiesen werden soll. Das ist ähnlich wie im klassischen Fall, in dem man meistens ebenfalls annimmt, dass logisch äquivalente Aussagen (oder Sätze) als Gegenstände der Überzeugungen gleich zu behandeln sind. Das ist natürlich selbst bereits eine gewisse Idealisierung, aber die erscheint notwendig, wenn man nicht in die Untiefen idiosynkratischer Redeweisen der einzelnen epistemischen Subjekte eintauchen möchte.

Eventuell gibt es bestimmte *probabilistische Unabhängigkeiten*, wodurch sich die Zahl der Wahrscheinlichkeiten, die ein epistemisches Subjekt kennen muss, wieder reduzieren ließe, doch das hängt natürlich von den speziellen Überzeugungssystemen ab. Sind etwa A und B probabilistisch unabhängig, können wir P(A&B) aus P(A) und P(B) nach (T12) berechnen und müssen sie nicht eigens hinzufügen. Sind A und B wenigstens noch *relativ zu C probabilistisch unabhängig*, gilt immerhin noch:

(T12a) $P(A\&B|C) = P(A|C) \cdot P(B|C)$

Diese Gleichung werden wir im Folgenden häufiger verwenden und auch sie reduziert letztlich die Anzahl der Wahrscheinlichkeiten, über die das epistemische Subjekt einzeln verfügen muss.

Die Regeln (T13) (und entsprechend auch (T14)) ergeben sich anhand des logischen Zusammenhangs $A \equiv (A\&B) \vee (A\&\text{non-}B)$ woraus folgt: $P(A) = P(A\&B) + P(A\&\text{non-}B)$ gemäß dem dritten Axiom, und dann folgt (T13), indem wir darauf (T10) anwenden.

Zunächst einmal dürfte jedenfalls klar sein, dass es sich beim Übergang zum Probabilismus um einen wirklichen Paradigmenwechsel gegenüber der klassischen Erkenntnistheorie handelt. Es wird sich weiterhin zeigen, dass es keine einfachen Übersetzungen zwischen diesen beiden Konzeptionen gibt. Versuche, eine solche vorzunehmen, wie die von Hawthorne (s.u.) führen zu höchst komplexen Theorien über den Zusammenhang zwischen Glaubensgraden und kategorischen Überzeugungen.

Rationale klassische Überzeugungssysteme. Wir können uns nun analog ein einfaches *klassisches Überzeugungssystem* folgendermaßen vorstellen: Es gibt zunächst eine endliche Basismenge von atomaren

Aussagen $X = \{A_1,\ldots,A_n\}$ und dazu bilden wir die Boolesche Algebra A von Aussagen, die sich aus den A_i durch Hinzufügen aller logischen Verknüpfungen der Aussagen aus X ergibt (A&B, A∨B, ¬A etc.), die damit unsere Sprache darstellt. Aus dieser Menge wird im klassischen Ansatz eine Überzeugungsmenge oder Glaubensmenge $G \subseteq A$ ausgesondert, die die Aussagen angibt, die wir akzeptieren. Um *rational* zu bleiben, sollte die Menge zumindest *konsistent* sein, d.h., es gibt keine Aussagen p, so dass p und ¬p aus G ableitbar sind. Idealisierend wird dann meistens noch angenommen, dass die Menge G *deduktiv abgeschlossen* ist. Wir sollten, um rational zu sein, im Prinzip bereit sein, die logischen Konsequenzen unserer Überzeugungen ebenfalls zu akzeptieren. Die Menge G stellt aus Sicht des Subjekts auch seine (vermutliche) Wissensmenge dar, auf die es sich in seinen Entscheidungen stützt, die es aber auch heranzieht, um neue Hypothesen und Informationen zu bewerten und neue Schlussfolgerungen zu ziehen. Dazu sind wir auf die *deduktive Abgeschlossenheit* mehr oder weniger angewiesen, denn sonst müssten wir uns bei jedem Schluss aufs Neue fragen, ob der durch unser Hintergrundwissen in irgendeinem Sinne abgedeckt ist oder eben nicht. Hier kommt wieder die plausible spezielle Konsequenzbedingung von Hempel zur Anwendung, die uns schon in Kapitel 3 begegnet ist.

Wenn wir etwa glauben, dass gekränkte Eitelkeit ein typisches Mordmotiv ist, und dann erfahren, dass Franz von Lisa schwer gekränkt wurde, so sollten wir schließen dürfen, dass Franz damit ein Mordmotiv hatte. Dürften wir so nicht schließen, könnten wir unser Wissen nicht mehr sinnvoll einsetzen. Obwohl die Forderung nach deduktiver Abgeschlossenheit zugleich eine recht starke weitere Idealisierung wie die nach *logischer Allwissenheit* beinhaltet, möchte ich an dieser Stelle den klassischen Ansatz ähnlich stark ausstatten wie den probabilistischen. Außerdem liegt gerade hier ein möglicher Vorteil des klassischen Ansatzes. Mit nur wenigen grundlegenden Annahmen sind schnell weitere Schlussfolgerungen möglich, die die Grundlage für Entscheidungen und Planungen bieten. Damit ist natürlich kein Dogmatismus verbunden. Erhalten wir neue Informationen oder entdecken Fehler in unseren bisherigen Überlegungen, sind selbstverständlich jederzeit Korrekturen möglich, und wir können bisher akzeptierte Annahmen wieder aufgeben.

Um den Vergleich mit dem probabilistischen Ansatz weiter zu befördern, können wir das klassische Überzeugungssystem nun in ähnlicher Form wie das probabilistische darstellen: ($\mathscr{A}$,B) mit einer Bewertungsfunktion B:$\mathscr{A}\rightarrow\{0; 0{,}5; 1\}$ auf einer Algebra von Aussagen $\mathscr{A}$, wobei die Aussagen der Glaubensmenge G gerade auf die 1 abgebildet werden und die neutral bewerteten Aussagen auf die 0,5. Die *Rationalitätsanforderungen* können wir dann zunächst so formulieren:

($\mathscr{A}$,B) ist ein ***rationales klassisches Überzeugungssystem***, wenn für alle Aussagen p, q $\in \mathscr{A}$ gilt:
(1) *Deduktive Abgeschlossenheit*: Wenn q aus p deduktiv folgt und B(p) = 1 ist, dann ist auch B(q) = 1.
(2) *Konsistenz*: Aus B(p) = 1 folgt B(¬p) = 0.

Dabei werden alle Tautologien der Glaubensmenge zugeordnet und alle Kontradiktionen anhand der Bedingung (2) auf die 0 abgebildet. Das geht etwas über die Konsistenzforderung hinaus, es erscheint aber natürlich nur sinnvoll, den Kontradiktionen niemals den neutralen Status zuzuweisen. Wir können dann weiter (idealisierend) festlegen, dass für andere Fälle für zwei Aussagen p, q $\in \mathscr{A}$ sich für die Konjunktion ergibt: B(p&q) = min{B(p), B(q)} und für die Disjunktion B(p∨q) = max{B(p), B(q)}. Das sind zumindest plausible Ergänzungen, wenn wir für das Akzeptieren von Aussagen verlangen, dass wir über gute Gründe für sie verfügen. Das liefert uns auch wieder die Regel: B(¬p) = 1 – B(p), ganz analog zum probabilistischen Fall, weil wir etwa für die neutral bewerteten Aussagen ihre Negationen ebenfalls mit 0,5 bewerten. Das führt dazu, dass wir im klassischen Fall nur die atomaren Aussagen zu bewerten haben – jedenfalls um die Menge G zu bestimmen, die uns vor allem interessiert –, weil sich die Bewertung der komplexeren Aussagen daraus nach den genannten Regeln ergibt. Das erkennen wir für eine Aussage A an ihrer Darstellung als *disjunktive Normalform* $A \equiv w_1 \vee \ldots \vee w_k$, wonach A genau dann zu G gehört, wenn mindestens ein w_j akzeptiert wird, d.h., wenn alle Basisaussagen bzw. ihre Negationen, die in einem der Disjunkte w_j auftreten, akzeptiert werden. Das wird aber bereits durch die Funktion B auf den n Basisaussagen festgelegt. Der

Aufwand für 50 elementare Aussagen besteht dann nur darin, für diese 50 Fälle eine Bewertung festzulegen, und nicht für alle 2^{50} Aussagen wie im probabilistischen Fall.

Zur Dynamik unserer Überzeugungssysteme. Ein weiterer grundlegender Unterschied zwischen probabilistischen und klassischen Ansätzen besteht darin, dass vor allem der Bayesianismus sich ganz auf die *Übergänge* zwischen zwei aufeinanderfolgenden Überzeugungssystemen konzentriert. Für sie liefert er uns eine zentrale Regel (die Konditionalisierungsregel), während für die momentane Ausgestaltung der Systeme nur die relativ schwachen Wahrscheinlichkeitsaxiome als Kohärenzforderungen dienen. Die probabilistischen Ansätze sind also vor allem *dynamische Erkenntnistheorien*, die Regeln für die Dynamik unserer Überzeugungssysteme aufstellen und nicht primär für ihre Statik. Sie erlauben als Startwerte sehr unterschiedliche und sogar intuitiv unplausible Wahrscheinlichkeitswerte, die erst durch wiederholtes »Updaten« mit aktuellen Daten einen objektiveren Charakter gewinnen.

Die rationalen Veränderungen unseres Überzeugungssystems durch neue Daten können wir natürlich auch im klassischen Rahmen so beschreiben, dass bestimmte Beobachtungsaussagen a neu in die Glaubensmenge G := {p|B(p) = 1} aufgenommen werden, bzw. B(a) = 1 gesetzt wird, und dadurch bestimmte weitere Anpassungen von G bzw. B erforderlich werden, die dann eine klassische dynamische Erkenntnistheorie beschreiben kann. Die Dynamik unseres Überzeugungssystems ist in dem Fall durch Übergänge

$$(\mathscr{A},B) \rightarrow (\mathscr{A},B^{+}) \rightarrow (\mathscr{A},B^{++}) \ldots$$

gekennzeichnet, die die jeweils neuen Bewertungen beschreiben. Die klassische Erkenntnistheorie ist normalerweise allerdings nicht darauf eingestellt, spezielle Regeln für diese Änderungen zu formulieren.

Es gibt allerdings insbesondere im Rahmen des »belief revision«-Ansatzes (vgl. Hansson 2006) den Versuch, vor allem die Veränderungen an der Menge G zu beschreiben und gewisse Regeln für die Aufnahme neuer Aussagen a explizit und sogar axiomatisch präzise zu formulieren. Im sogenannten AGM-Modell werden sowohl Regeln für die einfache Aufnahme bzw. Erweiterung von G um eine Aussage a (expansion) sowie

für eine Anpassung (revision) angegeben, bei der a mit den bisherigen Überzeugungen in Konflikt steht und daher bestimmte Aussagen aus G (möglichst schonend) wieder entfernt werden müssen (contraction).

Dieser Ansatz ist noch am ehesten vergleichbar mit der probabilistischen Vorgehensweise, bei der unser Überzeugungssystem ganz ähnlich durch ein Paar (L,P) wiedergegeben wird, nur dass es sich bei P nun um eine Funktion für Glaubensgrade handelt und die Rationalitäts- bzw. Konsistenzforderung nun besagt, dass es sich bei P um eine Wahrscheinlichkeitsfunktion handeln sollte.

(L,P) mit einer Funktion P:L→[0,1] ist ein ***rationales probabilistisches Überzeugungssystem***, wenn P die Wahrscheinlichkeitsaxiome erfüllt.

Auch hier erwarten wir wieder, dass sich vor allem die Übergänge zu einem neuen Überzeugungssystem anhand von neuen Daten als Änderungen der Funktion P beschreiben lassen:

$$(L,B) \rightarrow (L,P^{+}) \rightarrow (L,P^{++}) \ldots,$$

wobei der Bayesianismus gerade besagt, dass diese Übergänge rationalerweise durch die bereits genannte *Konditionalisierungsregel* $P^{+}(A) = P(A|E)$ bestimmt werden sollten, die angibt, wie sich alle Wahrscheinlichkeiten ändern, wenn ein neues Datum E auftritt. Dem Datum E wird dabei im neuen System die Wahrscheinlichkeit 1 gegeben: $P^{+}(E) = 1$. Gemäß der Konditionalisierungsregel legen die alten bedingten Glaubensgrade die neuen unbedingten fest, sobald entsprechende Daten vorliegen. Wir werden später noch darauf eingehen und als Alternative etwa die Konditionalisierungsregel von Jeffrey kennenlernen, für die auch andere Werte $r<1$ erlaubt sind: $P^{+}(E) = r$. Das neue Datum wird in diesen Ansätzen nicht als unbedingt sicher betrachtet.

Wieviele Festlegungen benötigen wir? Ein weiterer wichtiger Unterschied zwischen dem klassischen Ansatz und dem probabilistischen wurde bereits oben erwähnt. Im klassischen Ansatz müssen wir die Bewertungsfunktion nur für die n Basisaussagen X festlegen und können die Werte für die weiteren Aussagen aus (A,B) dann anhand der Rationalitätsforderungen erschließen. Das ist für probabilistische

Überzeugungssysteme leider nicht der Fall. Selbst wenn wir P(A) und P(B) schon kennen, legt das P(A&B) i.A. nicht fest. Wir wissen nur, dass P(A&B) irgendwo zwischen 0 und min{P(A), P(B)} (einschließlich) liegen muss. Im Falle, dass A und B nicht als probabilistisch unabhängig angenommen werden, müssen wir solche Wahrscheinlichkeiten zusätzlich festlegen und erhalten so schnell die immens hohen Anzahlen an neuen Wahrscheinlichkeiten, die wir für unser Überzeugungssystem benötigen.

Zum Beispiel im Rahmen der Fuzzy-Logik hat man es dagegen mit den sogenannten t-Normen zu tun, die den Fuzzy-Wert für die Konjunktion aus denen der beiden Konjunkte direkt berechnen. Eine wichtige t-Norm ist etwa die Wahl des Minimums, die wir im probabilistischen Fall nur dann finden, wenn die eine Aussage aus der anderen deduktiv folgt. So wird die Unsicherheit der Konjunktion auf einfache Weise aus denen der Konjunkte bestimmt und wir vermeiden die zahlenmäßige Explosion. Probabilisten wenden allerdings dagegen ein, dass man damit nicht alle entsprechenden Zusammenhänge von relativen Häufigkeiten korrekt nachzeichnen kann, weshalb wir zu den komplexeren Regeln der Wahrscheinlichkeitsrechnung greifen müssen.

Für die Festlegung der Wahrscheinlichkeitsfunktion P müssen wir – wie oben schon gesagt – zumindest alle Vollkonjunktionen der atomaren Aussagen betrachten: $\pm A_1 \& \ldots \& \pm A_n$, und für sie Wahrscheinlichkeiten festlegen. Dann lassen sich die Wahrscheinlichkeiten für alle anderen Aussagen φ unserer Sprache leicht berechnen, weil φ sich als kanonische disjunktive Normalform von Vollkonjunktionen schreiben lässt und die Vollkonjunktionen untereinander inkonsistent sind. P(φ) ergibt sich daher als die Summe der Wahrscheinlichkeiten der Vollkonjunktionen in der disjunktiven Normalform von φ.

Das können wir an einem einfachen Beispiel sicher am leichtesten verdeutlichen. Nehmen wir als atomare Aussagen nur A, B und C und als ihre Negationen a, b und c. Dann schreiben wir vereinfachend »AB« für die Konjunktion »A&B«. Dann lässt sich z.B. die Aussage B darstellen als: B ≡ ABC ∨ aBC ∨ aBc ∨ Abc, wobei die Disjunkte einander ausschließen. Daher gilt dann: P(B) = P(ABC) + P(aBC) + P(aBc) + P(ABc). Entsprechend können wir auch komplexere Aussagen in die kanonische disjunktive Normalform bringen wie etwa: (A&B) ∨ (A&C) ≡ (ABC ∨ ABc) ∨ (ABC ∨ AbC) ≡ ABC ∨ ABc ∨ AbC, wobei der Term ABC einmal

wegfällt, weil er sonst doppelt aufträte. Damit ist gezeigt, dass wir die 2^n Wahrscheinlichkeiten für die Vollkonjunktionen benötigen, aber auch nicht mehr, denn jede andere Aussage in unserem System lässt sich schreiben als Disjunktion dieser Vollkonjunktionen.

Bei 64 Basisaussagen wären es ca. 10^{19} Wahrscheinlichkeiten, die wir benötigen. Da wir ungefähr über 10^{14} Gehirnzellen verfügen, müsste also jede Zelle ungefähr zum Speichern von 10 000 Wahrscheinlichkeiten dienen. Das belegt noch einmal die Dimensionen, die dabei im Spiel sind. Auch die Aufgabe, so viele Wahrscheinlichkeiten in konsistenter Weise zu vergeben, ist deutlich schwieriger als die Vergabe konsistenter Wahrheitswerte im klassischen Fall.

Wir vergeben etwa die Wahrscheinlichkeiten P(A|B) = 0,7 und P(B|A) = 0,3 sowie P(B) = 0,7. Das sieht zunächst harmlos aus, aber wenn wir bedenken, dass P(A|B)·P(B) = P(A&B) = P(B|A)·P(A) ist, so verlangen die obigen Zuordnungen, dass P(A) größer als 1,6 sein müsste. Also war die ursprüngliche Wahrscheinlichkeitsverteilung bereits inkonsistent. Diese Konsistenz für 10^{19} Wahrscheinlichkeiten zu erhalten, ist eine Herkulesaufgabe, die als sehr starke Idealisierung erscheint und von realen Personen offensichtlich nicht zu leisten ist.

Wir können natürlich L verkleinern und müssen nicht darauf bestehen, dass es sich um eine boolesche Algebra handelt oder anhand bestimmter Kausalbeziehungen auf einige probabilistische Unabhängigkeiten schließen und etwa annehmen, dass unsere Überzeugungen ein *Bayessches Netz* bilden, das entsprechende Unabhängigkeitsbeziehungen modelliert und so die Anzahl der erforderlichen Wahrscheinlichkeiten wieder deutlich reduzieren kann (vgl. Beierle & Kern-Isberner 2008, Kap. 12). Darauf kommen wir wieder zurück. In der Praxis werden wir es oft auch mit kleineren Anwendungsgebieten und nicht sogleich mit ganzen Überzeugungssystemen zu tun haben, dann kann sich die Menge der benötigten Wahrscheinlichkeiten natürlich in Grenzen halten.

Lassen wir also zunächst einmal die Probleme der großen Anzahlen von Glaubensgraden, über die wir verfügen müssten, als eine Form von Idealisierung beiseite, und konzentrieren uns nur auf den intuitiven Unterschied zur klassischen Konzeption sowie auf die Frage nach der Begründung für diesen Paradigmenwechsel sowie auf die Frage nach den Anwendungsmöglichkeiten für beide Ansätze. Doch zunächst möchte ich

noch ein einfaches Beispielüberzeugungssystem vorführen, das sogleich belegt, wie schwierig es ist, überhaupt konsistente Wahrscheinlichkeiten für unsere Überzeugungen zu vergeben.

Ein Beispielsystem. Um die Schwierigkeiten und Möglichkeiten dafür, ein guter Bayesianer zu sein, einmal konkret aufzuzeigen, möchte ich ein sehr einfaches Beispiel-Überzeugungssystem X = {E, D, H, H'} diskutieren. Denken wir uns dazu zwei einfache Hypothesen H und H', die einander ausschließen sollen, von denen wir aber eine Hypothese für wahr halten, also P(H) = 1–P(H'). Außerdem besagen die Hypothesen, welche Wahrscheinlichkeit das Ergebnis Kopf bei einem Münzwurf hat: H=*Die Wahrscheinlichkeit von Kopf ist 0,7* und H'=*Die Wahrscheinlichkeit von Kopf ist 0,3*. Die Münze ist also in jedem Fall gefälscht, die Frage ist nur in welcher Richtung das der Fall ist. Außerdem haben wir zwei Daten E und D, mit denen ich in dieser Reihenfolge update und beide besagen, dass jeweils Kopf gefallen ist. Die neuen Wahrscheinlichkeiten nach dem Updaten werden mit P^+ und nach dem zweiten Updaten mit P^{++} beschrieben. Außerdem nehmen wir einfach bestimmte Grundwahrscheinlichkeiten als Vorgaben für unser Beispiel an, wie z.B., dass wir mit gleichen Wahrscheinlichkeiten für beide Hypothesen starten:

(1) P(H) = 0,5 = P(H')
(2) P(E|H) = P(D|H) = 0,7
(3) P(E|H') = P(D|H') = 0,3

Diese Wahrscheinlichkeiten müssen wir nun zu einem kohärenten System ergänzen, was nicht leicht ist. Dazu betrachten wir alle gemeinsamen Wahrscheinlichkeiten unserer vier Aussagen, wobei wir die Negationen jeweils durch kleine Buchstaben charakterisieren. Außerdem sind alle Wahrscheinlichkeiten gleich Null, bei denen H und H' zusammen auftreten. Dann müssen wir nur noch die folgenden Wahrscheinlichkeiten festlegen, wobei wir weitere Symmetrien zwischen E und D annehmen und anstatt H' auch einfach h schreiben können, da H und H' sich gegenseitig ausschließen (auf die Konjunktionszeichen verzichten wir zugunsten der besseren Übersichtlichkeit):

P(EDH) = xP(EDh) = t
P(EdH) = P(eDH) = yP(Edh) = P(eDh) = u
P(edH) = zP(edh) = v

Um die Größen x bis v nun bestimmen zu können, müssen wir zunächst unsere Vorgaben (1) bis (3) auswerten bzw. umsetzen:

Aus (1) folgt: x+2y+z = 0,5 = t+2u+v
Aus (2) folgt: x+y = 0,35
Aus (3) folgt: t+u = 0,15

Damit verträglich ist z.B. die folgende Verteilung:

P(EDH) = x= 0,25 P(EDh) = t = 0,05
P(EdH) = P(eDH) = y = 0,1 P(Edh) = P(eDh) = u = 0,1
P(edH) = z = 0,05 P(edh) = v = 0,25

Mit dieser Verteilung sind nun alle Wahrscheinlichkeiten in unserem System festgelegt. Wir können beispielhaft einige ausrechnen. Man bedenke dazu, dass z.B. DE ≡ DEH ∨ DEh ist etc.:

P(DE) = x+t = 0,3
P(EH) = x+y = 0,35
P(D|EH) = P(EDH) / P(EH) = x/(x+y) = 0,25/0,35 = 0,714
P(E) = x+y+t+u = P(D) = 0,5

Damit erhalten wir die upgedateten Wahrscheinlichkeiten:
$P^+(H) = P(H|E) = P(EH)/P(E) = (x+y)/(x+y+t+u) = 0,7$

Damit sinkt die Wahrscheinlichkeit für H': $P^+(H') = 0,3$
$P^+(D) = P(D|E) = P(ED) / P(E) = 0,3 / 0,5 = 0,6$
$P^+(D|H) = P(D|EH) = 0,714$ (s.o.)
$P^+(D|H') = P(D|Eh) = t / (t+u) = 0,333$
$P^+(E) = 1$ und im nächsten Schritt

$P^{++}(H) = P(H|DE) = P(EDH) / P(DE) = 0,25 / 0,3 = 0,833$
Damit sinkt die Wahrscheinlichkeit für H': $P^+(H') = 0,167$
$P^{++}(E) = P^{++}(D) = 1$

Das Problem, selbst in einem so überschaubaren Beispiel eine *konsistente* Verteilung zu finden, die unsere intuitiven Vorgaben erfüllt, belegt zugleich, wie schwierig es ist, ein guter Bayesianer zu sein. Der Leser möge einmal selbst versuchen, konsistente Wahrscheinlichkeiten für ein bestimmtes System von Aussagen zu vergeben. Dazu kommt noch, dass wir in den praktischen Anwendungen oft die entsprechenden bedingten Wahrscheinlichkeiten erst noch berechnen müssen. Das Eingangsbeispiel vom Dschungelfieber zeigt zugleich, dass wir das nicht immer auf intuitive Weise korrekt durchführen, sondern oft auf die konkrete Berechnung angewiesen sind. Damit wird deutlich, wie groß der Schritt von einem klassischen Ansatz und Überzeugungssystem zu einem probabilistischen ausfällt. Der Probabilist wird nun argumentieren, dass wir mit seinem System aber zumindest im Prinzip mehr epistemische Differenzierungen darstellen können, als mit dem »groben« klassischen System, aber der enorme Preis, den wir dafür zahlen müssen, sollte uns dabei immer gegenwärtig sein.

Es gibt natürlich auch einen leichten Weg, Wahrscheinlichkeiten für alle Vollkonjunktionen zu vergeben, indem wir etwa im Sinne eines weitgehenden Indifferenzprinzips allen Vollkonjunktionen denselben Wert (hier also 1/8) zuweisen. Doch das ist auch ein besonders langweiliges Überzeugungssystem, denn alle Überzeugungen erhalten damit den Wert 1/2 und auch alle Likelihoods der Art P(E|H) sowie P(H|E) erhalten diesen Wert. Beim Updaten ändern sich jeweils nur die Wahrscheinlichkeiten der Daten und wir finden keine induktiven Effekte für die anderen Größen. Unsere Hypothese verharrt trotz zweifachen Updatens bei Wahrscheinlichkeit 1/2. Man sieht sehr gut, wie wichtig die Ausgangswahrscheinlichkeiten sind. Sie beinhalten bereits alle inhaltlichen Zusammenhänge für zwischen den Aussagen unseres Systems, auf die wir uns in unseren induktiven Schlüssen stützen können.

5.3.1 Argumente für Glaubensgrade

Warum sollten wir nach Meinung der Probabilisten zu Glaubensgraden anstelle von kategorischem Glauben übergehen, wenn wir Erkenntnistheorie treiben möchten? Zwei Argumente dafür sollen kurz diskutiert werden. Die Diskussion um die Sinnhaftigkeit von Glaubensgraden

muss m.E. letztlich vor allem über die Erfolge bzw. Probleme der gesamten Konzeptionen im Hinblick auf die Anwendungen geführt werden. Trotzdem sollen erste intuitive Überlegungen an dieser Stelle bereits angesprochen werden (vgl. dazu auch Bartelborth 2013).

Eine Überlegung der Probabilisten ist, dass wir nicht einfach nur einen kategorischen Glauben (also ein unbedingtes Glauben bestimmter Aussagen) bzw. ein einfaches Akzeptieren von Aussagen kennen, sondern auch *intuitiv deutliche Unterschiede in unserem Vertrauen* in einzelne Überzeugungen feststellen können. Bestimmte einfache mathematische Behauptungen wie 2+2 = 4 oder logische Wahrheiten (Tautologien) erscheinen uns deutlich sicherer als praktisch alle empirischen Behauptungen; doch selbst für die setzt sich diese Einteilung weiter fort. Wir unterscheiden bei wissenschaftlichen Hypothesen, die wir inzwischen akzeptieren, zwischen ziemlich sicher wahren Theorien bis hin zu eher spekulativen Theorien. Einfache Zusammenhänge aus der Physik (Reibung erzeugt Wärme) erscheinen uns besser gestützt, als Aussagen darüber, dass ein hoher Anteil an LDL-Cholesterin im Blut zu Herzinfarkten führt.

Auch Behauptungen über den genauen Umfang des Klimawandels sind sicher zurückhaltender zu beurteilen als Theorien über die Bewegungen von Billiard-Bällen. Mit Theorien über das Higgs-Boson oder über den Aufbau von Baryonen durch Quarks gehen wir noch einen Schritt weiter ins Hypothetische und trotzdem würden viele Physiker vermutlich zustimmen, dass sie diese Theorien akzeptieren. Hier gibt es also ein Spektrum von Aussagen, die wir akzeptieren, aber ebenso ein Spektrum im Bereich der Aussagen, die wir ablehnen. Theorien über Klabautermänner erscheinen uns zu Recht als noch unwahrscheinlicher als Annahmen über die kalte Fusion.

Das kennen wir genauso im Alltag. Denken wir nur an die unterschiedlichen Quellen, aus denen wir unser Wissen beziehen. Unterschiedliche Zeitungen oder andere Medien, verschieden qualifizierte und vertrauenswürdige Gewährsleute, die etwas Bestimmtes beobachtet haben, sowie unterschiedlich gute Beobachtungssituationen, die selbst unser direktes Beobachtungswissen als unterschiedlich sicher erscheinen lassen. Jeder kann sich dazu viele Beispiele ausdenken, die das Spektrum seiner Überzeugungen beschreiben. Wir haben für diese Zwecke etwa

Beschreibungen wie »das ist ganz sicher« oder »Franz ist ein zuverlässiger Zeuge« oder »das können wir nur vermuten« etc. Leider werden die von verschiedenen Leuten recht unterschiedlich verstanden und ergeben kein klares Spektrum.

Deshalb schlägt der Probabilist vor, eine einheitliche Repräsentation für diese Unterschiede einzuführen, die dem klassischen Erkenntnistheoretiker mit seiner einfachen Dreiteilung entgehen. Die sicheren Aussagen sollen dabei den Wert 1 zugewiesen bekommen, die sicher falschen den Wert 0 und Fälle mit völliger Indifferenz den Wert 0,5. Die anderen Aussagen müssen entsprechend in den Zwischenräumen angesiedelt werden. Dazu gibt es z.T. weitere Regeln. Wir können uns z.B. an bestimmten relativen Häufigkeiten orientieren oder unterschiedliche Indifferenzprinzipien heranziehen etc. Doch dazu später mehr.

Ein anderer Grund für den Probabilisten, nun auf Glaubensgrade zu setzen, ist darin zu sehen, dass wir sonst gewisse innere Widersprüchlichkeiten akzeptieren müssten, die u.a. im *Lotterie-Paradox* sowie im *Vorwort-Paradox* zu finden sind. Das Vorwort-Paradox geht davon aus, dass wir als Autoren eines wissenschaftlichen Buches etwa der Ansicht sind, dass alle Aussagen $A_1,\ldots,A_n$ des Buches wahr sind, denn sonst würden wir sie schließlich verändern. Wir gehen jedenfalls einmal von der Situation aus, dass das Buch nicht als »schönfärberische Auftragsarbeit« für bestimmte Firmen angefertigt wird, sondern schlicht unser bestes Wissen wiedergeben soll. Also gilt $g(A_1)$ &…& $g(A_n)$. Die deduktive Abgeschlossenheit unseres Glaubens impliziert dann: $g(A_1\&\ldots\&A_n)$. Andererseits wissen wir aus zahlreichen Erfahrungen mit früheren eigenen Werken und denen von Kollegen, dass sich trotzdem noch Fehler in dem Buch finden werden. Zumindest *eine* dort getroffene Aussage wird wohl falsch sein. Also sollten wir außerdem glauben, dass eine der Aussagen im Buch falsch ist: $g(\neg(A_1\&\ldots\&A_n))$. Doch das passt nicht mehr zusammen. Unser Glaubensoperator g (bzw. unser Operator für das Akzeptieren von Aussagen) führt so in widersprüchliche Konsequenzen, die gegen die Konsistenzforderung verstoßen. Der Probabilist kann diese Situation besser darstellen und die Inkonsistenz vermeiden. Er wird sagen, dass wir zwar für jeden Satz unseres Buches eine hohe Wahrscheinlichkeit annehmen werden, aber uns seiner trotzdem nicht völlig sicher sein können. Wenn aber gilt: $P(A_1)$

$= \ldots = P(A_n) = 99\%$, so muss dass keinen Widerspruch zu der Aussage ergeben, dass $P(A_1 \& \ldots \& A_n)$ klein ist. Für hinreichend viele und viele unabhängige Aussagen ergibt sich das sogar rechnerisch aus dem neuen Ansatz. Der klassische Erkenntnistheoretiker hat keine Ausdrucksmittel, um diese etwas paradoxe Situation angemessen zu beschreiben, der Probabilist aber schon.

Ähnlich sieht es für das Lotterie-Paradox aus. Nehmen wir eine faire Lotterie mit eintausend Losen, von denen genau eines gewinnt. A_i sei nun die Aussage, dass das i-te Los eine Niete ist. Da nur eines von tausend Losen gewinnt, scheint es nur plausibel, A_i zu akzeptieren. Das gilt aber für jedes i in derselben Weise. Also sollte gelten: $g(A_1) \& \ldots \& g(A_{1000})$. Aber andererseits sind wir davon ausgegangen, dass die Lotterie fair ist, weshalb wir ebenso akzeptieren sollten, dass nicht alle Lose Nieten sind: $g(\neg(A_1 \& \ldots \& A_n))$. Auch hier kommen wir zu keiner adäquaten Beschreibung im klassischen Rahmen. Es scheint auch nicht begründbar zu sein, gegenüber A_i neutral zu bleiben, da wir schließlich gute Gründe für A_i haben. Wir könnten sie sogar noch verbessern, indem wir mit einer Million Lose arbeiten. Letztlich kommen wir nicht darum herum, A_i für so gut begründet zu halten, dass wir es akzeptieren müssen, und erhalten schließlich unsere Paradoxie. Wiederum verfügt der Probabilist über die bessere Beschreibung dieser Situation: $P(A_1) = \ldots = P(A_{1000}) = 99{,}9\%$, aber es gilt zugleich $P(A_1 \& \ldots \& A_{1000}) = 0$, denn die A_i sind nicht unabhängig voneinander. Jedes Los i, das sich als Niete entpuppt (A_i ist wahr), erhöht die Wahrscheinlichkeit für die anderen Lose keine Niete zu sein. Das kann der klassische Erkenntnistheoretiker mit seiner zweiwertigen Glaubenskonzeption nicht nachzeichnen.

5.3.2 Benötigen wir den zweiwertigen Glauben?

Benötigen wir dann überhaupt noch den kategorischen Glauben bzw. das Akzeptieren von Aussagen oder wissenschaftlichen Theorien? Viele Probabilisten scheinen nicht dieser Ansicht zu sein. Basaler und ausreichend sind ihrer Meinung nach die Glaubensgrade, und wir können somit auf eine »Rückkehr« zum kategorischen oder zweiwertigen Glauben ganz verzichten. Insbesondere spricht dafür, dass eine Schwellenwertkonzeption mit einem Schwellenwert q (z.B. mit $0{,}5 < q < 1$) des Glaubens uns

wieder in die genannten Paradoxien zurückführt, wie sich der Leser leicht überlegen kann, denn wir müssten wiederum alle entsprechenden Lose kategorisch für Nieten halten und trotzdem die Annahme, dass es alles Nieten sind, auf der kategorischen Ebene definitiv ablehnen.

Die Schwellenwertkonzeption des Glaubens
g(A) gdw. P(A) > q (oder P(A) ≥ q)

Wählen wir dabei q = 1 scheint das zu viel zu verlangen für einen begründeten Glauben. Dann wäre der gesamte Bereich zwischen 0 und 1 als neutral anzusehen, aber einer Aussage gegenüber, der wir 99% Wahrscheinlichkeit zubilligen, sollten wir unsere Einstellung wohl kaum als neutral betrachten.

Doch was schadet es schließlich, wenn es dem Probabilisten nicht gelingt, von den subjektiven Wahrscheinlichkeiten zum dreiwertigen Glauben zurückzukehren? Dazu müssen wir ermitteln, welche Anwendungen wir für die Konzeption von akzeptierten Aussagen im Auge haben und wie der Probabilist nun damit umgehen kann. Zumindest ist ein guter Teil unserer alltäglichen Erkenntnistheorie und auch die klassische Statistik im Rahmen einer Konzeption kategorischer Aussagen formuliert und funktioniert damit nicht schlecht. Insbesondere genügt es dann in unserem obigen Beispiel mit 64 Basisaussagen, 64 Einstufungen vorzunehmen und nicht 10^{19} Wahrscheinlichkeiten schätzen zu müssen. Lohnt sich also der hohe Mehraufwand? Für die Situationen des Lotterieparadoxes und des Vorwortparadoxes können wir auch im klassischen Rahmen zumindest Ad-hoc-Lösungen entwickeln. Immerhin kann der klassische Erkenntnistheoretiker ebenfalls mit objektiven Wahrscheinlichkeitsaussagen bzw. relativen Häufigkeiten operieren. Er glaubt im Lotteriefall etwa die Aussage: A ≡ Die Wahrscheinlichkeit dafür, dass ein Los eine Niete ist, beträgt 999/1000. Er darf das nur nicht in epistemische Wahrscheinlichkeiten übersetzen, die etwas darüber aussagen, dass eine bestimmte Aussage selbst wahrscheinlich ist. Daher steht ihm die Lösung des Probabilisten also nicht wirklich zur Verfügung. Was sind dann die Stärken der klassischen Position?

Akzeptierte Aussagen dienen u.a. dazu, bestimmte Dinge zu erklären bzw. zu verstehen, aber vor allem für Vorhersagen und die sich darauf

stützenden (rationalen) Entscheidungen und Handlungen. Wenn ich jemandem eine Medizin verordne, dann ist das nur sinnvoll, wenn ich Gründe zu der Vorhersage habe, dass die Nebenwirkungen erträglich sein werden und die Krankheit des Betreffenden dadurch geheilt werden wird. Überhaupt sind solche Vorhersagen, die als Grundlage für Handlungen bzw. unser Eingreifen in die Welt dienen, die wohl wichtigsten Aufgaben für induktives Schließen bzw. das Begründen unserer Behauptungen. Dabei stützen wir uns typischerweise auf die Aussagen, die wir akzeptieren. Dafür müssen wir einen Punkt festlegen, ab dem wir Aussagen akzeptieren, um sie als Grundlage unserer Vorhersagen und Handlungen einzusetzen.

Der Probabilist wird nun sagen, dass eine *bayesianische Entscheidungstheorie* statt des klassischen Ansatzes die optimalen Anknüpfungspunkte für ein probabilistisch konzipiertes Überzeugungssystem darstellt; doch es stellt sich die Frage, ob das tatsächlich der Fall ist. Betrachten wir ein Beispiel: Zu Anfang der 1980er Jahre gab es zwei hauptsächliche Theorien zu der Entstehung von Magengeschwüren: 1. Stress führt zu einer Übersäuerung des Magens und die zu Magengeschwüren, und 2. Bakterien (später Helicobacter Pylori genannt), die sich in der Magenschleimhaut eingenistet haben, führen zu einer Entzündung der Magenschleimhäute und einer Übersäuerung des Magens und so zu Magengeschwüren. Die erste Theorie empfiehlt gegen Gastritis und mögliche Magengeschwüre: *Beruflich und privat kürzertreten.* Die zweite Theorie empfiehlt dagegen eine *Antibiotika-Therapie* mit mehreren Antibiotika, da die Bakterien nur schwer auszurotten sind.

Wie sollen wir im Fall unsicheren Wissens nun vorgehen? Nehmen wir an, die Theorien würden einander ausschließen und es wäre also nur eine richtig. Außerdem sei die erste Theorie etwas schlechter gestützt und hätte nur noch eine Wahrscheinlichkeit von 0,3, während die zweite eine Wahrscheinlichkeit von 0,7 aufwiese. Es sind nun mehrere Wege denkbar, dieses Wissen zu nutzen. Nehmen wir an, wir könnten oder wollten nur eine der Therapien durchführen; sonst wäre ein einfacher Vorschlag natürlich, einfach *beides* durchzuführen. Den könnte allerdings der klassische Erkenntnistheoretiker genauso rechtfertigen. Eine weitere Möglichkeit bestünde rein theoretisch noch in der Mischung

der Therapien, aber das sieht nicht besonders hilfreich aus, zumal eine nur zu 70% durchgeführte Antibiotika-Therapie sogar gefährlich wäre.

Die Doppeltherapie lässt sich nicht immer realisieren oder ist wegen doppelter Nebenwirkungen auch nicht sinnvoll. Dazu noch ein anderes Beispiel: Wenn wir eine bemannte Rakete zum Mond senden möchten und haben dafür zwei unterschiedliche Gravitationstheorien (z.B. die Newtonsche mit $1/x^2$-Gesetz und eine mit einem $1/x^3$-Gesetz) zur Verfügung, die nun unterschiedliche Wege und Vorgehensweisen für den Flug vorschlagen, so können wir im Normalfall nur einen der beiden Wege wählen und nicht beide einschlagen. Wir sind in einem solchen Fall dann auf ein eindeutiges Entscheidungsverfahren festgelegt.

Gehen wir im Falle der Therapien nun so vor, dass wir die Theorie mit der höheren Wahrscheinlichkeit wählen (also in unserem Beispiel die Antibiotikatherapie), so sind wir genau genommen wieder zu einem kategorischen Ansatz zurückgekehrt. Wir wählen die am besten begründete Theorie und entscheiden uns so, als ob sie die allein akzeptierte Theorie wäre. Dass die Konkurrenztheorie ebenfalls noch eine Wahrscheinlichkeit von 0,3 aufweist, spielt dann keine große Rolle mehr. Dann hätten wir im Prinzip bereits einen Übergang zu einer klassischen Konzeption von Überzeugungen vollzogen – etwa im Sinne einer Schwellenwertkonzeption. Genauso würde der klassische Erkenntnistheoretiker vorgehen. Er würde sagen, lass uns die Theorie akzeptieren, für die die besseren Belege vorliegen und auf ihrer Grundlage entscheiden. Er kann genauso akzeptieren, dass es daneben noch Gründe gibt, die für die andere Theorie sprechen, aber wir müssen eben eine Gesamtabwägung vornehmen, etwa im Sinne eines umfassenden Kohärenzvergleichs.

Wenn der Probabilist hier auf einer eigenständigen Lösung besteht, müsste er uns neue Wege aufzeigen. Er könnte z.B. vorschlagen, dass wir in den betrachteten Fällen eine spezielle *gemischte Strategie* wählen. Die kann nun nicht so aussehen, dass wir die Therapien mischen – das wäre offensichtlich Unsinn –, sondern muss in einer probabilistischen Mischung der beiden Handlungsoptionen bestehen; so wie wir das aus der Spieltheorie kennen. Demnach wäre die rationalste Lösung die, zunächst mit den entsprechenden Wahrscheinlichkeiten eine der beiden Therapien auszuwählen und dieser dann zu folgen. Also wir wählen etwa

anhand eines Losverfahrens mit Wahrscheinlichkeit 0,3 die Antistresstherapie und mit Wahrscheinlichkeit 0,7 die Antibiotikatherapie.

Würde das dem Patienten (oder wenigstens dem Arzt) als rationales Verfahren einleuchten? Oder dem Raumfahrer in einem Fall mit entsprechenden Wahrscheinlichkeiten? Wohl kaum! Stellen wir uns vor, der Arzt würde uns nach einem Losverfahren nur die Antistresstherapie geben, obwohl es doch deutlich unwahrscheinlicher ist, dass sie erfolgreich ist, bzw. der Raumfahrer müsste der schlechteren Theorie folgen und den Weg wählen, der ihn mit deutlich höherem Risiko ins Verderben führt. Beide würden wohl auf dem besseren Weg bestehen. Die Auskunft des Arztes, die bessere Therapie hätte ja auch die höhere Wahrscheinlichkeit gehabt, ausgewählt zu werden, und wäre dadurch schon angemessen in unserer Entscheidung repräsentiert gewesen, müssten wir als offensichtlichen Humbug einstufen. Was nützt uns das noch, wenn das Los erst einmal auf die schlechtere Therapie gefallen ist? Sie ist jedenfalls für uns die schlechtere nach allem was wir wissen, denn die Theorie dazu ist nach unserem Kenntnisstand schließlich deutlich schlechter begründet.

In der Spieltheorie sind gemischte Strategien bereits umstritten, aber dort haben sie noch einen guten Zweck. In Fällen, in denen unser Kontrahent etwa nicht erfahren oder vorhersagen soll, was wir genau tun werden, können wir das mit Hilfe eines Losverfahrens erreichen. Gibt es z.B. drei Routen für einen Geldtransporter und man möchte nicht, dass die richtige Route den möglichen Räubern bekannt wird, dann ist es ein sinnvolles Verfahren, die Route erst direkt vor der Abfahrt auszulosen. Dann kann die richtige Route nicht durchsickern oder vorherberechnet werden. Wählen wir einfach die schnellste, so ließe sich das vorhersagen und der wichtigste Effekt wäre damit vielleicht verschenkt. Überhaupt könnten sich in jedem anderen Verfahren eher bestimmte Vorlieben des Entscheiders zeigen, die wiederum einem Kontrahenten als Anhaltspunkt dienen könnten und es ihm unter Umständen sogar erlauben würden, begründete Vorhersagen über eine bestimmte Routenwahl zu treffen.

In unseren obigen Entscheidungsbeispielen gibt es aber keinen solchen gegen uns kalkulierenden Gegner, den wir überlisten müssten, wenn wir nicht die Natur selbst als boshaft annehmen wollen. Somit entfällt der einzig triftige Grund für eine gemischte Strategie und da-

mit sind wir an dieser Nahtstelle der Entscheidung aufgrund unserer Überzeugungen doch wieder auf die klassische kategorische Position zurückgefallen. Wir müssen eine Theorie als die deutlich bessere auswählen und werden ihr gemäß entscheiden.

Was passiert aber, wenn das nicht möglich ist, weil die Theorien etwa beide gleich gut bestätigt sind bzw. jeweils eine Wahrscheinlichkeit von ca. 0,5 aufweisen? Dann allerdings haben beide Ansätze kein geeignetes Verfahren mehr vorzuschlagen. Tatsächlich könnten wir nun eine Variante auswürfeln. Wir haben es hier mit dem neutralen Zustand zu tun. Aber das ist der Sonderfall, in dem wir über keine klaren Vorteile von einer der Seiten mehr verfügen und auch hier sind wir nicht wirklich auf eine gemischte Strategie angewiesen, sondern könnten uns willkürlich für eine der Seiten entscheiden. Genau genommen kommen wir daher in diesen Beispielen in unseren Entscheidungen doch wieder zu einem dreiwertigen Glaubensbegriff zurück und der Probabilist sollte auf eine Art von Schwellenwertkonzeption dafür zurückgreifen, ab wann wir auf der Grundlage bestimmter Annahmen entscheiden sollten.

Zunächst denkt man, dazu müssten wir ein $q > 0{,}5$ als Schwellenwert festlegen und uns dann nach einer Theorie T richten, sobald diese diesen Wert überschreitet:

Schwellenwert für Glauben: $P(T) \geq q > 0{,}5$

Aber leider ist der Fall nicht so einfach. Zunächst könnte z.B. der Fall vorliegen, dass wir 7 konkurrierende Theorien haben mit folgender Verteilung:

$P(T_1) = \ldots = P(T_6) = 0{,}1$ und $P(T_7) = 0{,}4$

Nach welcher Theorie sollten wir uns in unseren Entscheidungen dann richten? Wenn wir keine weiteren Überlegungen zur Verfügung haben, scheint T_7 der klare Sieger zu sein, aber der liegt noch unterhalb von 0,5. Er ist nur eben deutlich besser als alle Konkurrenten. Haben wir also Grund zu der Annahme, dass die wahre Theorie in unserer Liste dabei ist, wäre die Wahl von T_7 durchaus rational begründet. Die vorliegende Wahrscheinlichkeitsverteilung betont eine bestimmte Theorie oder hebt sie hervor. Das auszunutzen scheint in solchen Fällen

unser bester Weg zu einer Entscheidung zu sein. Auch hier kehren wir genau genommen zu einem kategorischen Urteil zurück, sobald wir eine Entscheidung treffen möchten. Allerdings ist diese Rückkehr zu einer kategorischen Konzeption in unserem Beispiel komplexer als es die einfache Schwellenwertkonzeption darstellt. Außerdem spricht gegen die Schwellenwertkonzeption – wie schon erwähnt –, dass wir unsere Paradoxien letztlich zurückerhalten. Das löst bei Probabilisten natürlich keine Begeisterung für eine Festlegung auf diese Konzeption aus.

Nicht nur die zahlreiche Konkurrenz führt uns manchmal zu Theorien mit Wahrscheinlichkeiten unterhalb von 0,5, die wir trotzdem akzeptieren, sondern auch die *Tiefe der Theorien* (wie Popper das nannte) kann das bedingen. Vor allem in der Wissenschaft suchen nach gehaltvollen und erklärungsstarken Theorien (zumal nur die uns bei Entscheidungen wirklich helfen können) und dann stoßen wir wieder auf das Problem, dass diese keine ganz hohen Wahrscheinlichkeiten aufweisen. Wenn wir aus T etwa eine Reihe von voneinander probabilistisch unabhängigen Daten $E_1,\ldots,E_n$ ableiten können, so sollte eine Abschätzung wie die folgende gelten:

$$P(T) \leq P(E_1 \& \ldots \& E_n) = P(E_1) \cdot \ldots \cdot P(E_n)$$

Da auch die Daten selbst alle nicht sicher sind, wird somit die Theorie eine entsprechend niedrige Wahrscheinlichkeit aufweisen. Trotzdem möchten wir sie vielleicht als Entscheidungsgrundlage mit heranziehen. Also benötigen wir oft eine komplexere Entscheidungsregel als die einfache Schwellenwertkonzeption.

Das hat Probabilisten wie Mark Kaplan (2006) schon dazu geführt, keine einfachen Zusammenhänge zwischen Glaubensgraden und solchen kategorischen Entscheidungen mehr zu sehen. Aber das kann eigentlich nicht das Ziel der Probabilisten sein, denn sie sollten schließlich mit ihrer neuen Erkenntnistheorie auch Anknüpfungspunkte zu klassischen Überzeugungen und darauf basierenden praktischen Anwendungen in der Hand haben.

Der einzige Fall, in dem wir mit der bayesianischen Entscheidungstheorie einen Weg wählen können, der dem klassischen Erkenntnistheoretiker versperrt bleibt, ist der, in dem wir für alle potentiellen nützlichen

und alle schädlichen Wirkungen einer Handlung genaue (quantitative) *Nutzenwerte* angeben können, die wir mit den Wahrscheinlichkeiten und gegeneinander verrechnen können, um so zu einer Handlungsempfehlung zu gelangen. Haben die beteiligten Medikamente etwa beachtenswerte Nebenwirkungen, und auch ihre Hauptwirkungen treten nur mit bestimmten Wahrscheinlichkeiten ein, und können wir diese Wirkungen mit Nutzenwerten versehen, so können wir im Sinne der rationalen Entscheidungstheorie entsprechende Verrechnungen vornehmen, und betrachten dann den *Erwartungsnutzen* als das Entscheidungskriterium – jedenfalls dann, wenn wir den Nutzen in objektiver Weise quantifizieren können.

Viele Vertreter der klassischen Erkenntnistheorie werden hier allerdings generelle Bedenken gegen die bayesianische Entscheidungstheorie mit ihren subjektiven Schätzungen der Wahrscheinlichkeiten anmelden und eine entsprechende Entscheidungstheorie mit objektiven Wahrscheinlichkeiten bevorzugen. Außerdem werden sie einwenden, dass wir in vielen Fällen wie den obigen kaum sinnvolle gegeneinander verrechenbare Nutzenwerte angeben können.

Zunächst zur bayesianischen Entscheidungstheorie: Haben wir die Handlungsoptionen $h_1,\ldots,h_m$ und können jeweils Wahrscheinlichkeiten w_{ij} dafür vergeben, dass eine der Konsequenzen $k_1,\ldots,k_n$ auftritt, die für uns bestimmte Nutzenwerte $u(k_i)$ aufweisen, so können wir den folgenden Erwartungswert für den Nutzen u bilden:

Erwartungsnutzen: $EU(h_i) = \sum_j w_{ij}\, u(k_j)$

Die bayesianische Entscheidungstheorie rät uns dann: Wähle die Handlung h_i mit dem höchsten Wert $E(h_i)$. Hier werden alle Nutzen- bzw. Schadenswerte mit der Wahrscheinlichkeit gewichtet, dass sie auftreten werden, und diese Wahrscheinlichkeiten sollten sich aus unserem probabilistischen Entscheidungsmodell ergeben.

Der klassische Erkenntnistheoretiker kann natürlich solche Abwägungen auf etwas andere Weise ebenfalls durchführen und dafür plädieren, in dem Ausdruck auf der rechten Seite objektive Wahrscheinlichkeiten bzw. nach objektiven Regeln geschätzte objektive Wahrscheinlichkeiten einzusetzen. Weiterhin wird er nicht unbedingt die subjektiven Wahrscheinlichkeiten dadurch aufgewertet sehen, dass sie sich nun mit einem

anderen problematischen Begriff (dem des Nutzens) verrechnen lassen. Seine Ableitung aus bestimmten Präferenzen bietet ebenfalls zahlreiche Unklarheiten und neue Schwierigkeiten (vgl. etwa Nida-Rümelin & Schmidt 2000). Insbesondere müssen wir hier wieder auf die Probleme bei der Bestimmung der Glaubensgrade hinweisen. Bereits dort sind wir auf Nutzenwerte angewiesen und können eine gegenseitige Anpassung mit Nutzenwerten vornehmen (s.u.).

Die Glaubensgrade drohen damit eher zu etwas Praktischem zu werden als zu einem grundlegenden Bestandteil der Erkenntnistheorie, was ebenfalls ihren subjektiven Charakter betont. Sie beschreiben dann nicht mehr, wie stark ein bestimmtes Hintergrundwissen eine Überzeugung stützt, sondern eher, als wie nützlich wir bestimmte Überzeugungen erachten. Doch das sollte sich nach Ansicht des klassischen Erkenntnistheoretikers besser andersherum ausrechnen lassen. Wir bestimmen zunächst, für welche Überzeugungen wir gute Gründe haben, und die so ausgewählten Überzeugungen entscheiden dann, welche Konsequenzen und damit welchen Nutzen bestimmte Handlungen wohl mit sich bringen werden. Wir werden außerdem sehen, dass bereits bei der Bestimmung von Nutzenwerten und subjektiven Wahrscheinlichkeiten gegenseitige Verrechnungen möglich sind, die diese Art von subjektiven Wahrscheinlichkeiten für einen klassischen Erkenntnistheoretiker suspekt erscheinen lassen.

Auch für den klassischen Ansatz können (wie gesagt) durchaus Wahrscheinlichkeiten ins Spiel kommen, aber nur auf der objektsprachlichen Ebene der Aussagen, die wir jeweils akzeptieren oder zurückweisen. So können wir als klassische Erkenntnistheoretiker z.B. die Aussage *akzeptieren*, dass jemand mit einer bestimmten Krankheit eine Wahrscheinlichkeit von 0,3 hat, daran zu sterben, wie wir das schon im statistischen Syllogismus gesehen haben. Diese Wahrscheinlichkeit kann dann auch in eine Nutzenkalkulation eingehen, sofern man diese in solchen Fällen für sinnvoll hält, und weitere Wahrscheinlichkeiten auf der Metaebene darüber, wie wahrscheinlich diese Aussage wiederum ist, sind dann unnötig. Klassische Erkenntnistheoretiker können so ebenfalls Aussagen über objektive Häufigkeiten oder Wahrscheinlichkeiten nutzen, wie z.B. die folgenden: Etwa 60% der an Y Erkrankten sind nach Behandlung mit Medikament M genesen und in 30% der Fälle sind

schwere Nebenwirkungen aufgetreten. Dann kann eine Nutzenabwägung mit objektiven Wahrscheinlichkeiten erfolgen, die der klassische Erkenntnistheoretiker einer mit subjektiven Wahrscheinlichkeiten für die jeweils dahinterstehenden Theorien vorziehen wird. Die Debatte, welcher Ansatz für unsere Entscheidungen und Handlungen der geeignetere ist, ist zumindest nicht klar zugunsten der Probabilisten entschieden, wie diese gern behaupten. Denken wir dazu daran, wie gut kategorische Überzeugungen in unserer Alltagspsychologie verankert sind und welche Probleme die Probabilisten haben etwa mit der großen Anzahl der Glaubensgrade und den zahlreichen Paradoxien der bayesianischen Entscheidungstheorie. Beide Ansätze haben somit Licht- und Schattenseiten.

Insbesondere sind wir mit den kategorischen Überzeugungen sehr vertraut und verbinden sie leicht mit unseren Entscheidungen, während der Probabilist hier weiter zu zeigen hat, ob das mit Glaubensgraden ebenso gut oder sogar besser gelingen kann. Der Arzt wird jedenfalls nach einem Konsens verlangen, der bestimmte Behandlungsmethoden als die richtigen auszeichnet, und kann kaum etwas mit der Auskunft anfangen, wir hätten stattdessen eine komplizierte Wahrscheinlichkeitsverteilung auf unterschiedlichen Hypothesen zu dieser Krankheit. Wir alle werden vermutlich an dieser Stelle eine Form von Rückkehr zu der klassischen Überzeugungskonzeption erwarten. Der Probabilist wird noch einige Überzeugungsarbeit zu leisten haben, bis wir in diesem Punkt neue Wege bevorzugen werden. Eine weitere Frage ist außerdem, was Glaubensgrade überhaupt sind. Das Konzept erscheint zunächst als sehr anschaulich und klassische Probabilisten hielten es daher für einfach operationalisierbar. Doch diese Hoffnung erwies sich als trügerisch (s.u.).

Weisberg (2009) diskutiert auch noch andere Ansätze wie den einer friedlichen Koexistenz beider Begriffe. Doch er verweist zu Recht darauf, dass die Frager nach dem Zusammenhang beider Konzepte dadurch nicht wirklich zum Schweigen gebracht werden können. Natürlich muss es einen Zusammenhang geben, wenn beide Konzepte etwas mit dem Entscheidungsverhalten von Menschen zu tun haben und als Grundlage für weitere Inferenzen dienen sollen.

Leitgeb hat in seiner Stabilitätstheorie des Glaubens (s. etwa Leitgeb 2014) den Zusammenhang zwischen kategorischem Glauben und Glaubensgraden weiter analysiert. Betrachten wir zunächst die Glaubensmenge G in der aussagenlogischen Sprache L. Sie soll endlich, konsistent und deduktiv abgeschlossen sein. Dann können wir die Aussage $\mathbf{B}_{G,L}$ bilden, die die Konjunktion aller Aussagen aus G darstellt. Sie ist die stärkste Aussage, die wir glauben, und wegen der deduktiven Abgeschlossenheit glauben wir dann alles, was daraus folgt. Genauer gesagt gilt gemäß unseren Idealisierungen:

$\forall$ A $\in$ L gilt: A $\in$ G gdw. $\mathbf{B}_{G,L} \Rightarrow$ A.

Wenn wir nun noch eine Glaubensgradfunktion P auf L haben, können wir dazu einen Schwellenwert benennen, nämlich r = P($\mathbf{B}_{G,L}$). Tatsächlich gilt dann für alle geglaubten Aussagen, dass ihre Wahrscheinlichkeit größer oder gleich r sein muss. Das Umgekehrte muss natürlich noch nicht gelten. Dazu hat Leitgeb weitere Stabilitätsbedingungen eingeführt, die ich nicht im Detail verfolgen werde.

Jedenfalls lassen sich so zwischen einer kategorischen Glaubensfunktion und den Glaubensgraden einer Person bestimmte Zusammenhänge festmachen. Allerdings gibt es nicht einen bestimmten Schwellenwert, ab dem wir nun eine Aussage glauben *sollten*. Sondern wir können nur zu einer vorgegeben geeigneten Glaubensmenge und einer geeigneten Glaubensgradfunktion einen Zusammenhang beschreiben. Die Glaubensmenge wird dabei nicht überflüssig und die Glaubensgradfunktion liefert uns nicht von sich aus schon eine Auswahl der Aussagen, die wir akzeptieren sollten. Der genaue Wert von r ist von beiden Funktionen abhängig und wie Leitgeb selbst angibt, damit kontextabhängig. Je nachdem wie fein oder grob unsere Aussagen die Menge der möglichen Situationen aufteilen, erhalten wir jeweils andere Zusammenhänge, die möglicherweise auch nur trivialer Art sind. Ob wir auf diesem Wege im Allgemeinen von beliebigen Glaubensgradfunktionen zu sinnvollen Überzeugungen kommen oder umgekehrt, bleibt erst einmal abzuwarten.

5.3.3 Was sind Glaubensgrade?

Probabilisten versuchen vor allem deshalb das Konzept der Glaubensgrade zu operationalisieren, um es uns dadurch als unproblematisch

verkaufen zu können. Doch solche Operationalisierungen sind selbst unter Bayesianern keineswegs unumstritten. Allerdings haben sie eine lange Tradition innerhalb des Bayesianismus, weshalb wir sie in jedem Fall genauer unter die Lupe nehmen müssen. Die Grundidee besagt, dass die Glaubensgrade bestimmte Auswirkungen auf unser Verhalten haben sollten, ähnlich wie die zweiwertigen Überzeugungen. Natürlich vertrauen wir den Aussagen mehr, die einen höheren Glaubensgrad aufweisen. Die Frage bleibt dann aber, wie sich der genaue Glaubensgrad in unserem Verhalten manifestieren kann. Der Trick ist darauf zu schauen, welche *Wettquoten* auf eine Behauptung A jemand gerade noch akzeptieren würde bzw. welche er für fair hält. Die sollen dann seinen Glaubensgrad für A festlegen. Für de Finetti (1990) ergibt sich P(A) wie folgt:

Glaubensgrad: Glaubensgrad(A) = P(A) = r soll der Glaubensgrad von A sein gdw. man einen Einsatz von r Euro für eine Wette auf A für *fair* hält, die 1 Euro bei Gewinn auszahlt.
Allgemeiner ergibt sich: Wenn man gerade noch r gegen s auf A wettet, dann ist die subjektive Wahrscheinlichkeit P(A) = r/(r+s).

Das ist so gedacht, dass wir den Wettquotienten (r zu s) als *fair* betrachten, bei dem wir bereit sind, beide Seiten zu übernehmen. Das ist also genau der Betrag, bei dem der Wunsch, die Wette auf A einzugehen, in den Wunsch übergeht, auf ¬A zu setzen. Diese Stelle identifiziert so das Ausmaß, in dem wir an A glauben. Zahlen wir beispielsweise gerade noch 0,5 Euro für die Wette mit Auszahlung 1 Euro, so geben wir dem Auftreten von A genau eine 50-prozentige Chance; daher ist der Erwartungswert der Wette gerade 0,5 Euro, was wir eben noch zu bezahlen bereit wären. Halten wir dagegen 0,9 Euro für die Wette für angemessen, erscheint uns das Auftreten von A so wahrscheinlich (nämlich P(A) = 0,9), dass wir im Durchschnitt eine Auszahlung bei Wetten dieser Art von 90 Cent erwarten. Das geht aber nur, wenn diese Wette auch in 9 von 10 Fällen tatsächlich zur Auszahlung kommt. Das bedeutet aber gerade, dass A für uns eine Wahrscheinlichkeit von 0,9 besitzt. Für uns ist also ein Wettquotient fair, wenn bei entsprechenden

Wetten der Erwartungswert der Wette sich genau mit unserem Einsatz deckt, d.h., wenn wir im Durchschnitt bei solchen Wetten weder Gewinne noch Verluste zu verzeichnen haben.

Im Folgenden werde ich meist von Wetten ausgehen, die im Gewinnfall 10 Euro auszahlen. Das ist schön übersichtlich. Bin ich bereit, 6 Euro dafür zu bezahlen (also 6:4 zu wetten), so erwarte ich, dass ich die Wette im Durchschnitt in 60% der Fälle gewinne (habe also einen Glaubensgrad von 0,6 für den Gewinn der Wette). Für 10 Spiele müsste ich dann 60 Euro setzen und würde davon ungefähr 6 gewinnen und bekäme so auch 60 Euro wieder heraus – jedenfalls im Mittel. Das belegt noch einmal, was mit einer fairen Wette gemeint ist. Das ist eine Wette, bei der im Durchschnitt (bzw. im Erwartungswert) keine Seite einen Vorteil hat.

Der intuitive Zusammenhang sollte klar geworden sein. Umgekehrt ergibt sich dann der folgende Zusammenhang von Glaubensgraden und Wettquoten:

> **Faire Wettquoten**: Wenn mein Glaubensgrad für A gerade P(A) = r ist, so betrachte ich die Wettquotienten von r zu 1–r als fair und wäre demnach bereit, bis zum Betrag r auf A gegen 1–r zu setzen und oberhalb von 1–r (bzw. ab 1–r) auf ¬A zu setzen.

Auch hier ist wieder gemeint, dass nur die Zahlenverhältnisse zählen und nicht die Höhe der Einsätze. Bei P(A) = r wäre ich also genauso bereit $r \cdot 10^6$ Euro gegen $(1–r) \cdot 10^6$ Euro zu setzen.

Erste Probleme mit diesem Ansatz sind offensichtlich. Zunächst zählt dabei nur die Wettquote und nicht z.B. die Höhe der Einsätze, doch das ist eher unrealistisch. Nehmen wir an, ich glaube, dass Deutschland eine Chance von 33% hat, der nächste Fußballweltmeister zu werden. Sollte ich dann mein ganzes Vermögen oder eine Million Euro darauf wetten? Selbst wenn ich dabei über 2 Millionen gewinnen könnte, käme mir das unsinnig vor. Aber ich würde auch nicht von der anderen Seite her die 2 Millionen dagegen setzen wollen. Damit würde ich schließlich meine ganze Existenz aufs Spiel setzen. Warum sollte man eine solche Wette eingehen? Dazu muss man wohl schon ein Hasardeur sein. Welchen Glaubensgrad habe ich dann für die Aussage A: »Deutschland wird der nächste Fußballweltmeister.«? Das scheint auf diesem Weg nicht zu klären zu sein.

Beschränken wir den Einsatz daher lieber auf kleinere Beträge. Erhalten wir denn dann unseren Glaubensgrad anhand unseres Wettverhaltens? Nicht unbedingt. Wir müssen z.B. unterscheiden zwischen Personen, die Spaß am Wetten haben und solchen, die das nicht haben oder es sogar vehement ablehnen. Wer gerne wettet, ist vielleicht schneller dazu bereit, auf bestimmte Wetten einzugehen, ich bin das z.B. nicht. Warum sollte ich mein mühsam verdientes Geld beim Wetten aufs Spiel setzen, selbst wenn ich gewisse Chancen habe, etwas dabei zu verdienen? Diese Einstellung darf aber nicht schon ausschließen, dass ich über Glaubensgrade verfüge.

Spieltheoretiker diskutieren dazu noch besondere Situationen. Nehmen wir an, wir sind in einer Nachbarstadt und haben nur noch 12 Euro, benötigen aber 20 Euro für eine Rückfahrkarte und haben keine EC-Karte oder Ähnliches dabei und müssten daher einen weiten Weg nach Hause laufen. Nun wird eine Münze geworfen (die wir z.B. für fair halten), und wir können darauf setzen, dass Kopf kommt (A). Dann wären wir vermutlich bereit, unsere 12 Euro gegen 10 Euro auf Kopf zu setzen, wenn wir nur so eine Chance hätten, die erforderlichen 20 Euro zu gewinnen. Genauso wären wir aber auch bereit, unsere 12 Euro gegen 10 Euro auf Zahl (also nicht-Kopf) zu setzen. Was ist dann unsere Wettquote? (55% für Kopf und 55% für Zahl?) Wir wären in solchen Situationen sogar bereit, auch 14 Euro gegen 6 Euro zu setzen, wenn wir denn 14 Euro hätten oder sogar noch schlechtere Kurse zu akzeptieren. Das oben genannte Messverfahren scheint hier nicht mehr richtig zu funktionieren.

Um mit solchen Fällen besser umgehen zu können, wird normalerweise ein weiteres Konzept nämlich der *Nutzenbegriff* eingeführt. Es zählt für uns nicht einfach nur der Geldbetrag, sondern vielmehr der Nutzen oder Wert, den der Betrag jeweils für uns hat. Der steigt in diesem Beispiel genau bei 20 Euro stark an. Außerdem kommt hier noch eine Budgetgrenze ins Spiel. Mit Hilfe eines Nutzenbegriffs könnten wir womöglich wieder begründen, dass unser Glaubensgrad an A 0,5 sei. Das könnte man etwa aus der Symmetrie der Situation erschließen. Allerdings könnten wir auch tatsächlich der Meinung sein, dass die Münze unfair ist und z.B. nur eine Wahrscheinlichkeit von 0,4 besteht, dass Kopf kommt, und das wäre in diesem Fall unser tatsächlicher Glaubensgrad an A. Der würde sich durch das genannte

Messverfahren aber kaum ermitteln lassen, weil die Situation immer noch symmetrisch bliebe. Wir wären weiterhin bereit, auf oder gegen A ähnlich extreme Wettquoten zu akzeptieren. Außerdem liefert dann auch der Nutzenbegriff keine eindeutige Zuordnung der Glaubensgrade mehr, weil sich die Nutzenwerte der jeweils eingesetzten unterschiedlichen Geldbeträge unterscheiden sollten.

Es bleibt also hier in jedem Fall ein Problem für die Bestimmung des korrekten Glaubensgrades zurück. Dabei zeigt sich insbesondere, dass die Wettquote nun von nicht-epistemischen Aspekten mitbestimmt wird und sicher nicht direkt mit unserem Glaubensgrad zu identifizieren ist. Selbst als Messverfahren taugt die Wettquote in diesem Fall nicht mehr.

Außerdem ist der Nutzenbegriff seinerseits wieder schwer bestimmbar, so dass in einigen Ansätzen Nutzen und subjektiven Wahrscheinlichkeiten zugleich anhand unseres Verhaltens erraten werden müssen. Damit kommt eine weitere Variable ins Spiel, die nicht direkt mit unserem rein epistemischen Glaubensgrad zusammenhängt. Es zeigt sich wiederum, was wir eigentlich schon wissen sollten, nämlich dass Operationalisierungen nicht wirklich funktionieren.

Verfügen wir bereits über Glaubensgrade, dann lassen sich die Nutzenwerte noch einigermaßen gut anhand unserer Präferenzen empirisch ermitteln (vgl. Ramsey 1931). Nehmen wir an, wir hätten bestimmte Handlungen A_i, die mit bestimmten Wahrscheinlichkeiten $P(A_i)$ zu bestimmten Konsequenzen etwa k, l, m führen (z.B. in Abhängigkeit davon, welche der Situationen S_j auftritt, was eben mit der entsprechenden Unsicherheit behaftet ist). Nehmen wir außerdem an, wir könnten auf der Menge der Konsequenzen eine Präferenzrelation »<« (die u.a. vollständig und transitiv ist) ermitteln, wonach k < l besagt, dass ich die Konsequenz l der Konsequenz k gegenüber vorziehe. Nehmen wir weiterhin an, ich hätte zwei recht extreme Konsequenzen k und m ermittelt, so dass für die meisten meiner Konsequenzen l gilt: k < l < m. Dann kann ich eine Nutzenfunktion für diese Konsequenzen einführen, indem ich zunächst u(k) = 0 und u(m) = 1 wähle. Das soll meine Normierung für die Nutzenfunktion sein. (Andere Nutzenfunktionen u' mit anderen Randwerten können sich etwa durch affine Transformationen daraus ergeben: u' = au+b, mit reellen Zahlen a und b, wobei gilt a>0.) Jedenfalls kann ich dann auf folgende Weise einen Nutzenwert u(l) festlegen. Ich

biete dazu zwei Lotterien an: L_1 (man erhält in jedem Fall l) und L_2 (man erhält k mit Wahrscheinlichkeit p und m mit Wahrscheinlichkeit 1−p). Dann variiere ich p so lange, bis eine *Indifferenz* zwischen L_1 und L_2 vorliegt. Das interpretiere ich so, dass der *erwartete Nutzen* von L1 und L2 derselbe ist:

$$u(L_1) = u(l) = u(L_2) = p \cdot u(k) + (1-p) \cdot u(m) = 1-p$$

Also wähle ich einfach u(l) = 1–p und erhalte damit eine Nutzenfunktion, die tatsächlich einen Zusammenhang zwischen Nutzenwerten und subjektiven Wahrscheinlichkeiten herstellt. So versucht man manchmal, empirisch bestimmte Nutzenwerte zu ermitteln.

Doch im allgemeinen Fall, in dem weder Nutzenfunktion noch Wahrscheinlichkeiten vorliegen, müssen wir beide Funktionen *zugleich* aus den Präferenzen bestimmen. Dazu sind recht starke – und z.T. unrealistisch starke – Anforderungen an die Präferenzen auf der Menge der Handlungen A, B, … zu stellen, denn erst dann erhalten wir ein entsprechendes Repräsentationstheorem (vgl. Gintis 2009, Kap. 1). Handlungen sollen schließlich anhand ihres Erwartungsnutzens beurteilt werden:

Erwartungsnutzen: $EU(A) = \sum_i u(A,S_i)\, p(S_i)$

Dabei gibt $u(A,S_i)$ den Nutzen an, den die Handlung A hat, wenn die Situation S_i eintritt und $p(S_i)$ die subjektive Wahrscheinlichkeit dafür, dass S_i eintritt. Bei hinreichend starken Anforderungen F an die Präferenzenfunktion (u.a. Transitivität und Vollständigkeit) ergibt sich dann das Theorem:

Repräsentationstheorem: Wenn »<« die Forderungen F (wie Transitivität u.a.) erfüllt (hier nicht weiter ausgeführt), dann gibt es genau ein Paar von Funktionen (u,p), wobei p eine Wahrscheinlichkeit darstellt, so dass für alle Handlungsoptionen A und B gilt:
$A < B$ gdw. $EU(A) < EU(B)$.

Doch die rationale Entscheidungstheorie führt zu vielen neuen Problemen und Paradoxien, in denen unser Verhalten systematisch von den Vorhersagen der Theorie abweicht, die wir hier aber nicht ausführlich

behandeln können. Jedenfalls zeigen die vielen Paradoxien, dass auch diese Idee keineswegs unproblematisch ist (vgl. Gintis 2009), weil die Anforderungen F normalerweise nicht erfüllt sind. Geld war dagegen zunächst ein viel greifbareres Gut als die Nutzenwerte, das ähnlich wie die Nutzenwerte in allgemeiner Weise wertvoll ist. Es bietet von daher einen guten intuitiven Zugang zu unserem Wettverhalten. Wir wollen im Weiteren einfach davon ausgehen, dass unsere Wettsituationen eben keine solchen Besonderheiten wie in unserem letzten Beispiel aufweisen und wir daher ohne spezielle Nutzenfunktionen einfach mit den Geldwerten arbeiten dürfen.

Außerdem hat Lyle Zynda (2000) gezeigt, dass statt durch (u,p) unsere Präferenzen ebenso durch andere Funktionen (u*,p*) repräsentiert werden können, wobei wir nur die Annahme aufgeben müssen, dass p* Glaubensgrade repräsentiert, die sich *additiv* verhalten und damit nach den Wahrscheinlichkeitsaxiomen richten. Lassen wir auch andere Glaubensgrade zu und setzen nicht an dieser Stelle bereits voraus, dass sich unsere Glaubensgrade an den Wahrscheinlichkeitsaxiomen orientieren, können wir neue Funktionen für den Erwartungsnutzen EU* einführen, indem wir einfach eine beliebige nichtverschwindende Funktion $f(A,S_i)$ einführen, womit wir erhalten:

$$EU(A) = EU^*(A) = \sum_i [u(A,S_i)\cdot f(A,S_i)]\ [p(S_i)/f(A,S_i)]$$

Dann ergeben sich neue Nutzenfunktion $u^* = f\cdot u$ (s. erste eckige Klammer) und neue Glaubensgrade $p^* = p/f$ (s. zweite eckige Klammer). Lassen wir also nicht-additive Glaubensgrade zu, für die wir z.B. Vorbilder in der Fuzzy-Logik oder anderen Ansätzen finden, so geht die Eindeutigkeit verloren und damit liefert noch nicht einmal das Repräsentationstheorem mehr eindeutige subjektive Wahrscheinlichkeiten. Unser Wettverhalten lässt sich also auf unterschiedliche Weise interpretieren, und für die Wahrscheinlichkeitsaxiome als Rationalitätspostulate für unsere Glaubensgrade muss der Probabilist schließlich erst noch argumentieren.

Weiterhin bleiben die Fragen nach Glaubensgraden besonders problematisch, bei denen wir nach wissenschaftlichen Theorien fragen. Nehmen wir etwa: A ≡ *Die Quantenmechanik ist wahr*. Wie viel sollten

wir auf A setzen? Wetten auf A erscheinen schon deshalb nicht besonders sinnvoll, weil die Wahrheit von A nicht endgültig nachgewiesen werden kann (vgl. Earman 1992, Kap. 2.4). Eher schon könnte man feststellen, dass A falsch ist. Man denke hier an Poppers Überlegungen zur Falsifizierbarkeit. Da die Wette aber vermutlich niemals und jedenfalls nicht zu unseren Lebzeiten positiv für uns ausgehen kann, sollten wir lieber nicht auf A wetten. Zeigt das schon, dass wir ein Skeptiker gegenüber der Quantenmechanik sind? Wohl kaum. Gerade für unsere intendierten Anwendungsfälle in der Wissenschaft haben wir also besondere Probleme, das Operationalisierungsverfahren anzuwenden. Allgemein kann man sagen, dass Aussagen mit gemischter Quantorenstruktur (wie z.B. $\forall x \exists y\, Hxy$), wobei die Quantoren über einen potentiell unendlichen Bereich laufen, normalerweise weder verifizierbar noch falsifizierbar sind. Inwiefern ist es dann vernünftig oder auch nur sinnvoll darauf zu wetten?

Auch wenn man Wissenschaftler, die über bestimmte Theorien gesagt haben, sie seien sehr wahrscheinlich wahr, danach fragt, welche Wahrscheinlichkeit sie ihnen beimessen würden, weigern die sich meist standhaft, eine konkrete Zahl zu nennen. Nun könnte man das sicher ein wenig erleichtern, indem wir auch Intervalle zuließen oder sogar Fuzzy-Mengen von Wahrscheinlichkeiten, doch eigentlich wollten wir unsere epistemischen Unsicherheiten bereits durch die Punkt-Wahrscheinlichkeiten zum Ausdruck bringen und nicht noch auf weitere Hilfsmittel zurückgreifen. Solche zusätzlichen Ausdrucksmöglichkeiten würden unseren ganzen Kalkül jedenfalls erheblich erschweren (vgl. etwa den Dempster-Shafer-Ansatz u.a. in Beierle & Kern-Isberner 2008, Kap. 13). Bleiben wir daher bei Punkt-Wahrscheinlichkeiten als hinreichenden Idealisierungen. In jedem Fall stellen die wissenschaftlichen Theorien einen Problemfall für Glaubensgrade dar, den wir weiter im Auge behalten sollten.

Patrick Maher (2006) (früher selbst ein subjektiver Probabilist heute eher ein Vertreter der induktiven Logik) wendet gegen die Operationalisierungsversuche ein, dass beim Wetten viele artfremde Gesichtspunkte ins Spiel kommen können. Gesichtspunkte, die nicht direkt verknüpft sind mit unseren epistemischen Einstellungen, sondern eher einen praktischen Charakter haben. Möchte ich mich etwa selbst motivieren, nun mit dem Rauchen aufzuhören, setze ich vielleicht einen hohen

Betrag gegen einen kleinen, dass ich die nächsten 12 Monate clean bleiben werde. Doch das zeigt womöglich nicht, dass ich von meinem Erfolg tatsächlich sehr überzeugt bin, sondern hat vielmehr nur die Absicht, es etwas wahrscheinlicher zu machen, dass ich auch durchhalte. Der hohe Betrag ist hier vermutlich eher ein Indiz dafür, dass ich eigentlich nicht glaube, es zu schaffen und nun eine zusätzliche starke Motivation dafür benötige. Hier gehen unsere Wünsche ein in die Gestaltung der Wetten und in die Höhe der Wettbeträge.

Außerdem müssen wir noch eine weitere Unterscheidung berücksichtigen: Das Wettverhalten könnte zum einen als mehr oder weniger gutes *Messverfahren* für Glaubensgrade gedacht sein, die eigentlich nicht weiter definiert werden können, sondern einen Grundbegriff (einen theoretischen Term etwa) darstellen sollen. Dann bleibt allerdings weiter die Frage offen, was denn Glaubensgrade sind. Oder das Wettverhalten dient als echte *Operationalisierung* in dem Sinne, dass es uns tatsächlich sagt bzw. *definiert*, was ein Glaubensgrad ist. Bei de Finetti und vielen anderen Probabilisten dürfen wir wohl davon ausgehen, dass sie von einer echten Definierbarkeit ausgingen. Doch das kommt uns ganz besonders unplausibel vor, denn wir treffen hier auf die zahlreichen Einwände gegen den Operationalismus und im Speziellen gegen den logischen Behaviorismus, der versucht hat, intentionale Zustände durch Verhaltensmuster zu definieren (vgl. Eriksson & Hajek 2007).

Eriksson und Hajek (2007) haben noch einige weitere Probleme einer derartigen Operationalisierung zusammengestellt. Normalerweise erwarten wir von Messgeräten, dass ihre Messungen auch fehlerbehaftet sein können, doch das ist für den Operationalisten unmöglich, denn die Messung definiert gerade die gesuchte Größe. Wir würden eigentlich erwarten, dass die Glaubensgrade ein bestimmtes Verhalten erzeugen und hervorrufen und dabei Fehler z.B. durch Fehlschätzungen der Nutzenwerte auftreten können, doch das ist für einen Operationalisten ausgeschlossen. Speziell die Einwände gegen einen logischen Behaviorismus kommen hier ebenfalls zum Tragen. Die besagen z.B., dass Schmerz nicht einfach identifizierbar ist mit einem bestimmten Schmerzverhalten, denn Putnams Superspartaner könnte das Schmerzverhalten unterdrücken, obwohl er Schmerzen hat, oder ein guter Schauspieler kann sie vortäuschen, obwohl keine Schmerzen vorliegen. Ein Stück weit

können wir das alle, und das zeigt wiederum, dass wir Schmerzen haben nicht mit unserem Schmerzverhalten identifizieren dürfen, und ebenso wenig können wir Glaubensgrade mit einem bestimmten Wettverhalten identifizieren. Die Spieler einer Mannschaft wetten vielleicht auf ihren Gewinn, um damit ihre Loyalität der Mannschaft gegenüber und ihre Einsatzbereitschaft zu signalisieren. Das mag durchaus vernünftig sein, ist aber keineswegs ein zuverlässiger Gradmesser für ihre Glaubensgrade.

Außerdem dürfte klar sein, dass wir normalerweise bereits immer Glaubensgrade zuschreiben müssen, obwohl der Betreffende keine tatsächlichen Wetten abgeschlossen hat. Das Wettverhalten, auf das wir uns beziehen, wenn wir von Glaubensgraden sprechen, ist nur ein *hypothetisches*. Wir *würden* wohl bestimmte Wetten abschließen, wenn wir bestimmte Glaubensgrade aufwiesen. Man denke dazu noch an die extrem große Zahl von Wetten, die wir benötigen, um unsere Glaubensgrade zu manifestieren. Den Aktualisten kann das zumindest nicht zufriedenstellen. Hier werden hypothetische Glaubensgrade durch hypothetische Wetten definiert. Dazu kommen die vielen Abweichungen, die selbst unser vernünftiges Wettverhalten von unseren Glaubensgraden trennen, denn es gibt viele nicht epistemische Aspekte beim Wetten zu berücksichtigen. Das zeigen inzwischen zahllose Beispiele.

Eriksson und Hajek (2007) plädieren daher schließlich dafür, die *Glaubensgrade als Grundbegriff* zu betrachten, der sich eben nicht definieren oder anderweitig auf grundlegendere Begriffe reduzieren lässt. Unser Wettverhalten in bestimmten Situationen mag eine erste approximative Messung dafür abgeben, aber die Glaubensgrade sind nicht so einfach und intuitiv wie der Name vielleicht suggeriert. Allerdings ist dieser Schritt nicht ganz so unproblematisch, wie die beiden Autoren denken, denn die Analogien, die Eriksson und Hajek zu anderen basalen empirischen Konzepten in der Wissenschaft sehen, sind nicht unbedingt überzeugend, weil diese Konzepte vor allem dadurch gerechtfertigt sind, dass sie in *empirischen* Theorien ihre Erklärungskraft unter Beweis stellen. Doch für die subjektiven Wahrscheinlichkeiten ist nicht so klar, ob sie tatsächlich zu Erklärungszwecken taugen. Die Psychologie setzt nach vielen negativen empirischen Resultaten zumindest nicht mehr allzu große Hoffnungen in die rationale Entscheidungstheorie (oder die Wert-Erwartungstheorie, wie sie in der Psychologie genannt wird) als

erklärender Theorie (vgl. a. Gintis 2009). Sie ist wohl eher eine normative Theorie rationalen Entscheidens. Dann hinken allerdings die Analogien zu den basalen empirischen Begriffen in der Wissenschaft erheblich.

Von den genannten Problemen sind ebenso die *Dutch-Book-Argumente* betroffen, die eine zentrale Stellung für den Probabilisten einnehmen, denn sie beruhen zuallererst auf der Anknüpfung der Glaubensgrade an unser Wettverhalten. Scheitern Sie und wir verfügen nicht mehr über gute Gründe dafür, dass vernünftige Glaubensgrade den Wahrscheinlichkeitsaxiomen gehorchen müssen, lässt sich der ganze Apparat der Wahrscheinlichkeitsrechnung nicht ins Spiel bringen. Für den neuen Grundbegriff der Glaubensgrade wird letztlich also seine theoretische Fruchtbarkeit für die Erkenntnistheorie sowie seine Erklärungskraft im Rahmen von Handlungserklärungen ausschlaggebend sein. Wir werden uns hier nur mit der erkenntnistheoretischen Seite beschäftigen und müssen am Ende des Kapitels beurteilen, wie erfolgreich insbesondere der Bayesianismus in diesem Feld ist.

Zunächst müssen wir im Gedächtnis behalten, was der Probabilist annimmt. Er geht davon aus, dass wir für *alle* Überzeugungen und die Kombinationen daraus Glaubensgrade besitzen. Selbst für recht abstrakte Theorien aus der Wissenschaft nimmt er solche Glaubensgrade an. Diese Annahmen haben einen empirischen und einen normativen Aspekt und beide sind nicht immer strikt voneinander zu trennen. Der empirische besagt, dass wir unsere Überzeugungssysteme durch Glaubensgrade repräsentieren können und dadurch in die Lage versetzt werden, viele zumindest große Teile unseres Entscheidungsverhaltens zu erklären und nach Möglichkeit sogar vorherzusagen. Das bleibt sicher – trotz der zahlreichen Rückschläge – weiter ein spannendes Gebiet der empirischen Forschung, auf das ich hier aber nicht weiter eingehen kann. Der normative Aspekt besagt, dass wir mit Hilfe der Glaubensgrade besonders gut unsere epistemische Situation modellieren können. Dadurch können wir schließlich Konzepte wie Rechtfertigung und Wissen oder neue Ersatzbegriffe dafür explizieren, die helfen, unsere epistemische Situation zu erhellen, und uns Hilfen dafür geben, wann wir uns auf bestimmte Behauptungen in unseren Entscheidungen stützen dürfen bzw. stützen sollten. Mir geht es mehr um die normative Frage, was wir unter einem *vernünftigen* Entscheidungsverhalten zu

verstehen haben bzw. wie wir uns verhalten *sollten*. Das verweist auf ein weiteres Problem der Operationalisierung. Sie dient bestenfalls dazu, die tatsächlichen Glaubensgrade einer Person zu messen, aber das sind nicht unbedingt die *rationalen* Glaubensgrade, die wir ihr zuschreiben sollten, wenn wir sie rationalisieren. Auf das Problem kommen wir wieder zurück.

Außerdem trennen wir meist die kognitiven von den volitionalen Aspekten des Entscheidens. Oft nehmen wir sogar an, dass es für unsere Überzeugungen bzw. Glaubensgrade begründete Rationalitätsanforderungen gibt, während unsere Wünsche dem nicht unterliegen. Sie sind demnach eher Geschmackssache. Wenn jemand Schmerzen an Dienstagen gut und an allen anderen Tagen nicht so gut findet, obwohl wir keine physiologischen Unterschiede zwischen diesen Tagen ausmachen können, ist das demnach nicht irrational, sondern eben nur sein besonderer Geschmack. Von dieser Herangehensweise bin ich nicht wirklich überzeugt, aber das zeigt zumindest wiederum die Gefahren der Verquickung von Glaubensgraden mit Präferenzen, wenn es uns doch eigentlich um eine Konzeption *vernünftigen* induktiven Schlussfolgerns geht. Unser nächstes Thema ist nun vor allem die zweite Forderung der Probabilisten, dass Glaubensgrade den Wahrscheinlichkeitsaxiomen gehorchen sollten.

5.3.4 Dutch-Book-Argumente

Selbst wenn wir nun einmal davon ausgehen, dass Glaubensgrade mit Wettquoten zusammenhängen, ist nicht ausgemacht, ob sie den Wahrscheinlichkeitsaxiomen gehorchen. Das verlangen Probabilisten aber, damit sie ihren wahrscheinlichkeitstheoretischen Apparat zum Einsatz bringen können. Sie behaupten, dass zumindest *vernünftige* Glaubensgrade sich danach richten. Sollten sie das tun, nennen wir sie *kohärent*. Das heißt z.B., dass $P(A) = 1–P(\neg A)$ gilt, dass also die entsprechenden Glaubensgrade oder Wahrscheinlichkeiten für A und ¬A intern zusammenpassen. Die Wahrscheinlichkeitsaxiome erscheinen uns inzwischen recht natürlich, da wir den Umgang mit ihnen gewohnt sind, aber das war nicht immer so. Sie wurden erst spät entdeckt, was schon darauf hindeutet, dass sie nicht ganz so selbstverständlich

sind, wie es uns heute oft erscheint. Tatsächlich haben Psychologen wie Kahneman und Tversky (vgl. Gintis 2009) immer wieder zeigen können, dass Menschen in ihren Schlüssen von den Axiomen abweichen, und haben dazu die Prospect-Theorie entwickelt, die menschliches Entscheidungsverhalten besser beschreiben kann als die klassische rationale Entscheidungstheorie. In der Prospect-Theorie gehen die Wahrscheinlichkeiten jedoch nur noch in verzerrter Form in die Entscheidung ein. Auch die Fuzzy-Logik (vgl. Gottwald 1993) hat für die Konjunktion und Disjunktion eine ganze Reihe anderer Regeln (etwa die verschiedenen T-Normen und T-Konormen) vorgeschlagen, die auf andere Weise mit epistemischen Unsicherheiten umgehen. Warum sollten wir uns also im Sinne der Wahrscheinlichkeitsaxiome kohärent verhalten?

Probabilisten behaupten, dass wir sonst *ausbeutbar* sind, sobald wir unsere Glaubensgrade nicht kohärent gestalten, und das wäre offensichtlich irrational. Unter der Annahme, dass unsere Glaubensgrade sich in entsprechenden Wettquoten manifestieren und wir tatsächlich bereit sind, entsprechende Wetten einzugehen, lässt sich zeigen, dass wir dann schlechte Entscheidungen treffen. Diesee werden uns in Form von sogenannten Dutch-Books präsentiert. Man kann aber beweisen: *Es gibt kein Dutch-Book gegen uns genau dann, wenn unsere Glaubensgrade kohärent sind.* Dabei ist mit einem Dutch-Book ein System von Wetten gemeint, bei dem wir in jedem möglichen Fall nur verlieren können, wenn wir uns darauf einlassen. Es scheint jedenfalls offensichtlich irrational zu sein, sich auf ein Paket von Wetten einzulassen, wenn von vornherein feststeht, dass wir in jedem Fall Geld dabei verlieren werden, ganz gleich was passieren wird.

Nehmen wir ein einfaches Beispiel, um das zu illustrieren: Unser Glaubensgrad sei sowohl für A wie auch für ¬A gerade 0,8 (P(A) = P(¬A) = 0,8). Dann wären wir bereit, 8 Euro für eine Wette zu bezahlen, die 10 Euro auszahlt, falls A eintritt, und ebenso noch einmal 8 Euro für eine Wette zu bezahlen die 10 Euro auszahlt, falls ¬A eintritt.

Wette 1: 8 Euro auf A
Wette 2: 8 Euro auf ¬A

Insgesamt hätten wir also 16 Euro bezahlt, würden aber nur eine der beiden Wetten gewinnen, d.h., wir bekämen in jedem Fall genau 10 Euro ausgezahlt und blieben immer auf einem Verlust von 6 Euro sitzen. (Tritt A ein, gewinnen wir Wette 1, bei ¬A Wette 2.) Das sieht nicht sehr vernünftig aus, weil wir in jedem Fall Geld verlieren. Man kann in diesem Beispiel schön erkennen, dass wir für die zweite Wette nicht mehr als 2 Euro zahlen sollten, sonst tritt das Malheur ein. Bei einer 2-Euro-Wette wäre etwa P(¬A) = 0,2 und es wäre also P(A) = 1–P(¬A) genau erfüllt. Was würde aber passieren, wenn wir nun, um auf Nummer sicher zu gehen, nur 1 Euro für die Wette auf ¬A setzen würden? Das brächte uns wiederum in Schwierigkeiten, da der Probabilist davon ausgeht, dass wir dann ebenfalls bereit wären, die andere Seite der Wette zu übernehmen (es handelt sich ja um die für uns *fairen* Wettquoten) und wir müssten somit bereit sein, 9 Euro gegen ¬A zu setzen, für einen Pott von 10 Euro. Ebenso müssten wir bereit sein, die andere Seite von Wette 1 zu übernehmen. Damit ergäbe sich:

Wette 1: 2 Euro gegen A
Wette 2: 9 Euro gegen ¬A

Damit würden wir wieder nur eine Wette und damit 10 Euro gewinnen (im Falle, dass A eintritt, würden wir diesmal Wette 2 gewinnen), hätten dafür aber zusammen 11 Euro bezahlt. So sehen einfache Dutch-Books aus. Sie zeigen schon, dass wir auch in diesem Fall die Regel P(A) = 1–P(¬A) zu respektieren hätten, um keine Verluste zu machen.

Diese Idee können wir nun direkt auf die Axiome anwenden. Die ersten beiden Axiome sind relativ leicht nachvollziehbar. Es lohnt sich nicht, gegen Tautologien zu wetten, und es lohnt sich ebenso wenig, für eine Wette, die nur 10 Euro auszahlt, mehr als 10 Euro zu bezahlen. Etwas schwieriger ist wieder die Begründung der Additivitätsregel, von der wir eben einen Sonderfall betrachtet haben.

Nehmen wir wieder eine Reihe von Wetten, die alle im Gewinnfall 10 Euro auszahlen. Außerdem seien A und B inkonsistent. Man wähle dann P(A) = 0,4 und P(B) = 0,5 und P(A∨B) = 0,6. Dann sollten wir bereit sein, die folgenden drei Wetten zu akzeptieren:

Wette 1: 4 Euro auf A
Wette 2: 5 Euro auf B
Wette 3: 4 Euro gegen (A∨B)

Für die Wetten muss ich zusammen 13 Euro einzahlen, doch was bekomme ich heraus? Dazu gilt es drei Fälle zu betrachten: 1. A ist wahr: Dann ist B falsch und meine 3. Wette verliere ich auch. Ich gewinne also nur 10 Euro. 2. B ist wahr: Das ist analog, nur dass jetzt A falsch ist und ich auch meine 3. Wette verliere. 3. A ist falsch und B ist falsch: Dann gewinne ich endlich die dritte Wette, verliere aber die beiden anderen; also gehe ich wiederum mit einem Minus von 3 Euro nach Hause. Um dabei nicht zu verlieren, dürfte ich in der dritten Wette nur 1 Euro setzen, und das heißt im Umkehrschluss, dass mein Glaubensgrad für A∨B mindestens 0,9 sein sollte, weshalb dann gilt: P(A∨B) = P(A) + P(B). Weniger darf es aber auch nicht sein, sonst können wir wieder wie im ersten Beispiel zu den jeweiligen Gegenwetten übergehen und verlieren wieder mit Sicherheit Geld.

Das kann man natürlich etwas systematischer analysieren: Es sei P(A) = r, P(B) = s und P(A∨B) = m. Nehmen wir außerdem einmal an, nun sei m>s+r und wir wetten jeweils insgesamt um einen Euro, dann müssten wir die drei folgenden Wetten akzeptieren:

Wette 1: 1–r Euro gegen A (gegen r Euro)
Wette 2: 1–s Euro gegen B (gegen s Euro)
Wette 3: m Euro auf (A∨B) (gegen 1–m Euro)

Nun lässt sich bestimmen, welche Auszahlungen sich in den unterschiedlichen Fällen ergeben:

Situationen: Auszahlung:
1. Fall: A und B falsch: r+s–m < 0
2. Fall: A falsch, B wahr: r–(1–s)+(1–m) = r+s–m < 0
3. Fall: A wahr, B falsch:–(1–r)+s+(1–m) = r+s–m < 0

Also wird hier deutlich, dass wir in allen möglichen Fällen (man beachte, dass A und B sich gegenseitig ausschließen) immer einen Verlust zu gegenwärtigen haben. Ähnliche Probleme ergeben sich, wenn m<s+r gilt, nur dass wir nun die umgekehrten Wetten akzeptieren müssten:

Wette 1: r Euro auf A (gegen 1–r Euro)
Wette 2: s Euro auf B (gegen 1–s Euro)
Wette 3: 1–m Euro gegen (A∨B) (gegen m Euro)

Damit ergibt sich für die Auszahlungen:

Situationen: Auszahlung:
1. Fall: A und B falsch: –r–s+m < 0
2. Fall: A falsch, B wahr: –r+(1–s)–(1–m) = –r–s+m < 0
3. Fall: A wahr, B falsch: –(1–r)+s–(1–m) = –r–s+m < 0

Es lässt sich überdies die andere Richtung zeigen, so dass wir insgesamt das folgende Theorem erhalten:

Dutch-Book-Theorem: Es existiert kein Dutch-Book gegen uns gdw. unsere Glaubensgrade kohärent sind. (d.h., sich nach den Wahrscheinlichkeitsaxiomen richten)

Kaum jemand dürfte bezweifeln, dass es irrational ist, ein Dutch-Book zu akzeptieren. Kritik an der Argumentation kommt daher eher aus der Richtung, ob unser Wettverhalten im Falle bestimmter Glaubensgrade durch diese Argumentation richtig wiedergegeben wird. Der letzte Abschnitt sollte uns schon hinreichend dafür sensibilisiert haben, dass es eine deutliche Kluft zwischen Glaubensgraden und entsprechenden Wettquoten gibt. Dann ist das Argument nicht mehr so überzeugend, denn auf diesen Zusammenhang waren wir schließlich in unserem Argument angewiesen. Doch wir erkennen noch weitere Probleme in dem Argument.

Ist es denn automatisch vernünftig, ein *Paket von Wetten* zu akzeptieren, sobald es vernünftig ist, die einzelnen Wetten zu akzeptieren? Dieses »package-principle« hat schon Schick (1986) kritisiert. Das steckt aber hinter der Argumentation der Dutch-Books. Die Probabilisten behaupten nämlich, dass unsere Wetten einzeln deshalb nicht rational sein können, weil ein entsprechendes Paket von Wetten nicht rational sei. Das setzt aber genau das Paket-Prinzip voraus. Nur wenn wir zu den einzelnen Wetten diese auch noch im Paket akzeptieren sollten, dann geraten wir in die erwähnten Schwierigkeiten. Ein rationaler Entscheider könnte also genau an dieser Stelle »Halt!« rufen.

Das Paket-Prinzip: Wenn es *vernünftig* ist, bestimmte Wetten $w_1, \ldots, w_n$ jeweils für sich zu akzeptieren, so ist auch *vernünftig*, all diese Wetten zugleich als Paket zu akzeptieren.

Doch das Paket-Prinzip sieht nicht wirklich überzeugend aus. Denken wir uns z.B. folgende Situation: Wir haben eine potentiell tödliche Krankheit und verfügen über 700 Euro, benötigen zum Überleben aber ein Medikament, das 1000 Euro kostet. Stehlen können wir das Medikament nicht, da es gut gesichert ist. Dann wäre es durchaus rational, alles auf eine Karte zu setzen und die 700 Euro gegen 300 Euro auf »Kopf« bei einem Münzwurf zu setzen, selbst wenn wir die Münze für fair hielten. Genauso wäre es rational, die 700 Euro gegen 300 Euro auf »Zahl« bei diesem Münzwurf zu setzen. Doch offensichtlich wäre es völlig kontraproduktiv und irrational, beide Wetten zugleich einzugehen, auch wenn der Buchmacher gern bereit wäre, einem die 700 Euro für die zweite Wette (aber eben nur für die Wette) zu stunden, bis die Wetten zur Auszahlung kommen.

Ein pedantischer Skeptiker könnte vielleicht noch einwenden, dass die Gesamtwette zumindest keine Verschlechterung gegenüber dem Zustand darstellt, in dem wir überhaupt nicht wetten. Über 300 oder 700 Euro zu verfügen sei kein relevanter Unterschied für mein Überleben. Das würde sich allerdings auch noch ändern, wenn es eine winzig kleine Chance gäbe, ohne Wette noch 300 Euro geschenkt zu erhalten. Die wäre zu klein, um sich nicht auf die einzelnen Wetten einzulassen, aber würde zumindest noch einen Unterschied zwischen den 300 und 700 Euro Zuständen aufzeigen. (Hier sind natürlich viele Abänderungen des Beispiels möglich, um den Einwand zu entkräften.) Das Paket-Prinzip ist also in jedem Fall keine brauchbare Rationalitätsforderung, wodurch die Dutch-Book-Argumente klar unterminiert werden. Ein rationaler Wettender würde sich eben nicht auf Dutch-Books einlassen, müsste aber deswegen nicht gleich seine Glaubensgrade an die Wahrscheinlichkeitsaxiome anpassen, zumal er vielleicht auch nicht wüsste, welche seiner Glaubensgrade er dafür am besten ändern sollte (s.u.). Unser Wettverhalten kann also rational sein, weil es für die einzelnen Wetten gute Gründe gibt, während das entsprechende Paket von Wetten offensichtlich irrational ist.

Bei Arntzenius et al. (2004) findet sich noch ein hübsches Beispiel gegen das Paket-Prinzip im abzählbar additiven Fall. Gott hat einen Apfel in abzählbar unendlich viele Stücke zerlegt. Davon darf Eva endlich viele essen, aber niemals unendlich viele. Eva möchte möglichst viel vom Apfel essen. Für jedes einzelne Stück gilt dann, dass sie es zu sich nehmen sollte. Sie kann dabei ruhig immer wieder eines dazu nehmen. Trotzdem gilt nicht das Paket-Prinzip, denn sie darf nicht alle Stücke zu sich nehmen. Das wäre irrational, da sie dann von Gott bestraft würde.

Was bleibt schließlich als Fazit für die Einführung von Glaubensgraden? In jedem Fall sind *vernünftige* Glaubensgrade nicht operationalisierbar. Selbst messen können wir bestenfalls die tatsächlichen Glaubensgrade; die gehorchen aber oft nicht den Wahrscheinlichkeitsaxiomen. Insbesondere verlangen die *vernünftigen* Glaubensgrade logische Allwissenheit, die niemand leisten kann, denn Axiom 2 setzt z.B. voraus, dass wir alle logischen Tautologien identifizieren können.

Überhaupt bleibt völlig unklar, wie wir einer bestimmten Person, die nicht schon über kohärente Glaubensgrade verfügt, *vernünftige* Glaubensgrade zuschreiben sollen. Hat sie etwa zwei Glaubensgrade, die nicht zusammenpassen (etwa $P(A) = P(\neg A) = 0{,}4$), so ist nicht festgelegt, wie wir (oder sie selbst) das auflösen sollen. Welche Glaubensgrade sollte sie dann stattdessen annehmen? Soll sie etwa beide Glaubensgrade auf 0,5 anheben, oder soll sie einen auf 0,6 setzen oder beiden Aussagen ganz neue passende Werte zuweisen und nach welchen Regeln? Es wird nur die Kohärenz gefordert, aber es gibt keine Vorgaben, welche Glaubensgrade anzunehmen bzw. zu verändern sind, um diese Kohärenz zu erreichen. Das ist in einem solchen Fall kaum besonders hilfreich. Insbesondere die kohärenten Glaubensgrade sind daher für eine bestimmte Person nicht in irgendeiner Weise messbar und vermutlich noch nicht einmal seine tatsächlichen Glaubensgrade. Das Konzept der rationalen Glaubensgrade ist ein theoretisches Konzept, das durch seine Funktionen vor allem im Bayesianismus gekennzeichnet ist, das damit aber keineswegs so harmlos daherkommt, wie es zunächst klingt.

5.3.5 Wahrheitsnähe und epistemische Entscheidungstheorie

Ein anderer Weg, um die Wahrscheinlichkeitsaxiome für unsere Glaubensgrade zu rechtfertigen, wurde neuerdings u.a. von James Joyce (1998, 2009) beschritten. Dabei wird zunächst für Glaubensgradfunktionen eine Konzeption von *Genauigkeit* oder *Wahrheitsnähe* eingeführt, die zu einem Begriff von *epistemischem Nutzen* führt. Mit Hilfe des Nutzenbegriffs können wir dann auf bekannte entscheidungstheoretische Prinzipien zurückgreifen, um mit ihrer Hilfe verschiedene epistemische Normen begründen zu können (s. dazu Pettigrew 2011, 2013). Dabei ist die wichtigste zunächst die des Probabilismus, wonach unsere Glaubensgrade sich nach den Wahrscheinlichkeitsaxiomen richten sollten, wir finden aber auch weitere Argumente für andere epistemische Normen.

Der besondere Vorteil dieser Argumentationen für bestimmte epistemische Normen soll gerade sein, dass sie *nicht-pragmatisch* sind bzw. nur auf die epistemischen Ziele hin ausgerichtet sind. Sie fragen also nicht (wie z.B. die Dutch-Book-Argumente), welche praktischen Auswirkungen inkohärente Überzeugungssysteme hätten, sondern versuchen den Probabilismus und andere epistemische Normen ausschließlich unter Bezugnahme auf die epistemischen Ziele und bestimmte Entscheidungsprinzipien zu begründen. Das ist ein anspruchsvolles Unterfangen, denn worauf sollen wir uns stützen, wenn wir nicht mehr über die praktischen Auswirkungen der epistemischen Normen sprechen dürfen, sondern nur noch über Wahrheitsnähe? Doch zumindest die dabei eingesetzten rationalen Entscheidungsregeln sind selbst natürlich wieder begründungsbedürftig, und es ist keineswegs klar, dass wir dazu nicht wiederum auf ihre praktischen Konsequenzen schauen müssen. Das Problem wird sich besonders deutlich bei der Argumentation für ein Indifferenzprinzip zeigen. Doch zunächst zum Probabilismus.

Nehmen wir also an, unsere epistemischen Subjekte haben eine Glaubensgradfunktion c:$\mathscr{A} \rightarrow [0,1]$ von einer endlichen Algebra $\mathscr{A}$ von Aussagen X nach [0,1], wobei Kontradiktionen 0 und Tautologien 1 erhalten sollen. (Dabei steht »c« für »credence«.) Wir verlangen aber noch nicht, dass die Funktion sich nach den Wahrscheinlichkeitsaxiomen richtet.

Für alle möglichen Situationen oder möglichen Welten w∈W, die vorliegen könnten, definieren wir nun jeweils eine Wahrheitswertfunktion v_w:

$$v_w(X) = \begin{cases} 1 & \text{falls X wahr ist in w} \\ 0 & \text{falls X falsch ist in w} \end{cases}$$

Den Wahrheitsabstand von c in w können wir dann durch den Abstand zwischen c und v_w bestimmen. Die dafür meistgenutzte Bewertungsfunktion ist der sogenannte *Brier-Score* B(c,w), der sich als Negation aus dem quadratischen Abstand $d(v_w,c)$ zwischen den beiden Funktionen ergibt:

$$d(v_w, c) = \sum_{X \in \mathscr{A}} |v_w(X) - c(X)|^2$$

und dann $B(c,w) = 1 - d(v_w,c)$, womit wir eine Bewertungsfunktion für die Wahrheitsnähe von c in w gefunden haben. B(c,w) kann dann auch als *epistemischer Nutzen* von c in w betrachtet werden, denn unser epistemisches Hauptziel ist die Annäherung an die Wahrheit und B(c,w) stellt intuitiv diese Annäherung dar (vgl. Pettigrew 2013, 2014). (Man beachte aber, je kleiner der Abstand d ist bzw. umso größer B ist, umso näher sind wir an der Wahrheit und umso größer ist der epistemische Nutzen.) Es können aber durchaus auch andere Bewertungsfunktionen zum Einsatz kommen. Wenn wir geeignete Anforderungen an diese Funktionen stellen, werden sich jeweils entsprechende Ergebnisse ableiten lassen.

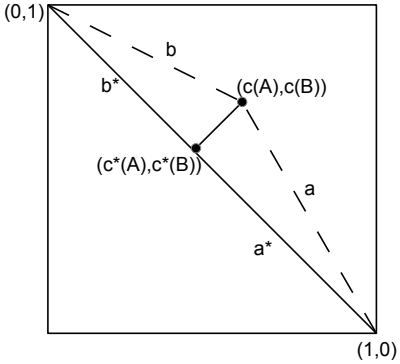
Ein Argument für den Probabilismus. Dazu ziehen wir als erstes die Dominanzregel aus der Entscheidungstheorie heran, wonach eine Option in jedem Fall vorzuziehen ist, wenn sie in allen möglichen Situationen, die auftreten können, höhere Nutzenwerte liefert als die anderen Optionen. Dazu gibt es ein wichtiges Theorem von de Finetti (1974, 87-91) und entsprechende moderne Varianten von Joyce (1998, 2009).

Theorem (von de Finetti)

(1) Wenn eine Glaubensgradfunktion c die Wahrscheinlichkeitsaxiome nicht erfüllt, dann gibt es eine Glaubensgradfunktion c*, die die Axiome erfüllt, mit: B(c*,w) > B(c,w) für alle w.
(2) Wenn eine Glaubensgradfunktion c* die Wahrscheinlichkeitsaxiome dagegen erfüllt, dann gibt es keine Glaubensgradfunktion c mit: B(c,w) > B(c*,w) für alle w.

Das bedeutet, dass es zu jeder inkohärenten Glaubensgradfunktion c (also eine, die die Wahrscheinlichkeitsaxiome nicht erfüllt) eine kohärente Funktion c* gibt, die sie dominiert im Sinne der epistemischen Nutzenfunktion. Oder man könnte auch sagen, es gibt zu einer inkohärenten Funktion c immer eine in allen Welten wahrheitsnähere kohärente Funktion c*. Kohärente Glaubensgradfunktionen werden dagegen nicht dominiert. Das zeigt, dass es *epistemisch irrational* ist, inkohärente Glaubensgrade aufzuweisen, da es dazu immer wahrheitsnähere kohärente Glaubensgrade gibt. Die rationale Entscheidungsregel für diesen Fall lautet: *Dominierte Strategien (oder Optionen) sind zu eliminieren*. Bevor wir eine Option auswählen, sollten wir in einem ersten Schritt schon einmal alle Optionen streichen, die von anderen verbliebenen Optionen dominiert werden. Damit bleiben nur noch die probabilistischen Glaubensgrade als rationale Lösungen unseres epistemischen Entscheidungsproblems übrig.

Wir können in einem einfachen Beispiel auch aufzeigen, wie man zu den wahrheitsnäheren kohärenten Glaubensgraden kommt. Nehmen wir an, wir hätten es nur mit zwei einander ausschließenden Aussagen A und B zu tun: $\mathscr{A}$ = {A,B} und es sei c(A) = 0,6 und c(B) = 0,7. Nun lassen sich alle Glaubensgradfunktionen jeweils durch einen Punkt (c(A),c(B)) im Einheitsquadrat darstellen (vgl. Pettigrew 2014):



Grafik 5.1: Wahrheitsabstände für Glaubensgradfunktionen

Im Bild haben wir die beiden Welten w1 = (1,0) und w2 = (0,1) und die Glaubensgradfunktion c, die den Wahrheitsabstand a^2 von w1 und b^2 von w2 aufweist. (Man beachte dabei, dass wir hier der Einfachheit halber gleich mit den quadrierten Abständen arbeiten und nicht jeweils erst die Wurzel ziehen. Außerdem setzen wir vereinfachend direkt die Welten w in c ein statt ihnen entsprechender Aussagen.) Auf der Diagonalen von (1,0) nach (0,1) liegen gerade die kohärenten Glaubensgradfunktionen. Fällt man vom Punkt (c(A),c(B)) das Lot auf diese Diagonale, erhält man als Schnittpunkt eine Glaubensgradfunktion c*, die zu beiden Welten offensichtlich näher liegt als c, da gilt: b > b* und a > a* (nach Phytagoras). Der epistemische Nutzen von c* in w1 ist dann: $B(c,w1) = 1–a^2 < B(c^*,w1) = 1–(a^*)^2$. Entsprechendes finden wir für die Abstände von w2. Als konkrete Werte ergeben sich damit:

c(A) = 0,6	c*(A) = 0,45
c(B) = 0,7	c*(B) = 0,55
a^2 = 0,65	$(a^*)^2$ = 0,605
b^2 = 0,45	$(b^*)^2$ = 0,405
B(c,w1) = 0,35	B(c*,w1) = 0,395
B(c,w2) = 0,55	B(c*,w2) = 0,595

Die Grafik belegt anschaulich, wieso die inkohärenten Glaubensgradfunktionen immer durch entsprechende kohärente Funktionen dominiert werden und wie man solche Funktionen finden kann.

Maher (2002) konnte allerdings zeigen, dass das Theorem nicht mehr gilt für die einfache Bewertungsfunktion, die nur die absoluten Abstände aufaddiert: $S(b,v) = 1/N \sum_n |b_n - v_n|$. Die ist nach Joyce aber auch ungeeignet, um als Bewertungsfunktion zu dienen. Sie ist nicht *konvex* (und nicht einmal »strictly proper«) und erfüllt auch andere plausible Anforderungen an Score-Funktionen nicht. Insbesondere kann es passieren (Joyce 2009, 282 f.), dass sogar ein logisch inkonsistentes Überzeugungssystem eine kohärente (probabilistische) Glaubensgradfunktion dominiert. Wenn wir das ausschließen möchten und zumindest einige schwächere Anforderungen an die Bewertungsfunktionen akzeptieren, dann kann Joyce zeigen, dass das oben genannte Theorem gilt, das die probabilistischen Glaubensgrade eindeutig positiv auszeichnet.

Insbesondere haben Predd et al. (2009) bewiesen, dass wir immer ein entsprechendes Theorem wie das von de Finetti erhalten, wenn wir eine *geeignete* (»strictly proper«) Bewertungsfunktion wählen, die einige elementare Eigenschaften erfüllt. In die längere Debatte über die Auszeichnung dieser Eigenschaften als erkenntnistheoretisch wünschenswert, kann ich hier nicht eintreten. Wir sollten jedenfalls zunächst konstatieren, dass insbesondere der Brier-Score eine naheliegende Bewertungsfunktion für Wahrheitsnähe darstellt (vgl. dazu Leitgeb & Pettigrew 2010a,b) und wir somit ein zusätzliches überzeugendes, erkenntnistheoretisches Argument für den Probabilismus gefunden haben.

Ein Argument für die Konditionalisierungsregel. Es lassen sich im Rahmen der epistemischen Nutzenfunktionen und der rationalen epistemischen Entscheidungstheorie aber auch noch weitere epistemische Normen begründen. Ab jetzt gehen wir immer davon aus, dass die Glaubengradfunktionen kohärent sind, es sich also um Wahrscheinlichkeitsfunktionen handelt. Wenn wir keine weiteren dominierten und dominanten Optionen mehr finden, dann ist eine bekannte entscheidungstheoretische Regel, die Option auszuwählen, die den höchsten Erwartungsnutzen aufweist.

Um den bestimmen zu können, müssen wir die Glaubensgrade $c(w)$ für die möglichen Welten w angeben können. Dafür müssen wir die Funktion für die entsprechenden Atome der Algebra $\mathscr{A}$ bzw. für die entsprechenden Vollkonjunktionen (wenn wir auf der syntaktischen

Ebene bleiben) hernehmen, die genau in w wahr sind (die also w entsprechen). Als *epistemischen Erwartungsnutzen* EU(c) erhalten wir dann:

$$EU(c) = \sum_{w \in W} c(w) \cdot B(c, w)$$

Damit können wir nun ein entscheidungstheoretisches Argument für die *Konditionalisierungsregel* erhalten. Greaves und Wallace (2006) haben bewiesen, dass die normale Update Regel: $c^+(X) = c(X|E)$ zu höheren erwartbaren epistemischen Nutzenwerten führt als eine beliebige andere Update-Regel, die zu einem kohärenten c* führt (mit $c^*(E) = 1$).

Theorem (von Greaves und Wallace)
Für das Updaten mit bestimmten Daten E gilt:
$\sum_{w \in W} c(w) \cdot B(c^+, w) > \sum_{w \in W} c(w) \cdot B(c^*, w)$ für beliebige $c^* \neq c^+ := c(\cdot|E)$

Greaves und Wallace haben das noch für allgemeinere Fälle bewiesen – nicht nur für den Brier-Score und auch für allgemeinere Formen von Daten. Der Beweis bietet immerhin einen interessanten neuen Weg zur Begründung der Konditionalisierungsregel. Wir könnten es auch so formulieren, dass auf der Grundlage der bisherigen epistemischen Wahrscheinlichkeiten c das klassische Updaten mit Hilfe der Konditionalisierungsregel die Glaubensgradfunktion mit der größten zu erwartenden Wahrheitsnähe (gegenüber anderen Update-Regeln) liefert, wobei wir den Durchschnitt über die Wahrheitsnähe in allen möglichen Welten w bilden. Wie zwingend diese Argumentation ist, wird sicher noch weiter zu diskutieren sein, da die epistemische Entscheidungstheorie noch in den Kinderschuhen steckt. Sie stellt allerdings eine interessante neue Entwicklung dar, die sich nicht nur auf den praktischen Nutzen unserer Glaubensgrade stützt, sondern stärker im erkenntnistheoretischen Bereich bleibt.

Ein Argument für das Indifferenzprinzip. Mich interessiert aber noch mehr eine andere Anwendung dieser Entscheidungstheorie. Richard Pettigrew behauptet in verschiedenen Aufsätzen (u.a. in 2014), dass wir hier auch noch ein entscheidungstheoretisches Argument für das Indifferenzprinzip gewinnen können, das stärker sei, als die bisher genannten Argumente. Darüber würde ich mich freuen, aber ich bin leider skeptisch, ob ihm das wirklich gelungen ist.

Er ist dafür auf ein anderes entscheidungstheoretisches Prinzip angewiesen, nämlich das *Minimax-Prinzip* (manchmal auch als Maximin-Prinzip bezeichnet), wonach wir jeweils die Option wählen sollten, die den maximalen Schaden minimiert. Das heißt, wir geben für jede Option (oder Strategie) an, wie groß der Schaden im schlimmsten Fall (wenn die Welt sich als für uns ungünstig erweist) ist, und wählen dann die Option aus, für die der Schaden (=negativer Nutzen) am kleinsten ist. (Man kann auch sagen, wir maximieren den Nutzen unter all den minimalen Nutzenwerten, die auftreten können.) In unserem Fall bedeutet das, dass wir die Glaubensgradfunktion c wählen, die im ungünstigsten Fall den kleinsten Wahrheitsabstand aufweist.

Minimax: Wähle c so, dass ($\min_{w \in W}$ B(c,w)) maximal wird.

Tatsächlich führt diese Strategie dazu, dass wir immer eine möglichst indifferente (falls noch weitere Nebenbedingungen vorliegen, muss das nicht die Gleichverteilung sein) Glaubensgradfunktion wählen sollten. Im Falle von nur zwei Aussagen A und B wie in unserem Beispiel oben ist das offensichtlich. Der maximal mögliche Abstand zu den beiden Welten w1 und w2 wird dann am kleinsten, wenn wir für die Glaubensgradfunktion den Wert in der Mitte der Diagonalen wählen. Wir gehen das kleinstmögliche Risiko ein, wenn wir unsere Glaubensgrade in der Mitte wählen, bzw. möglichst gleichmäßig verteilen, wenn es noch mehr als zwei einander ausschließende Aussagen gibt, die wir als grundlegende Möglichkeiten betrachten können.

Diese Minimax-Strategie ist die *pessimistische Strategie*, die es vor allem darauf anlegt, den Schaden klein zu halten. Wenn wir mit unseren Glaubensgraden schon danebenliegen, dann sollten wir selbst im schlimmsten Fall nicht mehr daneben liegen, als es unvermeidlich ist. Wir wissen aber aus der Spieltheorie, dass diese Strategie für Entscheidungen unter Unsicherheit auch ihre Schwächen hat und keinesfalls unumstritten ist. Sie verschenkt womöglich die Chance auf größere Gewinne, nur um den Preis, dass der Schaden dabei im schlimmsten Fall noch ein wenig größer ausfallen könnte. Entscheidungstheoretiker führen dazu gern Situationen wie die folgende an:

	w1	w2
s1	1	100
s2	2	3

Wir haben hier zwei Strategien s1 und s2 zur Auswahl und es gibt zwei möglichen Situationen (oder Welten), die auftreten können, nämlich w1 und w2. Dann ergeben sich jeweils die angegeben Nutzenwerte. Am schlechtesten ist es für beide Strategien, wenn w1 vorliegt. Also sollten wir nach Minimax s2 wählen, weil dann das schlechteste Ergebnis 2 ist und nicht 1 wie bei s1. Doch ist das wirklich rational? Wir lassen uns damit jede Chance auf den möglichen Gewinn 100 entgehen. Der Optimist wird in jedem Fall anders an die Sache herangehen: Er wird s1 wählen, denn dann kann er 100 Nutzen gewinnen und der im schlimmsten Fall (w1) auftretende Verlust gegenüber der Entscheidung für s2 beträgt nur 1 Nutzen und ist damit verschmerzbar.

Die Situation könnte etwa die folgende sein: Nehmen wir an, wir fragen uns, ob wir ein Medikament bei einer schweren Krankheit nehmen sollten (s1) oder eben nicht nehmen sollten (s2). Es können zwei Fälle auftreten: (w1) Das Medikament wirkt nicht, hat aber schwache Nebenwirkungen. (w2) Das Medikament wirkt und heilt die Krankheit. Die Nutzenwerte können dann wie in unserem Zahlenbeispiel oben sein. Es scheint kaum rational zu sein, sich die Chance auf Heilung entgehen zu lassen, selbst wenn wir nicht beziffern können, wie groß sie ist. Wir werden wegen des möglichen großen Gewinns und der nicht so großen Nebenwirkungen vernünftigerweise das Medikament einnehmen.

Der Optimist wird vielleicht sogar für die *Maximax-Strategie* oder eine komplexere Regel (etwa als Kombination aus den beiden Strategien) eintreten. Maximax besagt, man solle die Option wählen, die im besten Fall auch den größten Nutzen aufweist.

John Rawls stützt sich in seiner *Theorie der Gerechtigkeit* ebenfalls auf die Minimax-Regel. Es scheint für ihn rational zu sein, sich im Urzustand unter dem Schleier des Nichtwissens – in dem die grundlegenden Prinzipien für eine gerechte Gesellschaft ausgewählt werden sollen – eher risikoscheu zu verhalten. Ließe man etwa eine Sklavenhaltergesellschaft zu, müsste man damit rechnen, dass es einen auch selbst erwischen könnte und man in dieser Gesellschaft als Sklave leben müsste. Es

erscheint mir durchaus rational, dass man hier lieber auf Nummer sicher geht und Minimax zur Anwendung kommt.

Aber ist das auf unsere Anwendungen übertragbar? Zunächst einmal ist es tatsächlich so, dass wir das Indifferenzprinzip als Verfahren einsetzen möchten, um erste Glaubensgrade zu erzeugen. Also liegt auch hier eine Entscheidung unter Unwissenheit vor, bei der wir für die unterschiedlichen Situationen, die auftreten können, noch keine Wahrscheinlichkeiten angeben können. Aber ist es auch rational, sich in Bezug auf Wahrheitsabstände eher *risikoscheu* zu verhalten? Der auftretende Schaden ist nur epistemischer Natur. Dafür haben wir keine klaren Vorstellungen, wie das zu bewerten ist. Erst wenn wir diese Glaubensgrade wieder in anderen praktischen Entscheidungen zum Einsatz bringen, haben wir klare Vorstellungen, wie Schaden und Nutzen zu verrechnen sind. Dann handelt es sich aber nicht mehr um eine rein epistemische Entscheidungstheorie, sondern wir bringen wiederum pragmatische Aspekte ins Spiel.

Doch ohne eine Ankopplung an praktische Fragen sehe ich nicht, wie wir zu entsprechenden Einschätzungen gelangen können. Allerdings kommen wir dann wieder zurück in ähnliche Situationen, wie ich sie für das *Prinzip der epistemischen Gleichbehandlung* herangezogen habe, für das ich in Kapitel 5.3.15 noch ausführlich eintreten werde. Die Behandlung durch die epistemische Entscheidungstheorie ist dann nicht mehr ganz so revolutionär und überlegen, wie es sich Pettigrew wünscht.

Das Argument von Pettigrew leidet also darunter, dass das Minimax-Prinzip bekanntermaßen nicht unumstritten ist. Über die anderen eigesetzten Entscheidungsprinzipien herrscht dagegen eine weit größere Einigkeit. Pettigrew gesteht daher auch ein, dass es schwer ist, für das Minimax-Prinzip in der epistemischen Entscheidungstheorie überzeugend zu argumentieren, weil es sich um ein sehr grundlegendes erkenntnistheoretisches Prinzip handelt und wir in unseren Argumentationen kaum auf andere Prinzipien zurückgreifen können. Schauen wir uns dazu kurz an einem Beispiel an, weshalb er die anderen Argumente für das Indifferenzprinzip für schwächer als sein eigenes hält.

Kritik anderer Argumente für ein Indifferenzprinzip. Pettigrew (2014) geht noch auf andere Argumente für das Indifferenzprinzip (IP) ein und hält sie alle für schwächer als sein Nutzen-Argument. Vor

allem wendet er sich gegen ein Argument von Roger White (2010), das meiner Argumentation für ein (IP) in Kapitel 5.3.15 mit Hilfe eines epistemischen Gleichbehandlungsprinzips ähnlich ist. Danach sollte man gleiche Glaubensgrade für Aussagen A und B vergeben, wenn wir gleich starke (symmetrische) Gründe für A und B haben und sollte dann nicht eine der Aussagen bevorzugen. Was wendet nun Pettigrew (2014) dagegen ein?

Sein erster Einwand besagt, dass unser bestes Verfahren zur Beurteilung von Begründungen der Bayesianismus sei. Um die Voraussetzung des Indifferenzprinzip feststellen zu können, wären wir daher bereits auf Glaubensgrade angewiesen und die Argumentation für das (IP) würde damit zirkulär.

Das kann man kaum sehr ernst nehmen. Der Bayesianismus verhilft uns zwar zu einer Buchhaltung unserer Rechtfertigungsbeziehungen, aber die existieren nicht erst aufgrund des Bayesianismus. Der reproduziert nur bestimmte inferentielle Beziehungen zwischen unseren Aussagen – wenn z.B. bestimmte Datenaussagen deduktiv aus einer Theorie folgen –, und wir haben sehr wohl bereits ohne den Bayesianismus anzuwenden viele Situationen, in denen gemäß unserem Hintergrundwissen zwei Aussagen A und B gleich gut begründet sind.

Etwas gehaltvoller könnte da schon der zweite Einwand sein. Er fragt danach, was man eigentlich dabei gewinnt, wenn man sich an das Gleichbehandlungsprinzip hält. In der epistemischen Nutzentheorie geht es um ein klar definiertes epistemisches Ziel, dem wir uns nähern, nämlich der Wahrheitsnähe (oder Genauigkeit) bzw. wir streben nach höherem epistemischen Nutzen. Doch was ist der Gewinn beim Indifferenzprinzip? Wenn wir uns nicht auf den epistemischen Nutzen beziehen möchten, bleiben nach Pettigrew nur praktische Gewinne, die aber weniger direkt für erkenntnistheoretische Überlegungen geeignet erscheinen. Die Stoßrichtung der epistemischen Nutzentheorien war schließlich, von den pragmatischen Argumenten für epistemische Normen zu rein epistemischen Überlegungen überzugehen.

Da es sich beim Indifferenzprinzip um ein sehr grundlegendes epistemisches Rationalitätsprinzip ist, ist der Gewinn natürlich nicht leicht zu benennen, und ich werde daher auch in Kapitel 5.3.15 vor allem anhand der weiteren praktischen Folgen dafür argumentieren. Doch darauf ist

Pettigrew genaugenommen ebenfalls angewiesen. Sein Minimax Prinzip läuft doch nur darauf hinaus, dass wir im Rahmen einer pessimistischen Einstellung möglichst den größten Abstand von der Wahrheit, der auftreten könnte, klein halten wollen, dabei aber jede Chance verspielen, exakt richtig zu liegen. Denn diesen Abstand klein zu halten sei schließlich unser epistemisches Ziel.

Aber warum sollten wir statt für A und B je 0,5 als Glaubensgrad zu vergeben, nicht einfach P(A) = 1 und P(B) = 0 setzen? Ein Optimist könnte für das Prinzip Maximax plädieren und eine entsprechende Wahl genauso durch unser epistemisches Ziel begründen. Mit der neuen Vorgehensweise haben wir sogar die Chance, unser Ziel vollständig umzusetzen und nicht nur die Abstände nicht zu groß werden zu lassen. Um für Minimax gegenüber Maximax zu argumentieren, reicht es nicht, sich auf unser epistemisches Ziel zu berufen, denn das können beide Seiten für sich reklamieren. Es wird nur auf etwas anderen Wegen verfolgt. Einmal mit einer eher pessimistischen Einstellung und einmal mit mit einer optimistischeren. Außerdem sind noch Mischformen dazwischen denkbar, die in der Entscheidungstheorie auch ernsthaft diskutiert werden und vermutlich die besseren Ergebnisse bringen. Welche der Strategien tatsächlich rationaler ist, kann auch hier nur eine weitere Diskussion der daraus folgenden Konsequenzen zeigen. Die sind aber durchaus unklar, wie wir aus der Debatte um Entscheidungen unter Unsicherheit schon wissen. Es besteht sogar die Gefahr, dass wir hier keine Regel als beste Rationalitätsregel auszeichnen können.

Die Schwäche des Minimax Prinzips stellt damit eine Schwäche des pettigrewschen Arguments für das Indifferenzprinzip dar, während das Dominanzprinzip viel stärker ist und damit das entsprechende Argument für den Probabilismus deutlich überzeugender war. Daher glaube ich nicht, dass Pettigrew zu Recht behaupten kann, sein Argument wäre klar stärker als die anderen Argumente für das Indifferenzprinzip. Insbesondere werde ich in Kapitel 5.3.15 dafür eintreten, dass das *Prinzip der epistemischen Gleichbehandlung* selbst gut begründbar ist und daraus sowohl das klassische Indifferenzprinzip wie auch der statistische Syllogismus folgen, auf den wir in besonderer Weise angewiesen sind.

Eine weitere Argumentation für ein Indifferenzprinzip finden wir als dritten Punkt in Leitgeb und Pettigrew (2010), wo anhand von

geometrischen Überlegungen ebenfalls noch abgleitet wird, dass der beste Ansatz bei den Startwahrscheinlichkeiten der des objektiven Bayesianers ist, der ein Indifferenzprinzip (auf der Grundlage unserer Menge von möglichen Welten) annimmt. Das ist insgesamt sicher ein sehr kreativer Ansatz, gleichwohl müssen die Autoren selbst zugeben, dass sie sehr viele idealisierende Annahmen zugrunde gelegt haben, um die Übersetzung unserer Frage in ein geometrisches Problem zu bewerkstelligen, die alle für sich bestreitbar sind. Auch diese Überlegungen sind also weit davon entfernt, eine intuitiv zwingende Argumentation für den probabilistischen Ansatz zu ergeben. Außerdem führt sie uns nicht zum klassischen Bayesianismus. Es gibt somit insgesamt einige schwächere Gründe dafür, sich in seinen Glaubensgraden nach den Wahrscheinlichkeitsaxiomen zu richten, doch letztlich keine starken oder sogar endgültigen Begründungen. Die Anwendbarkeit und die Erfolge des Ansatzes müssen für sich sprechen.

An dieser Stelle kann uns auch der statistische Syllogismus weiterhelfen. Wenn ich meine Glaubensgrade an den relativen Häufigkeiten ausrichten möchte, so spricht das für die Annahme der Additivitätsregel. Nehmen wir an, wir haben einen verfälschten Würfel mit den Glaubensgraden oder Wahrscheinlichkeiten P(4) = 1/5 = P(5), die sich aus bestimmten relativen Häufigkeiten ergeben haben. Wie hoch sollte dann unser Glaubensgrad bzw. unsere Wahrscheinlichkeit dafür sein, dass die 4 oder die 5 beim nächsten Wurf oben liegen? Die relative Häufigkeit dafür wird bei 2/5 liegen und daher sollten wir diese auch für unsere Glaubensgrade übernehmen. Allerdings ist einschränkend zu ergänzen, dass der statistische Syllogismus ja nur für den Fall gilt, dass wir über keine anderen spezifischen Informationen über das Zutreffen der disjunktiven Behauptung verfügen.

5.3.6 Die Rigiditätsforderung

Es gibt eine Reihe von Ansätzen, die andere Axiome anführen, die uns intuitiver erscheinen sollen und die es letztlich gestatten, die Wahrscheinlichkeitsaxiome abzuleiten. Der bekannteste ist wohl der Ansatz von Cox (1946). Allerdings bleibt bei all diesen Ansätzen immer die Frage, ob die anfänglichen Axiome tatsächlich so viel plausibler sind

als die Wahrscheinlichkeitsaxiome selbst. Vermutlich bleibt uns letztlich zur Beantwortung der Frage nach der Plausibilität des probabilistischen Ansatzes nichts anderes übrig, als eine Gesamtbewertung seiner Erfolge in der Anwendung durchzuführen und dabei sind diese anfänglichen Einschätzungen zum Schluss schließlich nicht mehr von so großer Bedeutung.

Eine andere Idee zur Begründung der Kohärenzforderung finden wir in der sogenannten Rigiditätsforderung. Sie ist die entscheidende Voraussetzung, die uns zu Jeffreys-Regel der Konditionalisierung führt. Danach behalten wir beim Updaten, wenn wir ein neues Datum E lernen zumindest die konditionalen Wahrscheinlichkeiten $P(B|E)$ für alle Aussagen B bei:

> **Die Forderung der Invarianz oder Rigidität unseres probabilistischen Überzeugungssystems**: Wenn wir mit einem neuen Datum E von P zu P^+ updaten, gilt für alle weiteren Aussagen B unseres Überzeugungssystems: $P^+(B|E) = P(B|E)$.

Man sieht zunächst leicht ein, dass daraus die Regel von Jeffrey folgt, denn für die neue Wahrscheinlichkeitsfunktion P^+ muss gelten: $P^+(B) = P^+(B|E)\cdot P^+(E) + P^+(B|\neg E)\cdot P^+(E)$. Wenn wir nun an den alten bedingten Wahrscheinlichkeiten festhalten, erhalten wir genau die Formel für das Updaten von Jeffrey: $P^+(B) = P(B|E)\cdot P^+(E) + P(B|\neg E)\cdot P^+(\neg E)$ (s.Kap. 5.5.5) und im Falle, dass wir $P^+(E) = 1$ und damit $P^+(\neg E) = 0$ setzen, erhalten wir gerade die einfache Konditionalisierungsregel. (Für Jeffreys Regel sind wir dagegen nicht gezwungen die upgedatete Wahrscheinlichkeit für das Datum E auf 1 zu setzen.)

Die Rigiditätsforderung wird von Bayesianern üblicherweise als Rationalitätsforderung betrachtet und wird zum einen durch Dutch-Book-Argumente begründet, aber ebenso durch ein intuitives Verständnis des Updatens (vgl. Bradley 2005). Als neues Datum lernen wir nur, dass E der Fall ist (oder sich die Wahrscheinlichkeit von E verändert), dann müssen sich natürlich auch die Wahrscheinlichkeiten der Aussagen ändern, die aus E oder ¬E folgen. Aber andere Wahrscheinlichkeiten sollten wir beibehalten. Insbesondere sollten wir vernünftigerweise (aus Sicht der Bayesianer) nicht die Wahrscheinlichkeiten ändern, die das Verhältnis

von E zu anderen Aussagen betreffen. Die konditionale Aussage »wenn E, dann A« sollte keine neuen Wahrscheinlichkeitswerte erhalten, denn dazu haben wir eigentlich keine neuen Einsichten erhalten und sollten die Einschätzung dieser Aussage daher im Sinne einer möglichst konservativen Überzeugungsänderung auch beibehalten.

Beim klassischen Updaten mit einem Datum E mit $P^+(E) = 1$ und $P^+(\neg E) = 0$ werden zunächst für alle Aussagen A ihre Anteile, die aus ¬E folgen auch auf 0 gesetzt: $P^+(A\&\neg E) = 0$. Für die restlichen Anteile der Aussagen A&E gilt dann, dass ihre Wahrscheinlichkeit erhöht wird, wobei die weggefallen Wahrscheinlichkeitsanteile nun entsprechend den bisherigen Anteilen dieser Aussagen auf die Aussagen A&E und B&E verteilt werden, so dass sich aus der Rigiditätsforderung insgesamt ergibt:

$$\frac{P^+(A\&E)}{P^+(B\&E)} = \frac{P(A\&E)}{P(B\&E)}$$

Das sieht man leicht, anhand der Formel: $P(A\&E) = P(E)\cdot P(A|E)$ und der entsprechenden Formel für P^+. Das kann wiederum als eine Anwendung der Idee der minimalen Änderung betrachtet werden (vgl. Joyce 2009 und hier Kap. 5.5.5). Die Wahrscheinlichkeitsverhältnisse bleiben beim Updaten erhalten – insoweit das möglich ist. Es gibt nur Zugewinne durch die erforderliche Umverteilung, weil bestimmte Aussagen nun auf Wahrscheinlichkeit null gesetzt werden und die frei werdenden Wahrscheinlichkeitsanteile entsprechend auf die verbliebenen Aussagen verteilt werden.

Die Regel von Jeffrey, die wir später ausführlicher diskutieren werden, stützt sich auf die Rigiditätsforderung, erlaubt aber, dass die neuen Daten eine Wahrscheinlichkeit kleiner 1 erhalten ($P^+(E)<1$). Nur die Rigidität bleibt erhalten, womit wir (automatisch?) zu der genannten Update-Regel gelangen, auf die wir später (in Kap. 5.5.5) noch gesondert eingehen werden:

$$P^+(H) = P(H|E)\cdot P^+(E) + P(H|\neg E)\cdot P^+(\neg E)$$

Gegen diese Form von Rigidität argumentiert allerdings Bradley (2005).

5.3.7 Komparative Bestätigung und qualitative Wahrscheinlichkeit

Wir hatten schon gesehen, dass es keine einfachen Zusammenhänge zwischen kategorischem Glauben bzw. Akzeptanz von Aussagen auf der einen Seite und Glaubensgraden auf der anderen Seite gibt. Insbesondere akzeptieren Probabilisten typischerweise nicht die *locksche These* (so nennt sie z.B. Hawthorne 2009), wonach ein hinreichend hoher Glaubensgrad zugleich hinreichend für das Akzeptieren der entsprechenden Überzeugung sein soll, da wir sonst wieder in das Lotterie-Paradox zurückfallen.

Ein älterer, aber m.E. sehr interessanter Weg, um den Zusammenhang weiter aufzuklären, ist der über einen komparativen Bestätigungsbegriff zu einem (möglichst eindeutig bestimmten) quantitativen zu kommen. Das ist ähnlich wie für andere Konzepte in der Wissenschaft, bei denen wir mit einem komparativen Konzept starten und darauf aufbauend etwa anhand von Repräsentationstheoremen zu einem quantitativen Konzept gelangen. Das sagt uns mehr darüber, was wir für den quantitativen Begriff tatsächlich benötigen. Die Grundlagen finden sich bei Savage (1972), dann weitere Ergebnisse im Überblick bei Fishburn (1986) bis hin zu den Überlegungen von Hawthorne (2009). Bei Fishburn (1986) finden sich darüber hinaus Hinweise auf die mathematischen Zusammenhänge und Beweise für unterschiedliche Situationen sowie Gegenbeispiele für einige Fälle, in denen nicht alle Anforderungen erfüllt sind. Dieser Weg kann obendrein ganz neue Gründe dafür bieten, mit einem quantitativen Glaubensgradkonzept zu arbeiten. Deshalb verfolgen wir ihn an dieser Stelle.

Insbesondere könnten wir hier auch auf eine gute Begründung für die Geltung der Wahrscheinlichkeitsaxiome für Glaubensgrade stoßen. Aus unserer klassischen Erkenntnistheorie und aus unserem Alltag kennen wir zunächst Konzepte wie »A ist plausibler als B« oder »A ist besser begründet als B« und Ähnliches. Die Frage wird sein, ob wir von diesen Alltagsbegriffen in relativ intuitiver Weise zu unseren Wahrscheinlichkeitsaxiomen gelangen können, indem wir etwa naheliegende Rationalitätsforderungen an diese komparativen Begriffe stellen. Lassen sie sich dann repräsentieren durch (eindeutig festgelegte)

Wahrscheinlichkeitsfunktionen? Leider wird sich dieser Weg als steiniger erweisen, als man zunächst dachte.

Doch bevor ich darauf eingehe, möchte ich mit einem Unterthema starten, nämlich der Frage, ob Wissen und epistemische Rechtfertigung bzw. das Akzeptieren von Aussagen abgeschlossen unter Konjunktionen sind. Die Konjunktionen stellen das spezielle Problem beim Lotterieparadox dar. Wenn ich weiß, dass p und weiß, dass q, weiß ich dann auch, dass p&q? Wir hatten die generellen Probleme der deduktiven Abgeschlossenheit von Wissen schon (im Umfeld des Skeptikers) diskutiert, stoßen hier aber auf eine speziellere Frage. Schauen wir uns dazu die Wissensbedingungen kurz der Reihe nach an. Die Wahrheitsbedingung ist unproblematisch. Die Glaubensbedingung sollte es eigentlich ebenso sein, obwohl wir hier natürlich bereits einwenden können, dass dann für viele atomare Überzeugungen noch sehr viele Konjunktionen hinzukommen müssen, die es ebenfalls zu glauben gilt. Doch wir können zumindest sagen, dass wir diese konjunktiven Überzeugungen wenigstens *implizit* besitzen, wenn wir bereit sind, ihre Elemente auf Nachfrage hin auch explizit zu akzeptieren.

Größere Probleme machen aber die Begründungsfragen, das zeigte schon das Lotterie-Paradox. Wenn wir nämlich viele begründete Überzeugungen zusammenfügen, muss die Konjunktion nicht mehr genauso gut begründet oder überhaupt gut begründet sein. Das gilt letztlich genauso für unanfechtbare Begründungen, es sei denn, wir würden die gleichartigen Aussagen im Lotterie-Paradox, die aber doch nicht zusammen wahr sein können, als eine Art von Gegengrund gegen die Bestätigung der jeweiligen einzelnen Aussagen auffassen, was mir jedoch recht unplausibel erscheint. So war das Konzept der Gegengründe zunächst nicht gemeint.

Allerdings lässt sich das Lotterieproblem hier nicht direkt reproduzieren, weil Wissen verlangt, dass alle Konjunkte wahr sind. Trotzdem bleiben wir mit einem intuitiven Problem übrig: Einerseits sollte der Wissensbegriff so stark sein, dass er unter Konjunktionen stabil bleibt, andererseits scheinen die Bedingungen für Wissen das nicht zu garantieren. Einen möglichen Kompromiss zeigt Hawthorne (2009) auf. Jedenfalls würden wir uns eine entsprechende – vielleicht etwas eingeschränkte – Stabilität auch für Begründungen wünschen und wollen nun nachsehen,

inwiefern uns das Konzept der *komparativen Bestätigung* bzw. der *qualitativen Wahrscheinlichkeit* (so wird es oft genannt) dabei helfen kann, zumindest eine gewisse Stabilität der Begründungen und des Akzeptierens unter Konjunktionen zu garantieren.

Dazu gehen wir als neuem Grundbegriff der *komparativen Bestätigung* davon aus, dass wir »A $\succeq$ B« für zwei Aussagen A und B lesen als »*A ist mindestens genauso plausibel wie B*« oder »*A ist mindestens so gut begründet wie B*«, wobei das immer für ein bestimmtes epistemisches Subjekt S gemeint ist bzw. für das jeweilige Hintergrundwissen von S. Um hier aber nicht immer einen Index mitzuschleppen, lasse ich diese Bezugnahme auf S weg. Für diesen komparativen Bestätigungsbegriff können wir nun eine Reihe von Regeln angeben, die ein ideal rationales epistemisches Subjekt erfüllen sollte. Wir werden also eine Art von Logik der komparativen $\succeq$-Beziehung aufstellen.

Dafür wurden unterschiedliche Axiomensysteme vorgeschlagen. Ausgangspunkt waren die Axiome von de Finetti (von 1937), die zunächst recht übersichtlich aussehen. Dabei erhalten wir für eine vorgegebene Sprache L mit den Aussagen A, B, C,… die folgenden de Finetti Axiome:

(1) *Nichtnegativität*: $A \succeq \bot$ ($\bot$ Kontradiktion)
(2) *Nichttrivialität*: $\top \succ \bot$ ($\top$ Tautologie)
(3) *Vollständigkeit*: $A \succeq B$ oder $B \succeq A$
(4) *Transitivität*: $A \succeq B$ und $B \succeq C$, dann $A \succeq C$
(5) *Quasiadditivität*: Wenn A, C und B, C jeweils einander ausschließen, dann gilt: $A \succeq B \Leftrightarrow A \vee C \succeq B \vee C$

De Finetti hoffte, dass diese 5 Axiome bereits ausreichen, damit sich die Beziehung »$\succeq$« durch eine entsprechende Wahrscheinlichkeitsfunktion P repräsentieren lässt. Dabei sagen wir zunächst, dass P und $\succeq$ *übereinstimmen*, wenn für alle A und B gilt:

$A \succeq B$ gdw. $P(A) \geq P(B)$

Dann *repräsentiert* P unsere komparative Funktion $\succeq$, wenn P und $\succeq$ übereinstimmen. Die Axiome wirken auch relativ plausibel, wobei es aber einige Einwände vor allem gegen die Transitivitätsforderung gab. Doch dem möchte ich hier nicht weiter nachgehen, weil für die

Repräsentierbarkeit noch ein weit stärkeres Axiom erforderlich ist, dass sicherlich nicht als direkt einleuchtend zu betrachten ist (außerdem werden ähnliche Axiome weiter unten noch diskutiert).

Kraft et al. (1959) lieferten ein Gegenbeispiel zu de Finettis Vermutung, nämlich eine Struktur, die die 5 Axiome erfüllte, aber nicht probabilistisch repräsentierbar ist. Das Problem zeigt sich, wenn wir für (endliche) Aussagenmengen und endliche Mengen möglicher Welten, die Aussagen A durch die Menge ihrer möglichen Welten $w \in [A]$ darstellen, die Additivitätseigenschaft von P nachweisen möchten. Dafür muss dann nämlich die entsprechende Additivitätseigenschaft gelten:

$A \succeq B$ gdw. $\sum_{w\in[A]} P(w) \geq \sum_{w\in[B]} P(w)$

Dana Scott (1964) gab dann eine Version eines fehlenden verstärkten Additivitätsaxioms an. Um diese Ergänzungen besser verstehen zu können, benötigen wir zunächst noch einige Begriffe. Zunächst sei unsere Sprache L endlich mit endlichen vielen atomaren Aussagen $L = \{A_1,\ldots,A_n\}$, die zusammen mit allen logischen Kombinationsmöglichkeiten unsere Sprache L aufspannen sollen. Die schon bekannten Vollkonjunktionen $w = \pm A_1 \& \ldots \& \pm A_n$ von Aussagen aus X stellen dann alle möglichen Situationen (oder Welten oder Weltzuständen) dar. Nun betrachten wir zwei Folgen von Aussagen aus L: $X = \langle x_1,\ldots,x_k\rangle$ und $Y = \langle y_1,\ldots,y_k\rangle$ und nennen X und Y *ausgeglichen*, wenn für jeden Weltzustand w, die Anzahl aller wahren Aussagen in X gleich der Anzahl aller wahren Aussagen in Y ist. Dabei heißt x_j ist wahr in w gerade: $w \Rightarrow x_j$. Damit wird aus unserer Bedingung der Ausgeglichenheit:

X und Y sind *ausgeglichen* gdw. für all w gilt:
$Anz(\{j;\ w \Rightarrow x_j\}) = Anz(\{j;\ w \Rightarrow y_j\})$

Für solche ausgeglichen Folgen von Aussagen müssen wir dann verlangen, dass unsere komparative Wahrscheinlichkeit sie respektiert. Erst dann lässt sich nachweisen (s. etwa Scott 1964), dass sie auch numerisch repräsentierbar ist. Das Scott-Axiom wird auch manchmal als verstärkte Additivitätsbedingung bezeichnet:

(**Scotts-Axiom**) Für alle zwei Folgen von k Aussagen X und Y (und alle k) gilt:

(1) Wenn X und Y ausgeglichen sind und
(2) wenn für alle j < k gilt: $x_j \succeq y_j$, dann gilt:
(3) es ist nicht der Fall: $x_k \succ y_k$.

Sind die Folgen ausgeglichen, dann muss sich das auch in den komparativen Wahrscheinlichkeiten wiederfinden, die sollten dann jedenfalls nicht X an jeder Stelle als mindestens gleichplausibel und einer Stelle als echt plausibler betrachten. Leider können wir kaum behaupten, dass das Scott Axiom selbst offensichtlich wahr oder auch nur leicht einsehbar wäre. Es dokumentiert vielmehr, wie steinig der Weg von einer komparativen Plausibilitätseinschätzung zu einer numerischen probabilistischen Repräsentation ist. Trotzdem hoffe ich, dass die genannten Axiome unser Verständnis von quantitativen Wahrscheinlichkeiten verbessern können, weil sie zumindest z.T. recht gut verständlich sind und an unsere intuitiven Urteile im Alltag, aber auch in der Wissenschaft anknüpfen.

Für den unendlichen Fall hat Savage eine Art von archimedischem Axiom eingeführt, das für die probabilistische Repräsentierbarkeit erforderlich ist. Das hat Hawthorne (2009) aufgegriffen und darauf seine Glaubenskonzeption gestützt, die ich im Folgenden vorstellen möchte. Dazu definiert er zunächst ein darauf aufbauendes Konzept, nämlich das der Gewissheit »gewiss(A)«, das gewissermaßen unsere oberste Stufe an Plausibilität oder Begründetheit beschreibt. Es lässt sich definieren, durch einen Vergleich mit einer Tautologie:

$\text{gewiss}(A) \succeq A \vee \neg A$

Dann erhalten wir in einem ersten Schritt für eine vorgegebene Sprache L mit den Aussagen A, B, C,… die folgenden Axiome:

Basale komparative Bestätigungsbeziehung (BKB)
Für alle Aussagen A, B, C, D, der Sprache L gilt:
A1. Es gilt nie: $\neg(A \vee \neg A) \succeq (A \vee \neg A)$ (*Nichttrivialität*)
A2. $B \succeq \neg(A \vee \neg A)$ (*Minimalität*)
A3. $A \succeq A$ (*Reflexivität*)
A4. Wenn $A \succeq B$ und $B \succeq C$, dann gilt: $A \succeq C$ (*Transitivität*)
A5.1 Wenn gilt: gewiss[C≡D] und $A \succeq C$, dann gilt: $A \succeq D$ (*Rechtsäquivalenz*)

A5.2 Wenn gilt: gewiss[C≡D] und C ≥ B, dann gilt: D ≥ B (*Linksäquivalenz*)
A6.1 Wenn für ein E gilt: gewiss[¬(A&E)], gewiss[¬(B&E)] und (A∨E) ≥ (B∨E), dann gilt: A ≥ B (*Subtrahierbarkeit*)
A6.2 Wenn gilt: A ≥ B, dann gilt: für alle G mit gewiss[¬(A&G)] und gewiss[¬(B&G)], (A∨G) ≥ (B∨G) (*Additivität*)
A7. Wenn A eine logische Tautologie ist, dann gilt: gewiss[A] (*tautologische Gewissheit*).

Dazu können wir durch einige einfache Definitionen neue Konzepte einführen:

A = B (gelesen: A ist genauso plausibel/gut bestätigt wie B) gdw.
A ≥ B und B ≥ A
A ≻ B (gelesen: A ist echt besser bestätigt als B) gdw.
A ≥ B und nicht B ≥ A, und
A ≈ B (gelesen: As Bestätigung ist unbestimmt relativ zu der von B) gdw. nicht (A ≥ B) und nicht (B ≥ A)

Daraus lassen sich bereits etliche logische Schlussfolgerungen ziehen, d.h., wir verfügen mit (BKB) schon über einen recht beachtlichen Bestätigungsbegriff. Die Frage ist aber, ob diese Axiome als Rationalitätsanforderungen an eine komparative Bestätigungsbeziehung zu rechtfertigen sind. Dazu müssen wir uns im Einzelnen damit beschäftigen, wie intuitiv die jeweiligen Forderungen sind. Die Regel (1) dient dazu festzulegen, dass der Begriff der komparativen Bestätigung zumindest die grundlegenden Beziehungen zwischen Kontradiktionen und Tautologien respektiert. Das dürfte kaum zu bestreiten sein. Regel (2) legt fest, dass die Kontradiktionen den »Boden« unserer Vergleiche bilden, während in (7) festgelegt wird, dass die Tautologien jedenfalls (vielleicht mit anderen Aussagen) ganz oben stehen. Allerdings impliziert (7) zugleich, dass wir uns der Tautologien gewiss sind. Diese Art von logischer Allwissenheit ist natürlich keine realistische Forderung für reale epistemische Subjekte, sondern nur eine ideale Zielvorstellung für unser Konzept der rationalen komparativen Bestätigung. In (3) bestimmen wir nur, wie wir uns die komparative Bestätigung denken, nämlich als gleichstark bestätigte Aussagen einbeziehend. Die Regel (4) ist sicher schon etwas stärker, sollte

aber für einen Bestätigungsbegriff trotzdem relativ unproblematisch sein. Sonst hätten wir keinen komparativen Begriff mehr. Gehen wir aber davon aus, dass es ein komparatives Konzept von Bestätigung gibt, was zunächst recht plausibel erscheint, dann sollten wir wenigstens dessen Transitivität annehmen. Die Regeln (5.1) und (5.2) bestimmen den Begriff der Gewissheit noch weiter. Wenn wir uns der logischen Äquivalenz zweier Aussagen gewiss sind, dann dürfen wir die zumindest in einfachen Kontexten füreinander ersetzen. So bringen wir grundlegende Konzepte der komparativen Bestätigung auf intuitive Weise zusammen mit logischen Grundkonzepten. Gewisse Aussagen werden dabei schlicht als Basis für weitere Schlussfolgerungen betrachtet – ähnlich wie das bei Wissen der Fall ist.

Die Bedingungen (6.1) und (6.2) sehen etwas komplizierter aus, sind aber wieder ähnlich wie im Falle der vorigen Bedingungen zu verstehen. Zu (6.1): Wenn wir schon zu wissen glauben, dass E sowohl mit A wie auch mit B inkompatibel ist und $(A \vee E) \succeq (B \vee E)$, dann darf der mögliche Bestätigungsvorsprung des ersten Terms nur auf einen möglichen Vorsprung von A gegenüber B zurückzuführen sein. Das ist eine wesentliche Bedingung für den komparativen Bestätigungsbegriff, der eine Verbindung zum logischen »oder« zeigt. In (6.2) wird die umgekehrte Beziehung gefordert. Das disjunktive Hinzufügen von inkompatiblen Aussagen sollte die Richtung der komparativen Bestätigung nicht beeinträchtigen. Das kann man sich vielleicht am besten an einem Beispiel klarmachen.

Nehmen wir an, wir möchten für eine seltsam verbogene Münze bestimmen, welches die Wahrscheinlichkeit p für Kopf ist. Wir gehen hier davon aus, dass es sich um einen indeterministischen Prozess mit einem festen p handelt. Unsere drei Aussagen geben nun jeweils ein Intervall an, in dem sie p vermuten: A sagt in $I_1 = [0{,}2;0{,}4]$, B sagt in $I_2 = [0{,}3;0{,}6]$, E sagt in $I_3 = [0{,}6;0{,}8]$. Die Inkompatibilitätsbedingungen sind offensichtlich erfüllt. Dann sollte auch intuitiv gelten: $(A \vee E) \succeq (B \vee E)$ liegt genau dann vor, wenn $A \succeq B$ gegeben ist. Betrachten wir zunächst den Fall, dass $(A \vee E) \succ (B \vee E)$ ist. Dann muss diese bessere Bestätigung einen Grund haben, und der kann nur darin zu suchen sein, dass es uns plausibler erscheint, dass p in I_1 liegt als in dem Intervall I_2, etwa weil wir bisher nur kleine relative Häufigkeiten von Kopf beobachten konnten.

Sind wir uns dagegen schon ganz sicher, dass E zutrifft, könnte dieser Unterschied nicht auftreten. In jedem Fall ist klar, wenn (A∨E) = (B∨E) gilt, dann dürfen A und B keinen solchen Unterschied in der Bestätigung aufweisen, sonst wäre etwa p eher im ersten Intervall als im zweiten zu vermuten und damit müsste wieder gelten: (A∨E) ≻ (B∨E). Also sollte auch A = B sein. In der anderen Richtung scheinen uns die Beziehungen ähnlich plausibel zu sein, so dass die Axiome (6) durchaus sinnvolle Anforderungen an eine komparative Bestätigung sind. Übrigens sind die folgenden ebenso naheliegenden Anforderungen nun ableitbar und werden daher nicht als Axiome eingeführt:

(a) gewiss(A→B), dann gilt A ≽ B, und
(b) A ≽ B, dann gilt ¬B ≽ ¬A

Weiterhin ist klar, dass jede Wahrscheinlichkeitsverteilung P auf L eine solche basale komparative Bestätigungsbeziehung (BKB) impliziert: Erinnern wir uns, dass wir von einer derartigen Wahrscheinlichkeitsverteilung verlangen, dass für alle Q und R gilt: (1) P(R) ≥ 0, (2) P(t) = 1 für Tautologien t und schließlich die einfache Additivität (3) falls R und Q inkompatibel sind: P(Q∨R) = P(Q) + P(R). Daraus folgen die Axiome, wenn wir aus dem »≥« für die Wahrscheinlichkeiten das »≥« für die Aussagen gewinnen. (Der Leser kann hier aus dem Kontext hoffentlich leicht ermitteln, welche der beiden Relationen gemeint ist, obwohl dasselbe Zeichen für beide Relationen verwandt wird.)

Für P(A) ≥ P(B) setzen wir also: A ≽ B,

dann erfüllt eine so definierte Beziehung »≽« offensichtlich die Axiome von (BB). Allerdings leistet P natürlich noch viel mehr. Es liefert einen Vergleich der Bestätigung bzw. Plausibilität für *alle* Paare von Aussagen Q und R. Um ein Repräsentationstheorem zu erhalten, wird man also (BKB) verstärken müssen. Wir suchen nach Anforderungen, so dass wir schließen dürfen, dass es zu einer komparativen Bestätigungsbeziehung »≽« mindestens eine *passende* Wahrscheinlichkeitsverteilung P gibt, mit:

(PW) P(A) ≥ P(B) genau dann, wenn A ≽ B

Solche passenden Wahrscheinlichkeiten P lassen sich zu einer BKB »≥« natürlich nur finden, wenn die BKB zumindest in konsistenter Weise zu einer vollständigen Ordnung erweiterbar ist. Eine *vollständige* komparative Bestätigungsbeziehung (VKB) ist eine BKB, bei der das schwache bisherige Reflexivitätsaxiom (3) nun zu einem Vollständigkeitsaxiom (3*) verstärkt wurde:

(3*) Für alle A und B gilt: $A \succeq B$ oder $B \succeq A$

Hawthorne (2009) argumentiert nun dafür, dass wir nur dann eine BKB als *rational* bezeichnen sollten, wenn sie zumindest zu einer VKB erweitert werden *kann*. Dabei kann es möglicherweise unterschiedliche Erweiterungen geben. Gibt es hingegen überhaupt keine solche Erweiterung, so versteckt sich bereits eine Art von Inkonsistenz in unserer ursprünglichen BKB und wir sollten zu einer kohärenten neuen BKB wechseln, die zumindest im Prinzip erweiterbar ist. Das scheint mir recht überzeugend.

Leider genügt selbst das noch nicht, um eine passende Wahrscheinlichkeitsverteilung zu erhalten, sondern es muss noch eine weitere starke Annahme darüber hinzukommen, dass die Ordnung durch die neue VKB auch noch *ordnungsdicht* ist, d.h., dass es zu zwei Aussagen A und B mit $A \succ B$ immer auch noch eine dazwischenliegende Aussage C gibt mit $A \succ C \succ B$. Wir verlangen noch etwas mehr, um schon eine Art von Maßstab auf unserer Menge von Aussagen zu gewinnen, nämlich, dass es eine *n-stufige gleichplausible Zerlegung* mit bestimmten Eigenschaften gibt, mit deren Hilfe wir unsere Aussagen letztlich einstufen können. Dafür wird zur Not für die Einbettung der BKB in eine entsprechende VKB noch eine Erweiterung der Sprache L zu einer Sprache L* zugelassen.

Wir sagen dann, eine BKB $\succeq$ ist *geeignet erweiterbar*, wenn es eine Erweiterung $\succeq^*$ in einer Sprache L* gibt, mit den folgenden Eigenschaften:

Vollständig erweiterbare Bestätigungsbeziehung (VEB)
Die BKB »≥« ist eine VEB gdw. es eine BKB »$\succeq^*$« in einer erweiterten Sprache L* gibt, die die folgenden Bedingungen für alle A und B aus L erfüllt:

(1) Für alle A und B aus L: $A \succeq B \Rightarrow A \succeq^* B$ (*Erweiterung*)
(2) Für alle A und B aus L*: $A \succeq^* B$ oder $B \succeq^* A$ (*Vollständigkeit*)

(3) Für alle A und B aus L* gibt es ein $n \in \mathbb{N}$ mit: Wenn $A \succ^* B$, dann gibt es eine *n-stufige gleichplausible Zerlegung* $Z = (R_1,\ldots,R_n)$ bzgl. $\succeq^*$, für die gilt: $A \succ^* (B \vee R_i)$ für alle i. (*n-Zerlegbarkeit*)

Was ist nun mit einer solchen Zerlegung gemeint, die schon auf Savage zurückgeht? Es ist eine Einteilung für unsere Aussagen in L*, die sehr fein ist; so fein, dass der »Abstand« zwischen A und B jeweils »größer« ist als die der einzeln R_i. Dabei ist eine Zerlegung Z eine Menge von Aussagen mit den folgenden Eigenschaften:

Z ist eine n-stufige gleichplausible Zerlegung in L* **bzgl.** $\succeq^*$ gdw.

(1) $Z = (R_1,\ldots,R_n)$ und
(2) die R_i sind *gegenseitig inkompatibel* (gewiss*$(\neg(R_i \& R_j))$ für alle ungleichen i und j)
(3) die R_i sind *zusammen erschöpfend* bzw. gewiss (gewiss*$(R_1 \vee \ldots \vee R_n)$)
(4) die R_i sind *gleichplausibel*: $R_i =^* R_j$ für alle i und j.

Typische Beispiele für solche Zerlegungen wären etwa Lotterien mit n Losen, von denen nur eines gewinnt. R_i ist dann die Aussage, dass das i-te Los gewinnt. Bei einer fairen Lotterie sind die Bedingungen für eine Zerlegung erfüllt. Um die Zerlegungen so fein zu gestalten, dass die Bedingung $A \succ^* (B \vee R_i)$ erfüllt ist, müssen wir n nur groß genug wählen. Auch unser Beispiel mit den Intervallen lässt sich dazu ausbauen. Wir benötigen eine Zerlegung in n-Intervalle, die alle dieselbe Plausibilität dafür aufweisen, dass p dort hineingehört. Bei einer Gleichverteilung auf dem Einheitsintervall würde man gleichlange Intervalle wählen, allgemeiner gilt das aber natürlich nicht. Die Intervalle müssen nur jeweils als gleichplausibel ausgewählt werden, und indem wir sie verkürzen, können wir auch eine beliebig feine Verteilung erreichen, so dass $A \succ^* (B \vee R_i)$ für alle i erfüllt ist. Ist dabei das zu R_i gehörige Intervall in B enthalten, d.h., folgt R_i logisch aus B, so gilt gerade $(B \vee R_i) =^* B$, aber es gibt natürlich auch Intervalle, für die das nicht der Fall ist (schließlich ist B nicht gewiss), und daher gilt für mindestens ein j: $(B \vee R_j) \succ^* B$, weshalb wir nun zugleich die Ordnungsdichtheit gewährleistet haben. Außerdem liefern uns die Zerlegungen insgesamt eine Art von Maßstab

für die Wahrscheinlichkeiten. Sie können zumindest eine erste Einteilung und Einordnung der anderen Aussagen vornehmen, die schließlich in der Hierarchie zwischen bestimmten Disjunktionen von R_i's angesiedelt sind.

Hawthorne (2009) gibt dazu zwei *Theoreme* an. Erstens gilt, dass jede Wahrscheinlichkeitsverteilung auf L eine BKB induziert, die eine vollständig erweiterbare Bestätigungsbeziehung darstellt. Zweiten gilt ein *Repräsentationstheorem* in der anderen Richtung: Zu jeder BKB »≥«, die zu einer vollständigen Bestätigungsbeziehung erweiterbar ist, die die Axiome für VEBs erfüllt, gibt es mindestens eine Wahrscheinlichkeitsverteilung, die sie induziert. Man könnte nun sagen, die Menge aller zu »≥« passenden Wahrscheinlichkeitsverteilungen repräsentiert genau den epistemischen Zustand, der durch die BKB »≥« gegeben ist. Die Vagheiten, die in dem Zustand enthalten sind, werden dadurch wiedergegeben, dass wir eine Menge von Verteilungen betrachten. Damit verliert das Argument, dass *eine* solche Verteilung übergenau sei in der Repräsentation unseres epistemischen Zustands etwas an Kraft.

Wir können dann *Eindeutigkeit* für die Wahrscheinlichkeitsfunktion P erreichen, wenn die Axiome für VEBs bereits für unsere Relation »≥« gelten. Allerdings müsste unser epistemischer Zustand dann schon so fein ausdifferenziert sein, dass es zu je zwei Aussagen mit $A \succ B$, beliebig feine Zerlegungen gibt. Das ist eine enorm idealisierende Vorstellung und es ist alles andere als klar, welche intuitiven Anforderungen damit für »≥« außer der zunächst genannten Konsistenzforderung für unsere Überzeugungssysteme entstehen. Die Konsistenzbedingung sorgt zunächst nur dafür, dass wir die Aussagen aus L irgendwie im Einheitsintervall ansiedeln können. Dabei können aber noch alle als sehr pausibel oder als sehr unplausibel eingestuft werden. Erst die Einführung der Zerlegungen als eine Art von Maßstab sorgt dann für eine sinnvolle Verteilung auf die konkreten Werte im Einheitsintervall. Die komplexe Konstruktion belegt geradezu, wie groß der Schritt von einer komparativen Plausibilitätseinstufung zu einer absoluten Skala wie der Wahrscheinlichkeitsskala noch ist.

5.3.8 Hawthornes n-stufige Glaubenslogik

Interessant an der ganzen Konstruktion ist auch noch eine Verbindung zum kategorischen Glauben g, den Hawthorne hier im Sinne der Lockeschen Idee findet. Diese Verbindung wird so gestaltet, dass sie eine erste Antwort auf das Lotterie-Paradox anzubieten hat. Deshalb möchte ich sie zumindest kurz besprechen. Dazu müssen wir zunächst zwei sinnvolle Anforderungen stellen:

A8. Wenn gilt: gewiss(A), dann gilt: g(A)
(Gewissheit führt zu Glaube) und
A9. Wenn gilt: $A \geq B$ und g(B), dann gilt: g(A)
(Glaube besitzt einen Schwellenwert)

Damit gelangen wir schließlich zu einer Schwellenwertkonzeption des Glaubens für eine Wahrscheinlichkeitsverteilung P. Die könnte eine der folgenden Formen annehmen:

Schwellenwertkonzeption kategorischen Glaubens
(i) g(R) gdw. $P(A) > q$ oder
(ii) g(R) gdw. $P(A) \geq q$

Dazu schlägt Hawthorne (2009) eine interessante Zusatzregel für unsere komparative Plausibilitätsrelation vor:

A10. (n-Regel): Für alle $A_1,\ldots,A_n$ gilt: Wenn gewiss$(A_1 \vee \ldots \vee A_n)$, dann gilt nicht: $g(\neg A_1)$ &…& $g(\neg A_n)$

Das bedeutet, wenn man sich einer n-Disjunktion sicher ist, darf man höchstens n-1 der Negationen der Disjunkte glauben. Das vermeidet zumindest entsprechende Widersprüchlichkeiten bis zur Stufe n. Für Disjunktionen mit mehr als n Aussagen sagt das allerdings nichts.

Wählen wir nun $q > (n\text{-}1)/n$, so erfüllt die Schwellenwertkonzeption des Glaubens automatisch die n-Regel. Wenn etwa gilt gewiss$(A_1 \vee \ldots \vee A_n)$, so muss $P(A_1 \vee \ldots \vee A_n) = 1$ sein, da Gewissheit ebenso große Wahrscheinlichkeit verlangt wie die einer Tautologie. Nun ergibt ein wenig Rechnen:

$1 = P(A_1 \vee \ldots \vee A_n) \leq P(A_1) + \ldots + P(A_n) = [1 - P(\neg A_1)] + \ldots + [1 - P(\neg A_n)]$
$= n - \sum P(\neg A_i)$, also: (*) $\sum P(\neg A_i) \leq n-1$.

Das wäre aber nicht erfüllbar, wenn für all i gelten würde: $P(\neg A_i) \geq q > (n-1)/n$, denn dann wäre $\sum P(\neg A_i) \geq nq > n\ (n-1)/n = n$ im Widerspruch zu (*).

Also könnten wir nicht alle $\neg A_i$ im Sinne der Schwellenwertkonzeption glauben, wenn wir uns der Disjunktion $A_1 \vee \ldots \vee A_n$ sicher wären, gerade so, wie es die n-Regel verlangt.

Angewandt auf das Vorwort-Paradox (mit A_i = i-te Aussage ist falsch) bedeutet das: Wenn wir uns sicher sind, dass bei n Aussagen unseres Buches eine falsch ist, so dürfen wir nicht mehr alle Aussagen glauben. Bis zu n Aussagen sind wir durch die n-Regel also vor dem Vorwort-Paradox geschützt, bei einem Schwellenwert von $q>(n-1)/n$, allerdings nicht mehr darüber hinaus. Für mehr als n Aussagen kann es wieder auftreten.

Außerdem soll eine entsprechende Regel für Konjunktionen gelten:

A11. (n*-Regel): Für alle $A_1, \ldots, A_{n+1}$ gilt: Wenn für alle $i \neq j$ gilt: gewiss$(\neg[A_i \& A_j])$, dann gilt: $g(\neg A_1) \& \ldots \& g(\neg A_{n+1})$

Wenn wir uns für die n+1 Aussagen bereits sicher sind, dass höchstens eine davon wahr sein kann, dann sollten wir wenigstens eine der Negationen der Aussagen glauben. Auch das ist eine weitere Regel zur Vermeidung von Widersprüchlichkeiten zwischen »gewiss« und »g«.

Mit Hilfe der Axiome A1-A11 definiert Hawthorne (2009) dann seine n-stufige Glaubenslogik:

Basale n-stufige Glaubenslogik (nach Hawthorne)
Ein Paar von Operatoren $(\geq, g)_n$ definiert über einer Sprache L, das die Axiome A1-A11 erfüllt, heiße eine basale n-stufige Glaubenslogik.

Hawthorne zeigt dafür, dass man mit dieser Glaubenslogik Widersprüche wie im Lotterieparadox oder im Vorwortparadox bis zur Stufe n vermeiden kann. Außerdem gibt es hier auch wieder passende Wahrscheinlichkeitsverteilungen P zu solchen Glaubenslogiken. Wählen wir etwa:

g(A) gdw. P(A) ≥ q, mit n/(n+1) > q > (n–1)/n,

so ergibt sich eine basale n-stufige Glaubenslogik. Ist umgekehrt eine solche Logik wieder konsistent erweiterbar (in ähnlicher Weise, wie wir das schon für die BKBs kennengelernt haben), so ist sie auf entsprechende Weise durch eine Wahrscheinlichkeitsverteilung P repräsentierbar. Hawthorne stellt also auf recht komplizierte Weise wieder einen Zusammenhang zwischen Wahrscheinlichkeitsverteilungen und Glaubenslogiken her, der einige der Vorteile der probabilistischen Behandlung der Paradoxien (bis zur Stufe n) erhalten kann.

Das ist ein hilfreicher erster Schritt zur Untersuchung des Zusammenhangs der beiden Ansätze, aber sicher noch nicht das Ende der Debatte. So wird u.a. deutlich, dass die Stufe der Glaubenslogik von der jeweiligen Sprache abhängt. Ein Wechsel der Sprache etwa zu einer feineren Beschreibung derselben Sachverhalte (man denke nur an mein Beispiel der Intervalle, in dem offensichtlich ist, wie wir aus einer Aussage mehrere machen können) lässt die grundlegenden Zusammenhänge leider nicht invariant. Trotzdem zeigen sich hier interessante Zusammenhänge zwischen beiden Ansätzen und wir können schon erkennen, dass die nicht einfach beschaffen sind, auch wenn wir uns wieder einer Schwellenwertkonzeption verschrieben haben. Außerdem sind der komparative Bestätigungsbegriff und die darauf aufbauende Glaubenslogik auch für sich genommen interessante Vorstufen zu probabilistischen Ansätzen, die es weiter zu erkunden gilt. Insbesondere belegen die Überlegungen aber auch, was alles in einem probabilistischen Überzeugungssystem enthalten ist. Es ist keineswegs harmlos oder inhaltsarm, von einer probabilistischen Bewertung all unserer Überzeugungen auszugehen. Das ist im Gegenteil ein anspruchsvoller Schritt über den klassischen Ansatz hinaus und selbst für einen komparativen Ansatz sind einige Anforderungen zu erfüllen, bis wir sicherstellen können, dass er zumindest eine probabilistische Repräsentation besitzt.

5.3.9 Bayessche Netze

Ein schwerwiegendes Problem des probabilistischen Ansatzes war die unübersehbar große Anzahl an Glaubensgraden bzw. subjektiven Wahrscheinlichkeiten, die wir benötigen, um ein solches Überzeugungssystem

vollständig zu beschreiben. Eine Frage ist deshalb, wie wir etwa bestimmtes schon vorhandenes Wissen z.B. über kausale und andere inferentielle Zusammenhänge einbringen können, um die Anzahlen wieder deutlich zu verringern.

Probabilistische Überzeugungssyteme werden dazu oft mit einer gewissen Struktur gedacht. Wir haben nicht nur eine amorphe Masse von Aussagen vor uns mit einer gemeinsamen Wahrscheinlichkeitsverteilung darauf, sondern eine komplexere Struktur von speziellen Beziehungen und dementsprechenden (probabilistischen) Abhängigkeiten und Unabhängigkeiten zwischen diesen Aussagen. Dabei haben sich vor allem *bayesianische Netze* (bzw. *bayessche Netze*) als Modell für diese Zusammenhänge durchgesetzt, auf die wir im Rahmen des kausalen Schließens noch zurückkommen werden. Jedenfalls hat es sich als sehr hilfreich erwiesen, die (kausalen) Beziehungen zwischen bestimmten Faktoren (oder Aussagen) und damit auch die Unabhängigkeitsbeziehungen durch entsprechende Graphen darzustellen. Auch das Problem, überhaupt erst einmal konsistente Wahrscheinlichkeiten für eine Menge von Aussagen zu erhalten, lässt sich mit Hilfe der bayesschen Netze lösen (s. Williamson 2005).

Bayessche Netze enthalten also zunächst einen Graphen G. Ein Graph G = <V, E> besteht aus einer Menge V von Knoten oder Ecken (vertices) und einer Menge E von Kanten (edges), die wir formal etwa durch geordnete Paare (v,w) von Knoten beschreiben können, wenn die Kanten gerichtet sind und z.B. durch Doppelpaare (v,w) und (w,v), wenn die Kanten ungerichtet sind. Anschaulich lässt sich die Konzeption von Graphen leicht verstehen und die gerichteten Kanten sollen inhaltlich etwa die Richtung von direkten Kausalbeziehungen (oder eventuell auch anderen Begründungszusammenhängen) wiedergeben. Allgemein werden die Aussagen durch die Knoten des Graphen oder Netzes angegeben und die Abhängigkeiten als die Kanten des Graphen. Dann lassen sich selbst komplexere Unabhängigkeitsbeziehungen als topologische Eigenschaften des Graphen verstehen. Man spricht zu diesem Zweck davon, dass im Zusammenhang des Graphen bestimmte Knoten andere Knoten voneinander trennen oder *separieren* (vgl. etwa Beierle & Kern-Isberner 2008, Neapolitan 2004, 2009, Korb & Nicholson 2010). Dadurch lassen sich z.B. kausale Unabhängigkeitsbeziehungen

anschaulich darstellen. Kausalbeziehungen werden typischerweise durch einen *gerichteten* Graphen wiedergegeben, bei dem die Kanten jeweils orientiert sind und eine Ursache-Wirkungsbeziehung repräsentieren.

Jedenfalls stellt man sich bestimmte Zusammenhänge zwischen Aussagen – gerade im Bereich der Wissenschaften – als eine *gerichtete Struktur* vor, bei der auch die Begründungsbeziehungen in einer Richtung verlaufen und sich etliche statistische Unabhängigkeiten dadurch ergeben. So sinkt die Informationsmenge bzw. Menge an Wahrscheinlichkeiten, die wir zur Verfügung haben müssen, um über alle Zusammenhänge im Netz Bescheid zu wissen, dramatisch. Das lässt das ganze Modell wieder etwas humaner und realistischer erscheinen, denn ich hatte bereits darauf hingewiesen, dass wir für n Aussagen ansonsten über 2^n Wahrscheinlichkeiten verfügen müssten, was sich schon für wenige Aussagen als extrem große Menge an Informationen entpuppt. Im einfachsten Fall ausschließlich probabilistisch unabhängiger Aussagen lassen sich etwa alle Konjunktionen durch die entsprechenden Produkte von Wahrscheinlichkeiten berechnen und wir benötigen nur noch n Wahrscheinlichkeiten.

Das bayessche Netz N = <P, G> besteht dann aus einem Graphen G und einer dazu passenden Wahrscheinlichkeitsverteilung P auf den Knoten des Graphen. Für ein Netz von n Knoten bzw. Aussagen, von denen jeder höchsten drei Eltern hat, benötigen wir für jeden Knoten höchstens 2^3 bedingte Wahrscheinlichkeiten und daher höchstens 8n Wahrscheinlichkeiten, um damit die Wahrscheinlichkeitsverteilung für unsere n Aussagen festzulegen. Für n = 10 benötigen wir dann weniger als 80 Wahrscheinlichkeiten statt der 1023 für eine völlig amorphe Menge von Aussagen und für n = 20 nur 160 statt über eine Million Wahrscheinlichkeiten, für n = 30 nur 240 statt über eine Milliarde Wahrscheinlichkeiten usf. (Die Zahlenwerte ergeben sich aus der Kettenregel für bayessche Netze s.u.) Die jeweilige Ersparnis ist gewaltig und bringt uns erst in den Bereich von tatsächlich noch beherrschbaren Größenordnungen. Damit eröffnet sich erst durch die bayesianischen Netze die Möglichkeit, ein wahrscheinlichkeitstheoretisches Modell zu entwerfen, mit dem wir in der Praxis arbeiten können.

Üblicherweise dienen bayessche Netze dazu, bereits bekannte kausale (oder andere inferentielle) Zusammenhänge zu modellieren und man

denkt sich dann die Begründungszusammenhänge bzw. den Informationsfluss im Netz als parallel zu den kausalen Zusammenhängen gestaltet. Zur Vereinfachung stellt man sich die kausale Struktur dabei als gerichtetes Netz vor, das keine Rückkopplungen oder Schleifen enthält, das also *azyklisch* ist. Diese Netze werden auch als DAGs bezeichnet für »directed acyclic graph«. Die probabilistischen Unabhängigkeitsbeziehungen können wir dann anhand der sogenannten *d-Separation* (d für »directed«) aus der Topologie des Netzes ableiten (s. ausführlicher Kap. 7.3.8) und somit die Anzahl der benötigten Wahrscheinlichkeiten erheblich reduzieren.

Für eine Reihe von Anwendungen konnte man auf diesem Wege erfolgreiche Expertensysteme aufbauen (vgl. Beierle & Kern-Isberner 2008, Neapolitan 2004, Korb & Nicholson 2010). Diese bayesianischen Netze stellen letztlich das bayesianische Schließen für den etwas komplexeren Fall als den im Beispiel des Dschungelfiebers dar, und wir werden uns das an einem Beispiel von Neapolitan (2004, 4 ff.) kurz anschauen. Doch dazu benötigen wir zunächst noch einige formale Konzepte, die dabei zum Einsatz kommen.

Zufallsvariablen (im Unterschied zu reinen Aussagenvariablen, die nur zwei Werte annehmen vgl. dazu Beierle & Kern-Isberner 2008, Anhang A) sind zunächst Funktionen $X:\Omega\rightarrow\mathbb{R}$ eines Grundraumes von Ereignissen Ω etwa in einen Zahlenraum $\mathbb{R}$ meist die Menge der reellen Zahlen oder in einen Raum, der auf andere Weise die Ausprägung bestimmter Eigenschaften darstellt. Die Zufallsvariablen dienen u.a. dem Zweck, bestimmte Ergebnisse von Zufallsprozessen zu beschreiben. So kann einem Wurf mit drei Würfeln z.B. die Summe der Augen der Würfel zugeordnet werden. Dann erhalten wir eine Wahrscheinlichkeitsverteilung für die Zufallsvariable X als Verteilung auf der Menge der Ergebnisse, die X annehmen kann. So schreiben wir etwa $(X = 3)$ für das Urbild der 3 unter X, also für die Menge aller Würfe mit insgesamt drei Augen: $\{\omega\in\Omega; X(\omega) = 3\}$. Da es dafür nur ein Elementarereignis gibt, gilt $P(X=3) = 1/216$, wenn es sich um faire Würfel handelt.

Oft bestimmen wir den Grundraum Ω aber überhaupt nicht mehr, sondern wenden uns direkt den Wahrscheinlichkeiten für die Werte der Zufallsvariablen X zu. Wir benötigen eigentlich nur die Wahrscheinlichkeiten dafür, dass X bestimmte Ergebnisse annimmt, um mit diesen

Zufallsvariablen rechnen zu können (oder die Wahrscheinlichkeiten für bestimmte Aussagen, die solche Ergebnisse repräsentieren). Dann sparen wir uns gerne die Konstruktion eines ursprünglichen Grundraums Ω (vgl. Neapolitan 2004, Kap. 1.2.2), der oftmals nicht leicht zu konstruieren wäre. Die Zufallsvariablen stehen dann einfach für kausale Faktoren, die bestimmte Ausprägungen annehmen können. So könnte X für die Stärke oder Häufigkeit der Nebenwirkungen stehen oder für die Anzahl der eingenommenen Tabletten etc. und unsere Wahrscheinlichkeitsverteilung sagt uns, wie häufig diese Nebenwirkungen im Durchschnitt auftreten. Jedenfalls benötigen wir nur noch die Wahrscheinlichkeiten des Typs P(X=x), die danach fragen, wie wahrscheinlich es ist, dass unsere Zufallsvariable X genau den Wert x annimmt, ohne das über die Wahrscheinlichkeiten im Grundraum berechnen zu wollen.

Unsere bisherigen dichotomen Variablen (oder Aussagenvariablen) A sind eine Art von Sonderfall davon und lassen sich durch eine Zufallsvariable mit den Werten 0 und 1 charakterisieren. Die Wahrheit von A wird dann etwa durch X=1 und die Falschheit von A durch X=0 wiedergeben, aber wir wissen eben nicht genau, welcher Wert in unserem Fall zutrifft und das wird durch die Wahrscheinlichkeitsverteilung für X angegeben. Statt mit P(A) arbeiten wir dann mit P(X=1). Außerdem werden bestimmte Konstellationen von Zufallsvariablen $X_1,\ldots,X_n$ auf einfache Weise dargestellt, so schreiben wir vereinfachend:

$$P(x_1,\ldots,x_n) := P(X_1=x_1 \,\&\ldots\&\, X_n=x_n)$$

für die Wahrscheinlichkeit, dass unsere Zufallsvariablen gerade die Werte $x_1,\ldots,x_n$ annehmen, bzw. für die gemeinsame Verteilung der Zufallsvariablen $X_1,\ldots,X_n$. Insbesondere müssen wir nun nicht mehr nur einfache Korrelationen von Zufallsvariablen berücksichtigen, sondern haben auch eine komplexere Definition der statistischen Unabhängigkeit anwenden:

Danach sind zwei Zufallsvariablen X und Y *statistisch unabhängig*, wenn für *alle* Werte x und y aus ihren jeweiligen Wertebereichen gilt: $P(x|y) = P(x)$ bzw. äquivalent dazu: $P(x,y) = P(x)\cdot P(y)$, d.h., dass die Wahrscheinlichkeit mit der die eine Zufallsvariable bestimmte Werte annimmt, völlig unabhängig davon ist, welche Werte die andere gerade annimmt. Wir schreiben dafür auch: $X \perp Y$ (oder auch $P(X,Y) = P(X)\cdot P(Y)$)

Außerdem schreiben wir später auch noch $(X \perp Y|Z)$, wenn wir damit zum Ausdruck bringen möchten, dass X und Y statistisch unabhängig sind gegeben die Variable Z. Für jeden vorgegebenen Wert von Z sind also die Werte von X und Y statistisch unabhängig.

Damit es sich bei P um eine solche *gemeinsame Wahrscheinlichkeitsverteilung* der n (logisch unabhängigen) Zufallsvariablen $X_1,\ldots,X_n$ handelt, sind noch zwei Bedingungen zu erfüllen:

(1) $0 \leq P(x_1,\ldots,x_n) \leq 1$ für alle möglichen Werte $x_1,\ldots,x_n$
(2) $\sum_{\text{gesamt}} P(x_1,\ldots,x_n) = 1$

Dabei wird in (2) die Summe über alle Wertekombinationen $x_1,\ldots,x_n$ gebildet, die unsere Zufallsvariablen annehmen können. Nehmen wir als einfaches Beispiel zwei Zufallsvariablen X und Y, die einfach die Ergebnisse zweier Würfel darstellen. Sie können dann die 36 Kombinationen (1,1), (1,2),..., (6,6) annehmen und sollten für jede Kombination den Wahrscheinlichkeitswert 1/36 für zwei faire Würfel ergeben.

Wir können natürlich auch eine neue Zufallsvariable Z = X+Y konstruieren, die die Werte von 2 bis 12 annehmen kann und dann eine etwas komplexere Wahrscheinlichkeitsverteilung aufweist, die sich aber leicht ausrechnen lässt. So ist etwa P(3) = 1/18, weil es zwei Kombinationen von X und Y gibt, nämlich (1,2) und (2,1), die gerade als Summe 3 ergeben. Kennen wir die gemeinsame Verteilung von X und Y können wir leicht alle Wahrscheinlichkeiten im Zusammenhang mit diesen beiden Zufallsvariablen ausrechnen. Wie hoch ist etwa die Wahrscheinlichkeit P(x<3, y>4|y ist gerade)? Wenn y gerade ist, dann kann sich die vordere Bedingung nur auf die Fälle (1,6) und (2,6) beziehen und ergibt: 1/18.

Allgemeiner lassen sich also neue Zufallsvariablen einschließlich ihrer Wahrscheinlichkeitsverteilung konstruieren. Nehmen wir an, dass wir die 4 Zufallsvariablen X, Y, Z und W kennen, und die können jeweils bestimmte Werte x_i, y_j, z_k und w_l annehmen (die jeweiligen Wertebereiche sollen diskret und endlich sein, sind aber ansonsten beliebig zu denken). Außerdem kennen wir bereits die gemeinsame Verteilungsfunktion P für unsere Variablen, dann lassen sich damit z.B. die folgenden Wahrscheinlichkeiten ausrechnen:

(1) $P(z_k) = \sum_{i,j,l} P(x_i, y_j, z_k, w_l)$

Das heißt, die (marginale) Wahrscheinlichkeit dafür, dass Z gerade den Wert z_k annimmt, ergibt sich, indem wir über alle Werte der gemeinsamen Verteilung aufsummieren, in denen dieser Wert für Z realisiert ist. Das sind einfach alle disjunkten Fälle, in denen Z diesen Wert annimmt. Die zusammengenommen sind für uns alle Fälle, in denen z_k auftritt. (Das ist eine Variante des Theorems der totalen Wahrscheinlichkeit.) Entsprechend können wir auch alle bedingten Wahrscheinlichkeiten im Prinzip gemäß der klassischen Formel ausrechnen:

$$(2)\ P(y_j|z_k,w_l) = P(y_j,z_k,w_l)/P(z_k,w_l) = \sum_i P(x_i,y_j,z_k,w_l)/\sum_{i,j} P(x_i,y_j,z_k,w_l)$$

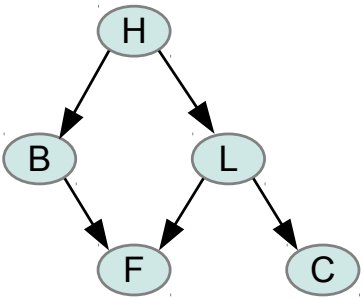
Umgekehrt können uns manchmal bestimmte bedingte Wahrscheinlichkeiten bereits genügen, um die gemeinsame Wahrscheinlichkeitsverteilung festzulegen, was wir uns an einem Beispiel von Neapolitan (2004, 4 ff.) anschauen wollen. Dazu betrachten wir ein kleines bayesianisches Netz (1) mit 5 Zufallsvariablen, die wir zunächst in einer Tabelle darstellen wollen:

Variable	Werte	Bedeutung der Werte
H	h1	Es gibt eine Rauchervorgeschichte.
	h2	Es gibt keine Rauchervorgeschichte.
B	b1	Bronchitis liegt vor.
	b2	Bronchitis liegt nicht vor.
L	l1	Lungenkrebs liegt vor.
	l2	Lungenkrebs liegt nicht vor.
F	f1	Müdigkeit liegt vor.
	f2	Müdigkeit liegt nicht vor.
C	c1	Die Röntgenaufnahme des Thorax ist positiv.
	c2	Die Röntgenaufnahme des Thorax ist negativ.

Tabelle 5.2: Ein kleines bayesianisches Netz

Dazu gibt es eine kleine kausale Geschichte, die auch die probabilistischen Beziehungen beschreibt. So erhöht die Rauchervorgeschichte das Risiko für Lungenkrebs und ebenfalls für Bronchitis. Sowohl Lungenkrebs wie auch Bronchitis führen häufig zu Müdigkeit, aber

nur der Lungenkrebs erhöht die Wahrscheinlichkeit für eine »positive« Röntgenaufnahme des Thorax. Damit können wir die Zusammenhänge als ein Netz (1) wie folgt darstellen:



Bayessches Netz 1: Lungenkrebs

Die hier auftretenden bedingten Wahrscheinlichkeiten liefern uns dann eine gemeinsame Wahrscheinlichkeitsverteilung unserer 5 Zufallsvariablen, weil die anderen möglichen Zusammenhänge als Fälle von probabilistischer Unabhängigkeit gedacht werden. Darauf kommen wir in Kapitel 7 im Zusammenhang mit Kausalschlüssen noch ausführlicher zurück. In unserem Beispiel starten wir mit einer »Vorher-Wahrscheinlichkeit« für H und dann entsprechenden bedingten Wahrscheinlichkeiten.

Wahrscheinlichkeiten in unserem Netz (1)		
P(h1) = 0,2	P(b1|h1) = 0,25	P(f1|b1,l1) = 0,75
P(h2) = 0,8	P(b1|h2) = 0,05	P(f1|b1,l2) = 0,10
	P(l1|h1) = 0,003	P(f1|b2,l1) = 0,5
	P(l1|h1) = 0,0005	P(f1|b2,l2) = 0,05
		P(c1|l1) = 0,6
		P(c1|l2) = 0,02

Tabelle 5.3: Wahrscheinlichkeiten für ein kleines bayesianisches Netz

Mehr als diese 12 (bedingten) Wahrscheinlichkeiten benötigen wir nun nicht mehr. (Für den Wurzelknoten könnten wir sogar noch eine einsparen.) Wenn etwa X z.B. ein direkter Vorfahr von Y im Netz ist, d.h. X ist direkte Ursache von Y, so geben die Werte jeweils an, wie

wahrscheinlich die verschiedenen »Ursachenwerte« x für X sind, und legen dann noch zusätzlich fest, mit welcher Wahrscheinlichkeit der Knoten Y in den jeweiligen Fällen bestimmte Werte y annimmt, wenn seine Ursache gerade x ist. Dadurch ist mit Hilfe der Kausalbeziehung jede Auftretenswahrscheinlichkeit im Netz eindeutig bestimmt.

Nur 12 Werte festlegen zu müssen, bietet immerhin schon eine deutliche Ersparnis gegenüber den $2^5 = 32$ Werten, die wir ohne unsere Struktur angeben müssten. Auf der Grundlage dieser bedingten Wahrscheinlichkeiten können wir dann weitere Wahrscheinlichkeiten gemäß unseren Formeln ausrechnen. Hilfreich ist dazu die allgemeine Kettenregel, die sich schnell aus der wiederholten Anwendung der Produktregel für Wahrscheinlichkeiten ergibt.

Kettenregel: $P(x_1, \ldots, x_n) = P(x_n|x_{n-1}, x_{n-2}, \ldots, x_1) \cdot \ldots \cdot P(x_2|x_1) \cdot P(x_1)$

Für *bayesianische Netze* gilt außerdem die Markov-Bedingung (vgl. auch Kap. 7.3.8 ff.), wonach die Wahrscheinlichkeiten für eine Variable nur von den bedingten Wahrscheinlichkeiten (gegeben sind dabei die Elternvariablen der Variablen) im Netz abhängig sind. Oder anders ausgedrückt: Die Eltern pa_j (mit pa für *parents*) der Variablen X_j schirmen die Variable X_j statistisch von allen anderen Variablen ab (außer von den Nachkommen von X_j). Das heißt, es gilt dann die folgende Regel, die bayesianische Netze charakterisiert, die sich durch geschickte Anordnung der Faktoren mit Hilfe der Kettenregel ableiten lässt:

Markov Bedingung:
$P(x_n, \ldots, x_1) = P(x_n|pa_n) \cdot P(x_{n-1}|pa_{n-1}) \cdot \ldots \cdot P(x_1|pa_1)$,

für den Fall, dass die bedingten Wahrscheinlichkeiten auf der rechten Seite alle existieren.

Diese Forderung entstammt unserer Kausalvorstellung, wonach bereits die direkten Ursacheneiner Wirkung W diese festlegen und die indirekten Ursachen nur über die direkten ihre Wirkung entfalten. Das scheint uns für kausale Beziehungen geradezu selbstverständlich zu sein. Wenn wir alle direkten Ursachen U einer Wirkung W kennen, dann sollte uns gemäß der Markov Bedingung die Kenntnis der vorhergehenden Ursachen V von W keine zusätzlichen Informationen über das Auftreten

von W mehr geben. Wir können dann ausrechnen, mit welcher Wahrscheinlichkeit die entsprechenden Wirkungen eintreten. Das entspricht ziemlich gut unserer intuitiven Vorstellung von kausalen Strukturen. Damit erhalten wir:

> **Bayessche Netze B = (G,P)** sind gerichtete azyklische Graphen G mit Zufallsvariablen als Knotenpunkten und einer gemeinsamen Wahrscheinlichkeitsverteilung P darauf, die die Markov-Bedingung erfüllt, d.h., dass die Unabhängigkeitsbeziehungen im Graphen G von der Verteilungsfunktion P respektiert werden. (Weiteres im Rahmen der Debatte um Kausalität s. Kap. 7)

Ein bayessches Netz ist also ein azyklischer gerichteter Graph mit einer *dazu passenden* gemeinsamen Wahrscheinlichkeitsverteilung auf den Knoten, die uns bereits optisch die probabilistischen Unabhängigkeitsbeziehungen von P erkennen lässt. Für den einfachen Fall eines kleinen Netzes (2) X→Y→Z ergibt sich damit:

(für Netz 2) $P(x, y, z) = P(z|y) \cdot P(y|x) \cdot P(x)$

Im Falle unseres Netzes (1) ergibt sich:

(für Netz 1) $P(f,c,b,l,h) = P(f|b,l) \cdot P(c|l) \cdot P(b|h) \cdot P(l|h) \cdot P(h)$,

Jetzt beherrschen wir alle Informationsflüsse in unseren Netzen und können beliebige Wahrscheinlichkeiten darin im Prinzip berechnen. Wenn wir etwa berechnen möchten, wie hoch die Wahrscheinlichkeit für Lungenkrebs ist, für jemanden, der eine Rauchervorgeschichte hat und ein positives Ergebnis beim Röntgen, so können wir das im Prinzip nun ausrechnen. Allerdings sind immer noch sehr viele Rechenschritte nötig.

In Neapolitan (2004) und Beierle & Kern-Isberner (2008) werden einige Algorithmen vorgestellt, mit denen sich die Anzahl der Rechenschritte weiter reduzieren lässt. Dabei werden z.B. mehrere Knoten zusammengenommen, um die Komplexität des Netzes zu reduzieren. Leider bleiben die Verfahren für tatsächliche Berechnungen trotzdem noch recht umständlich, weshalb man typischerweise eine entsprechende Software für bayesianische Netze zu diesem Zweck einsetzt (vgl. a. Korb

& Nicholson 2010). Der Leser kann sich einmal an der Berechnung von $P(l_1|h_1,c_1)$ versuchen, um die Berechnungsproblematik ernst zu nehmen. Uns genügt es aber hier zu zeigen, wie wir im Prinzip solche Berechnungen durchführen können.

Wie sich die Berechnungen tatsächlich für einfache Netze vereinfachen, mag das folgende Beispiel von Neapolitan (2009, 100 ff.) ein wenig veranschaulichen (Netz 3):

X	→	Y	→	Z	→	W
$P(x_1) = 0{,}4$		$P(y_1\|x_1) = 0{,}9$		$P(z_1\|y_1) = 0{,}7$		$P(w_1\|z_1) = 0{,}5$
		$P(y_1\|x_2) = 0{,}8$		$P(z_1\|y_2) = 0{,}4$		$P(w_1\|z_2) = 0{,}6$

Netz 3: eine einfache (kausale) Kette

Dabei gehen wir davon aus, dass jede Zufallsvariable jeweils zwei Werte annehmen kann. Dann sind das die einzigen Wahrscheinlichkeiten, die wir benötigen, da sich die restlichen Werte wie etwa $P(x_2) = 1–P(x_1)$ jeweils daraus ergeben. Außerdem können wir nun leicht ermitteln, wie der Informationsfluss im Netz aussieht. In Pfeilrichtung können wir von *kausalen Schlüssen* sprechen und in der Gegenrichtung spricht man von *diagnostischen Schlüssen*. Zunächst bietet es sich an, die unbedingten Wahrscheinlichkeiten der Knoten mit Hilfe des Theorems der totalen Wahrscheinlichkeit zu berechnen:

$P(y_1) = P(y_1|x_1)\cdot P(x_1) + P(y_1|x_2)\cdot P(x_2) = 0{,}9\cdot 0{,}4 + 0{,}8\cdot 0{,}6 = 0{,}84$
und entsprechend
$P(z_1) = P(z_1|y_1)\cdot P(y_1) + P(z_1|y_2)\cdot P(y_2) = 0{,}652$
$P(w_1) = P(w_1|z_1)\cdot P(z_1) + P(w_1|z_2)\cdot P(z_2) = 0{,}5348$

Als nächstes können wir nun verfolgen, wie sich bestimmte Informationen im Netz ausbreiten. Nehmen wir etwa an, wir erfahren, dass für X gerade x_1 realisiert wurde. Was bedeutet das für die Wahrscheinlichkeit der anderen Faktoren? Was passiert etwa für Y und was für Z und W? Das berechnen wir mit Hilfe der Kettenregel und der Markov-Bedingung:

$P(y_1|x_1) = 0{,}9$
$P(z_1|x_1) = P(z_1|y_1,x_1)\cdot P(y_1|x_1) + P(z_1|y_2,x_1)\cdot P(y_2|x_1) =$
$P(z_1|y_1)\cdot P(y_1|x_1) + P(z_1|y_2)\cdot P(y_2|x_1) = 0{,}67$

$P(w_1|x_1) = P(w_1|z_1,x_1)\cdot P(z_1|x_1) + P(w_1|z_2,x_1)\cdot P(z_2|x_1) =$
$P(w_1|z_1)\cdot P(z_1|x_1) + P(w_1|z_2)\cdot P(z_2|x_1) = 0{,}533$

Im ersten Berechnungsschritt nutzen wir nur die Regel der totalen Wahrscheinlichkeit, während wir uns im zweiten dann auf die Markov-Bedingung für Aussagen A, B und C stützen:

Die eingesetzten Rechenregeln:
(1) $P(A|C) = P(A|C,B)\cdot P(B) + P(A|C,\neg B)\cdot P(\neg B)$ und dann
(2) $P(A|C,B)\cdot P(B) + P(A|C,\neg B)\cdot P(\neg B) = P(A|B)\cdot P(B) + P(A|\neg B)\cdot P(\neg B)$,
falls A unabhängig ist von C gegeben B.

Dabei ergibt sich also im zweiten Schritt ein einfaches und nachvollziehbares Muster, das auch für längere Pfade Bestand hat. In der anderen Richtung können wir diagnostische Schlüsse mit Hilfe des bayesschen Theorems vollziehen. Wenn wir etwa wissen, dass w_1 realisiert wurde und möchten wissen, was das für Z und Y und X bedeutet, können wir so vorgehen:

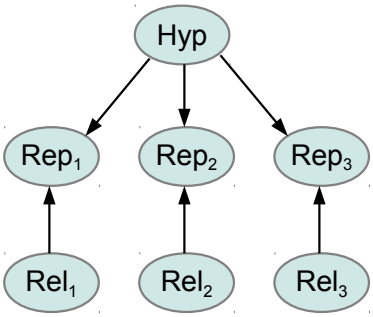
$P(z_1|w_1) = P(w_1|z_1)\cdot P(z_1) \,/\, P(w_1) = 0{,}6069$ und entsprechend:
$P(y_1|w_1) = P(w_1|y_1)\cdot P(y_1) \,/\, P(w_1)$,

hierbei fehlt uns allerdings noch die bedingte Wahrscheinlichkeit $P(w_1|y_1)$, die wir erst noch gemäß dem oberen Verfahren ermitteln müssten usf.

Damit ist zumindest für einfache Pfade deutlich geworden, wie wir hier rechnen können. Für komplexere Fälle gibt es – wie gesagt – noch speziellere Algorithmen, um die Berechnungen zu vereinfachen, aber das überlassen wir lieber dem Computer. Uns genügt es zu wissen, wie diese Rechnungen im Prinzip funktionieren, wenn wir ein bayessches Netz vor uns haben.

Bayessche Netze werden allerdings nicht nur eingeführt, um kausale Beziehungen wiederzugeben, sondern dienen ebenso dazu, bestimmte inferentielle Zusammenhänge und Unabhängigkeiten anschaulich darzustellen. In Bovens & Hartmann (2006) wird z.B. untersucht, wie sich *verschiedene Zeugenaussagen* oder wie sich unterschiedliche Messungen auf die Wahrscheinlichkeit einer Hypothese auswirken. Dabei wird

jeweils eine (oder mehrere) Zufallsvariable(n) Rel eingeführt, die die jeweilige Zuverlässigkeit der Zeugen bzw. die Zuverlässigkeit der Messungen (bzw. Hilfshypothesen) repräsentieren soll. Damit wird natürlich nicht unbedingt behauptet, dass Rel einen eigenständigen kausalen Faktor wiedergibt, sondern nur, dass es sich dabei um einen in einer bestimmten Modellierung der Zusammenhänge relevanten Parameter handelt, der von anderen Parametern oder Aspekten der Situation (kausal) abhängig bzw. unabhängig ist. Dessen Abhängigkeit von anderen Faktoren bzw. die Unabhängigkeiten verschiedener Zeugen in puncto Zuverlässigkeit können dann in Form eines bayesschen Netzes sichtbar gemacht werden, wodurch die unterschiedlichen Modelle, die in dem Buch diskutiert werden, leichter nachvollziehbar sind. Die jeweiligen Unabhängigkeitsannahmen sind direkt aus dem Netz ersichtlich.



Bayessches Netz 4: Drei Zeugenaussagen

Dabei soll Hyp die Behauptung der drei Zeugen darstellen (bzw. eine wissenschaftliche Hypothese) und Rep die Aussagen der Zeugen (bzw. die Messergebnisse), während der Faktor Rel die jeweilige Zuverlässigkeit der Zeugen (bzw. die der Messgeräte) angeben soll. Unser Netz besagt dann z.B., dass in dieser Modellierung ihre jeweilige Zuverlässigkeit voneinander unabhängig ist und sowohl die Wahrheit der Behauptung wie auch die Zuverlässigkeit der Zeugen Einfluss darauf hat, was die Zeugen behaupten. Wir können in einem solchen Netz nicht nur auf die Wahrscheinlichkeit für die Behauptung der Zeugen schließen, sondern es auch in der anderen Richtung nutzen und etwa auf die Zuverlässigkeitsfaktoren schließen (vgl. Bovens & Hartmann 2006). Wir werden später auf diese Anwendungen noch einmal zurückkommen,

wenn wir besprechen, wie sich Zeugenaussagen im bayesianischen Rahmen einbinden lassen.

Der Einsatz dieser bayesschen Netze ist m.E. aber schon deshalb so bedeutsam, weil er nun das einbringt, was nach meinen früheren Ausführungen von entscheidender Bedeutung ist für das induktive Schließen, was aber im Bayesianismus bisher nicht zu finden war, nämlich unser *kausales Hintergrundwissen* darüber, wie bestimmte Faktoren in unserer Welt untereinander kausal zusammenhängen. Man erkennt das sehr schön in den Überlegungen von Bovens & Hartmann (2006), in denen die Autoren unterschiedliche Situationen, die sich durch Zeugenaussagen oder Messungen ergeben können, zunächst durch ein bayessches Netz modellieren, das unsere kausalen Kenntnisse über die Zusammenhänge in diesen Beispielen einbringt, und dieses dann als Grundlage für ihre bayesianischen Berechnungen verwenden.

Genau genommen können wir sagen, dass wir uns damit schon wieder ein gutes Stück weit vom klassischen Bayesianismus entfernen, denn es zählen nun nicht mehr nur unsere einfachen Wahrscheinlichkeitsschätzungen, sondern es werden explizit kausale Annahmen eingespeist – erst durch das Netz und dann durch den entsprechenden Einfluss, das unser Netz auf unsere Berechnungen hat. Aus meiner Sicht versucht der Bayesianismus an dieser Stelle ein Versäumnis zu berichtigen, dem er bis dahin unterliegt. Dass es sich dabei um ein Versäumnis handelt, dafür habe ich bereits in Kapitel 1.4.3 anhand von Beispielen argumentiert, in denen deutlich wurde, wie insbesondere unsere kausalen Annahmen unsere weiteren induktiven Schlüsse beeinflussen. Hier gehen solche kausalen Annahmen in die Gestaltung unserer Netze und damit in unsere Modellierung der Situation ein, und beeinflussen so massiv die bayesianischen Berechnungen der Wahrscheinlichkeiten.

Eine spannende Teilaufgabe für entsprechende bayesianische Berechnungen nennt sich auch im Rahmen der bayesschen Netze *abduktive Schlüsse*. Dabei geht es darum, zu bestimmten Symptomvariablen $S_1,\ldots,S_r$ die Krankheitsvariablen $D_1,\ldots,D_n$ zu finden, die dazu die wahrscheinlichsten Erklärungen bieten. Das wird in Neapolitan (2004, Kap. 4.3) aber so gedeutet, dass wir die Einstellungen d suchen, die zu den vorgegeben Symptomen d am wahrscheinlichsten werden:

Bayesianische Abduktion: Wähle d so, dass P(d|m) maximal ist.

Das ist eine typisch bayesianische Herangehensweise an die Abduktion. Die klassischen abduktiven Ansätze würden stattdessen verlangen, dass d so zu wählen ist, dass vor allem P(m|d) möglichst groß wird. Dabei sind die d_i etwa als (Krankheits-)Hypothesen konzipiert und m repräsentiert jeweils die Daten bzw. Symptome. Dann werden abduktiv die Hypothesen ausgewählt, die die Daten am wahrscheinlichsten erscheinen lassen, während der Probabilist obendrein den Hypothesen Wahrscheinlichkeiten zuordnet und daher auch diesen anderen Weg einschlagen kann. Warum man den unbedingt als Abduktion bezeichnen sollte, ist aber keineswegs so klar. Allerdings gibt es natürlich gewisse Zusammenhänge zwischen beiden Wegen. Nur dass für die bayesianische Abduktion wieder die Vorher-Wahrscheinlichkeiten das Verfahren prägen, während für die klassische Abduktion allein die Frage im Vordergrund steht, wie gut die Theorien die Daten erklären können.

5.3.10 Das Dogmatismusverbot

Oft wird dem Bayesianismus vorgeworfen, dass er zu subjektiv sei und die Ausgangswahrscheinlichkeiten bzw. die Ausgangsglaubensgrade zu willkürlich bestimmt würden. Deshalb suchen Bayesianer immer wieder nach sinnvollen Möglichkeiten, diese »Vorher-Wahrscheinlichkeiten« weiter zu begrenzen. Eine solche Forderung findet sich im sogenannten *Dogmatismusverbot*.

Darin wird verlangt, dass nur analytisch wahren oder falschen Aussagen die Wahrscheinlichkeiten 1 oder 0 zugeschrieben werden dürfen.

Dogmatismusverbot: Ist A weder eine analytisch noch logisch wahre bzw. falsche Aussage, so gilt: $0 < P(A) < 1$.

Das sieht zunächst recht plausibel aus und setzt nur den Fallibilismus um, den wir üblicherweise mit unserer wissenschaftlichen Methodologie verbinden. Wir müssen danach zumindest immer die Möglichkeit berücksichtigen, dass sich eine empirische Behauptung als falsch (oder auch als wahr) erweisen kann. Für den Bayesianer ist jedenfalls klar, dass eine einmal auf 1 oder 0 gesetzte Wahrscheinlichkeit nicht wieder durch

irgendwelche Daten E zu revidieren ist. Dazu können wir schlicht auf die folgende Gleichung schauen:

Bayesianisches Updaten: $P^+(H) = P(H|E) = P(H) \cdot [P(E|H)/P(E)]$

Dabei bezeichnen wir mit »P^+« immer die Nachher-Wahrscheinlichkeiten, also die Wahrscheinlichkeiten, die unsere Hypothesen nach dem Updaten mit den Daten annehmen, während »P« die Vorher-Wahrscheinlichkeiten meint. Den Faktor in den eckigen Klammern nenne ich den *Update-Faktor*, da die Vorher-Wahrscheinlichkeit P gerade damit multipliziert wird, um die Nachher-Wahrscheinlichkeit zu erhalten. Für $P(H) = 0$ gilt offensichtlich $P^+(H) = 0$. Für $P(H) = 1$ folgt ebenso $P^+(H) = 1$. Man erhält nämlich durch Einsetzen der entsprechenden Formeln: $P^+(H) = P(E\&H)/P(E)$ und das ist 1, denn P&H ist logisch äquivalent zu E, weshalb Zähler und Nenner gleich sind. (Oder man betrachtet einfach ¬H, das schließlich eine Wahrscheinlichkeit von 0 haben muss, woraus folgt $P^+(\neg H) = 0$ etc.) Also zählt gegen eine einmal auf 1 gesetzte Aussage kein Datum mehr, sie ist tatsächlich *unrevidierbar* im klassischen bayesianischen Apparat. Das Entsprechende gilt für eine auf 0 gesetzte Aussage. Das sollte für einen Bayesianer einen guten Grund darstellen, dem Dogmatismusverbot zuzustimmen. (Erst wenn wir als Update-Regel nicht die klassische Kondiitionalisierungsregel, sondern die MaxEnt-Regel (s.u.) wie bei Jon Williamson 2010 nutzen, lassen sich die Wahrscheinlichkeiten 0 und 1 wieder revidieren.)

Insbesondere würde eine dogmatisch vertretene Aussage A mit P(A)=1 überdies dazu führen, dass jede Beobachtung, die wir anstellen, die scheinbar gegen sie spricht, nun durch A entkräftet würde. Deshalb sind wir normalerweise nicht bereit, dogmatisch vertretene Aussagen hinzunehmen. Allerdings wird der klassische Bayesianer bald schon eine Ausnahme gestatten und Beobachtungsaussagen auf 1 setzen. Das betrachtet er normalerweise als hilfreiche Idealisierung. Aber dazu später mehr.

Bedauerlicherweise hat das Dogmatismusverbot bereits einige Konsequenzen, die uns nicht alle als plausibel erscheinen. Schon Popper hatte darauf hingewiesen, dass ein Naturgesetz wie das newtonsche Gravitationsgesetz behauptet, dass alle Körper gemäß der bekannten

Formel eine Gravitationskraft F auf alle anderen Körper ausüben. Das hat etwa die logische Struktur:

Logische Allstruktur des Gravitationsgesetzes: $G \equiv \forall x \forall y\ (Fxy)$

Solche Allaussagen sollen für alle Objekte zu allen Zeiten gelten. Damit scheint das Gesetz G eine Behauptung über zumindest potentiell unendlich viele Objekte aufzustellen. Das hat Popper bewogen, dafür einzutreten, dass es eigentlich die (Ausgangs-)Wahrscheinlichkeit 0 haben sollte. Dafür gibt es unterschiedliche Überlegungen. Recht einfach gedacht könnte man so argumentieren: Nehmen wir zunächst als Grundlage ein noch einfacheres Gesetz vom Typ: $\forall x\ Fx$ (Für alle Objekte x gilt F) und schreiben es anhand von Termen a_i, die für bestimmte Objekte oder Situationen stehen sollen. Dann ist das intuitiv äquivalent zu einer unendlichen Konjunktion:

Gesetze als unendliche Konjunktionen: $H \equiv \forall i\ Fa_i \equiv Fa_1 \& Fa_2 \& \ldots$

Dann finden wir für die Teilaussagen $F^n \equiv Fa_1 \& \ldots \& Fa_n$ einige erste Abschätzungen. Im einfachsten Fall können wir die Fa_i als voneinander statistisch unabhängig zu unserem Überzeugungssystem P betrachten (im allgemeinen Fall können wir die Kettenregel anwenden) und erhalten:

(a) Unabhängige Daten: $P(F^n) = P(Fa_1) \cdot \ldots \cdot P(Fa_n)$
(b) Allgemeiner Fall (Kettenregel):
$P(F^n) = P(Fa_1) \cdot P(Fa_2|Fa_1) \cdot P(Fa_3|Fa_2 \& Fa_1) \cdot \ldots \cdot P(Fa_n|Fa_1 \& \ldots \& Fa_{n-1}) =$
$= P(Fa_1) \cdot P(Fa_2|F^1) \cdot P(Fa_3|F^2) \cdot \ldots \cdot P(Fa_n|F^{n-1})$

Im Fall (a) wird nun am deutlichsten, was gemeint ist (vgl. a. Kap. 3.1). Wenn wir für jede Aussage Fa_i etwa annehmen, dass ihre Wahrscheinlichkeit eben nicht 1 ist, sondern unterhalb einer bestimmten Schranke $q<1$ liegt, dann ergibt sich schnell:

(c) $P(F^n) \leq q^n$

Dann würde die Wahrscheinlichkeit $P(F^n)$ gegen 0 konvergieren für wachsendes n, wie es Popper uns prophezeite. Um das Dogmatismusverbot auch für H einzuhalten, darf es also kein solches q für alle n geben. Das ginge nur, wenn die $P(Fa_i)$ schnell gegen 1 konvergieren würden. Wir müssten also aufgrund des Dogmatismusverbots schon davon ausgehen, dass die späteren Fälle a_i mit hoher Wahrscheinlichkeit F wären, was wir eigentlich erst per Induktionsschluss zeigen möchten.

Noch klarer wird die Situation für die bedingten Wahrscheinlichkeiten im Falle (b). Sollten die alle echt kleiner 1 sein und würden eine ähnliche Abschätzung erfüllen, wonach es ein q<1 gibt, so dass für alle n gilt:

(d) $P(Fa_n|F^{n-1}) < q < 1$ gilt,

dann würde auch die Ungleichung (c) gelten und damit wäre:

(e) $P(H) = \lim_{n\to\infty} P(F^n) = 0$
(was voraussetzt, dass unser Glaubensgrad P σ-additiv ist)

Das wird aber gerade durch das Dogmatismusverbot ausgeschlossen, denn es besagt schließlich, dass selbst H nicht die Wahrscheinlichkeit 0 erhalten darf. Es darf also keine entsprechende Schranke q<1 geben und das würde bedeuten, dass dann der Term $P(Fa_n|F^{n-1})$ sehr schnell gegen 1 konvergieren müsste für wachsendes n (vgl. Earman 1992). Damit gilt: Gemäß unserem Überzeugungssystem P erscheint es fast ganz sicher, dass das nächste Objekt a_n für genügend großes n auch die Eigenschaft F hat, wenn die Vorgänger Objekte $a_1,\ldots,a_{n-1}$ sich alle als F erwiesen haben. Das ist aber genau die Induktionseigenschaft, die wir schon oft vergeblich beweisen wollten. Das Dogmatismusverbot ist in diesem Fall also so stark, dass hier ein Induktionsprinzip für bestimmte Anwendungen sogleich mit impliziert wird, das wir direkt eigentlich nicht begründen können. Das Dogmatismusverbot entpuppt sich also als nicht so harmlos, wie es zunächst aussieht. Es impliziert die folgende Induktionseigenschaft:

(IE) Induktionseigenschaft: Für (hinreichend großes) n gilt:
$P(Fa_{n+1}|F^n) > P(Fa_{n+1})$ bzw.
$P(Fa_{n+1}|F^n) \to 1$ für wachsendes n, bzw. $\lim_{n\to\infty} P(Fa_{n+1}|F^n) = 1$

Das besagt im Anwendungsfall, dass wir von dem nächsten noch nicht beobachteten Raben mit hoher Wahrscheinlichkeit wissen, dass er auch schwarz sein wird, wenn wir bisher viele schwarze Raben beobachtet haben.

Was uns daran stutzig machen sollte, sind zwei Dinge: Zunächst ist auffällig, dass wir quasi durch die Hintertür ein Induktionsprinzip geschenkt bekommen, obwohl wir nur ein relativ harmlos erscheinendes Verbot für extreme Wahrscheinlichkeiten aussprechen wollten. Vielleicht sollten wir das Dogmatismusverbot daher doch lieber auf Aussagen über *endlich* viele Individuen bzw. Anwendungsfälle einschränken. Dann tritt dieser eher seltsame Nebeneffekt nicht auf. Die Bayesianer müssten sich freilich überlegen, wie sich das präzisieren ließe.

Außerdem ist seltsam, dass es keine Beschränkungen für die Art der Eigenschaften gibt. Für alle Eigenschaften (also z.B. auch goodmansche Eigenschaften wie »grue«, die offensichtlich nicht projizierbar sind) gilt die Induktionseigenschaft (IE). Damit haben wir offensichtlich zu viel bewiesen. So einfach sollten wir ein Induktionsprinzip nicht postulieren. Es tritt auch nur in Kraft, wenn wir Aussagen über einen abzählbar unendlichen Individuenbereich machen, die sich in abzählbar unendlich viele Einzelaussagen zerlegen lassen. Es sieht eher wie ein formales Artefakt unserer Darstellung aus, als der lang ersehnte Durchbruch in der Debatte um das humesche Induktionsproblem. Trotzdem wird er von Bayesianern manchmal als solcher gefeiert (vgl. Howson 2000, 72).

Außerdem können wir hier eigentlich nur das Problem beobachten, dass wir für unendlich viele voneinander probabilistisch unabhängige Aussagen $A_1, A_2, \ldots$ keine Möglichkeit sehen, allen Aussagen eine feste Wahrscheinlichkeit kleiner als q mit $0 < q < 1$ zuzuweisen, so dass auch noch die Allaussage $\forall i\, A_i$ eine Wahrscheinlichkeit echt größer null erhalten kann. Damit wir das Dogmatismusverbot für die Allaussage aufrechterhalten können, müssen die A_i nun Wahrscheinlichkeiten erhalten, die sehr schnell anwachsen. Das gilt für beliebige Aussagen A_i und hat daher nicht viel mit der gesuchten Induktionseigenschaft zu tun, denn für unsere Aussagen A_i wird nicht verlangt, dass sie in irgendeiner Weise über eine bestimmte Art von Individuen sprechen.

Interessant ist noch, dass das Problem dort in gewisser Weise wieder verschwindet, wo wir *überabzählbar* viele Aussagen erhalten, die in einer

Aussage zusammenkommen. Nehmen wir an, wir haben eine Hypothese $H_{[2;4]}$, wonach die uns unbekannte Masse des nächsten Granitblocks, den wir finden, zwischen 2 und 4 kg liegt. Diese Hypothese – wie beliebig viele andere Hypothesen – muss dann eine nichttriviale Wahrscheinlichkeit aufweisen, sagen wir einfach $P(H_{[2;4]}) = 0{,}6$. Dazu gibt es eine Reihe von Hypothesen H_x, die besagen, dass die gesuchte Größe gerade x ist. Auch diese Hypothesen sollten nach dem Dogmatismusverbot alle eine von Null verschiedene Wahrscheinlichkeit aufweisen. Doch wenn wir die Menge $M = \{H_x: x \in [2;4]\}$ betrachten, sollte diese wieder die Wahrscheinlichkeit 0,6 aufweisen. Das ist aber für eine solche überabzählbar große Menge nicht möglich. Jede sinnvolle Summenbildung würde zu unendlich großen Werten führen. Wir haben aber jetzt wieder (anders als im abzählbar unendlichen Fall) die Möglichkeit, eine Gleichverteilung auf der Menge zu definieren, nämlich durch eine Dichtefunktion auf unserem Intervall. Doch die besagt gerade, dass sogar für alle x gilt: $P(H_x) = 0$. Für überabzählbare Mengen ist das Dogmatismusverbot also nicht mehr konsistent anwendbar. Solche Menge erhalten wir aber schon dann, wenn wir unendliche Folgen von Daten zulassen und nun die Menge aller solcher Folgen betrachten. Die meisten Folgen müssten dann wieder die Wahrscheinlichkeit 0 erhalten.

Das Resultat ist, dass das Dogmatismusverbot überraschend starke Konsequenzen hat und daher kaum als sinnvolle generelle Anforderung gelten darf. Trotzdem scheint es auf endlichen Mengen durchaus plausibel zu sein, so dass wir es nicht ganz aus dem Blick verlieren sollten. Jedenfalls dürfen wir es nicht einfach als Sieg über den Induktionsskeptiker feiern, da es viel zu schwach begründet ist, um solch weitreichenden Folgen rechtfertigen zu können. Es wurde vielmehr eingeführt zur Abwehr überzogener epistemischer Ansprüche und sollte dann nicht plötzlich zu unserer stärksten Waffe gegen den Skeptiker mutieren, wenn wir erkennen, dass es auf abzählbar unendlichen Mengen solch überraschende Konsequenzen besitzt.

5.3.11 Wahrscheinlichkeitskoordinierungsprinzipien: Statistischer Syllogismus und Likelihoodanbindung

Einen anderen Weg, die Vorher-Wahrscheinlichkeiten einzuschränken bzw. zu objektivieren, finden wir in den sogenannten *Wahrscheinlichkeitskoordinierungsprinzipien* (so hat Michael Strevens sie zumindest getauft), die alle eine Form von Anbindung der subjektiven an die objektiven Wahrscheinlichkeiten verlangen. Ein solches Prinzip hatten wir schon kennengelernt, nämlich den statistischen Syllogismus (vgl. Kap. 5.2).

> **Statistischer Syllogismus**: Ist die relative Häufigkeit von Fs in der Grundgesamtheit G gerade r und wir betrachten ein a aus G, von dem wir keinen Grund zu der Annahme haben, dass seine Auswahl speziell mit dem Haben der Eigenschaft F verknüpft ist, dann sollte unsere subjektive Wahrscheinlichkeit P(Fa) = r sein.

Wir hatten mit Hilfe des epistemischen Gleichbehandlungsgrundsatzes dafür argumentiert, dass es sich schon um eine logische Wahrscheinlichkeit handelt, die wir hier erhalten, und daher auch der Probabilist gut daran täte, in den Fällen, in denen entsprechende relative Häufigkeiten vorliegen, diese als seine subjektiven Wahrscheinlichkeiten zu übernehmen. Ohne diese Regel ginge den Bayesianern sonst die entscheidende Anbindung ihrer Startwahrscheinlichkeiten an empirische Daten verloren. Insbesondere diese Form der empirischen Kalibrierung von Glaubensgraden lässt den probabilistischen Ansatz etwa gegenüber rein komparativen Ansätzen wie dem der Rangfunktionen (vgl. Kap. 5.12) so attraktiv erscheinen, da diese anderen Ansätze keine Möglichkeit aufweisen, die relativen Häufigkeiten direkt einzubeziehen.

Betrachten wir nur kurz noch einmal ein Beispiel dazu: Nehmen wir an, wir wissen, dass im Durchschnitt 30% aller Patienten mit Lassa-Fieber sterben. Tom hat Lassa-Fieber und wir haben keine weiteren Kenntnisse über Toms Abwehrkräfte oder seine spezielle Behandlung. Was ist dann der vernünftigste Tipp dafür, dass auch Tom sterben wird? Es scheint offensichtlich zu sein, dass wir uns in solchen Fällen für die 30% entscheiden sollten, da das die einzige Information ist, die wir

effektiv über Tom haben. Somit bietet der statistische Syllogismus (wie in unserem Beispiel des Dschungelfiebers) einen ersten Weg an, um bei der Bestimmung von Vorher-Wahrscheinlichkeiten auf objektive Daten zurückzugreifen. Da uns in einigen Fällen entsprechende Statistiken vorliegen, handelt es sich tatsächlich um ein praktikables Verfahren zu ersten Wahrscheinlichkeiten zu gelangen, die wir als Bayesianer anschließend mit weiteren Informationen updaten können.

Allerdings setzt der statistische Syllogismus voraus, dass wir keine weiteren relevanten Informationen über Tom haben, was seine Anwendung einschränkt. Natürlich können wir sogar versuchen, diese Situation hypothetisch zu betrachten und dann erst unsere weiteren Informationen über Tom in Form bayesianischen Updatens einzubringen. Abgekürzt hat der statistische Syllogismus die folgende Form:

Statistischer Syllogismus: $P(Fa|h_n(F/G) = r \;\&\; Ga \;\&\; \neg Ia) = r$

wobei mit ¬Ia gemeint ist, dass wir über keine weiteren Informationen über a verfügen im Hinblick darauf, ob a F ist, außer den hier angeführten.

Die Likelihoodanbindung. Ein weiteres zentrales Koordinierungsprinzip können wir als *Likelihoodanbindung* bezeichnen. Starten wir wieder mit einem Beispiel. Unsere Hypothese H besage, dass eine Münze gefälscht ist und die Wahrscheinlichkeit für Kopf gerade 0,4 betrage. Wie hoch sollte dann unsere subjektive Wahrscheinlichkeit dafür sein, dass im nächsten Wurf mit dieser Münze Kopf kommt, wenn wir die Hypothese H dabei als wahr annehmen? Das heißt, wie sollen wir P(Kopf|H) wählen?

Auch hier liegt es nahe, P(Kopf|H) = 0,4 zu wählen. Zumindest diese spezielle bedingte subjektive Wahrscheinlichkeit könnten wir so objektivieren. Allerdings ist ein Probabilist darauf nicht zwingend festgelegt, sondern kann sich nur entschließen, seine Position dadurch zu objektivieren und damit zugleich für den klassischen Statistiker attraktiver zu gestalten, denn der akzeptiert nur solchen objektiven (direkten) Likelihoods. Allgemeiner könnten wir die Likelihood der Daten E, die durch unsere Hypothese vorgegeben ist [manchmal sagt man auch die

Likelihood der Hypothese H relativ zu einem Datum E (so irreführend redet man oft im Umgang mit Likelihoods, doch dem möchte ich nicht folgen)] mit $P_H(E)$ bezeichnen, um sie an dieser Stelle von den bedingten Wahrscheinlichkeiten des Probabilisten zu unterscheiden. Damit ist die Wahrscheinlichkeit gemeint, die E nach Auskunft der Hypothese H hat, sollte die Hypothese so stark sein, eine solche Likelihood zu bestimmen. (Im Englischen spricht man manchmal von »direct likelihood inference«.) Die Wahrscheinlichkeit $P_H(E)$ bringt für uns zum Ausdruck, welche Bedeutung die Hypothese H besitzt und daher sollte der Probabilist sich in seinen Wahrscheinlichkeiten daran auch halten, weil er sonst der Hypothese einen anderen *empirischen Gehalt* zuweisen würde.

Die Likelihoodanbindung: $P(E|H) = P_H(E)$

Hawthorne (u.a. in 2011c) verweist darauf, was passieren würde, wenn sich zwei Wissenschaftler an diesem Punkt uneins wären. Der erste würde etwa $P(E|H) = 0{,}8$ und der zweite $P(E|H) = 0{,}1$ ansetzen. Dann würde im Normalfall das Auftreten von E die Hypothese H für den ersten Wissenschaftler bestätigen, während es für den zweiten Wissenschaftler in den meisten Fällen wohl gegen H sprechen würde. Die Wissenschaftler wären sich jedenfalls offensichtlich uneinig darüber, welchen *empirischen Gehalt* die Hypothese hat, d.h. sie verstehen die Hypothese H jeweils anders. Sie benutzen zwar denselben syntaktischen Ausdruck, um ihre Hypothese darzustellen, aber sie interpretieren sie so unterschiedlich, dass wir von zwei Hypothesen sprechen sollten. Im ersten Fall lässt uns die Hypothese H das Ereignis E erwarten, während es für den zweiten Wissenschaftler ganz unwahrscheinlich ist, wenn H wahr ist. Wir können dafür natürlich keine sinnvollen Induktionsregeln mehr erwarten. Eine Voraussetzung unserer Fragestellung nach geeigneten Induktionsverfahren ist geradezu, dass wir uns auf *eine* Hypothese und *eine* Interpretation dieser Hypothese geeinigt haben. Nur für die eine Hypothese fragen wir dann, wie gut sie durch bestimmte Daten bestätigt wird.

Dann sollten wir zumindest in den Likelihoods $P(E|H)$ *übereinstimmen*, die den empirischen Gehalt der Hypothese zum Ausdruck bringen.

Die Likelihoods drücken gerade aus, was uns die Hypothesen über die Daten zu sagen haben. Nur bei gegebener Übereinstimmung über diese Likelihoods können Bayesianer jedenfalls erwarten, dass wir mit Hilfe der Funktion P die Bestätigung in objektiver Weise angeben können, die H durch die Daten erfährt. Gehören dann sogar konkrete Aussagen über bestimmte Wahrscheinlichkeiten zum Inhalt der Hypothese, dann müssen wir das ernst nehmen und in Form der Likelihoodanbindung umsetzen, sonst werden wir keine sinnvolle Bestätigungsbeziehung mehr erhalten.

Auch hier müssen wir aber als strikte Bayesianer wieder verlangen, dass wir über keine anderen relevanten Informationen für das Auftreten von E verfügen. Wir können zunächst nur erwarten, dass die Likelihoodanbindung für die Startwahrscheinlichkeiten funktioniert. Das wird etwa angenommen, in dem wir als Vorwissen explizit nur von H ausgehen. Es könnte aber sogar Unterminierer zu H geben oder Gegengründe gegen E, die wir dann berücksichtigen müssten, um Es bedingte Wahrscheinlichkeit zu bestimmen. Sie können etwa schon Eingang in unsere Plausibilitätseinschätzung P gefunden haben und wären dann im Sinne des Bayesianismus zu berücksichtigen. Solange wir die nicht kennen, scheint es jedoch auch für den Bayesianer rational zu sein, sich ganz auf die vorgegebenen Likelihoods zu stützen. Das kann für den Bayesianer allerdings bedeuten, dass er noch andere Anpassungen für P vornehmen muss, die er vielleicht aus anderen Gründen nicht vornehmen möchte. Darauf werden wir zurückkommen.

Allerdings muss diese objektive Likelihood nicht immer existieren. Prominente Problemfälle sind *disjunktive Hypothesen* wie die folgenden:

H1: Die Wahrscheinlichkeit für Kopf beträgt 0,3 oder 0,4.
H2: Die Wahrscheinlichkeit für Kopf liegt im Intervall [0,2;0,4]

In beiden Fällen ist $P_H(E)$ nicht direkt bestimmt, weil keine definitive Aussage darüber getroffen wird, welche Wahrscheinlichkeit E nach Annahme von H nun zukommt. Der klassische Statistiker kommt an dieser Stelle nur mit gewissen (bayesianischen) Tricks noch einen Schritt weiter. Er könnte ein Indifferenzprinzip anwenden und im Falle von H1 vermuten, dass beide Möglichkeiten dieselbe Chance haben, weshalb

dann $P_{H1}(\text{Kopf}) = 0{,}35$ zu wählen wäre. Im Falle von H2 könnte das Indifferenzprinzip besagen, dass alle Werte zwischen 0,2 und 0,4 dieselbe Chance haben (wir hätten dann eine konstante Dichtefunktion auf dem Intervall) und entsprechend $P_{H2}(\text{Kopf}) = 0{,}3$ gewählt werden könnte. Doch die Anwendung solcher Indifferenzprinzipien lehnt der klassische Statistiker eigentlich ab, denn sie gehören nur in das Repertoire der Vertreter einer induktiven Logik und stellen zumindest epistemische Wahrscheinlichkeiten dar.

Der Probabilist kann dagegen versuchen, diese Informationen weiter auszuwerten. Er kann etwa die folgenden Unterhypothesen für sich betrachten und ihnen jeweils eigene Wahrscheinlichkeiten zuordnen:

H1a: Die Wahrscheinlichkeit für Kopf beträgt 0,3.
H1b: Die Wahrscheinlichkeit für Kopf beträgt 0,4.

Unter der Annahme, dass H1 gilt, sollten diese sich zu 1 aufsummieren, also könnte etwa gelten:

$P(\text{H1a}|\text{H1}) = 0{,}2$
$P(\text{H1b}|\text{H1}) = 0{,}8$

Dann erhalten wir $P_{H1}(\text{Kopf}) = 0{,}2 \cdot 0{,}3 + 0{,}8 \cdot 0{,}4 = 0{,}38$, denn es gilt: $P(\text{E}|\text{H1}) = P(\text{E}|\text{H1a}) \cdot P(\text{H1a}|\text{H1}) + P(\text{E}|\text{H1b}) \cdot P(\text{H1b}|\text{H1})$, da H1a und H1b einander ausschließen. Für die Hypothese H2 könnte der Probabilist mit Apriori-Dichten arbeiten und entsprechende Ergebnisse erzielen. Auf solche Fälle gehen wir später noch ein. Damit werden die bekannten Informationen über die Wahrscheinlichkeit von Kopf dann doch noch ausgewertet. Es dürfte allerdings klar sein, dass die strikte Likelihoodanbindung, wie sie zunächst vorgestellt wurde, auch dem klassischen Statistiker gefällt, während die erweiterten Ideen, die wir dann eingeführt haben, eigentlich nicht mehr seine Zustimmung finden, weil wir dort sogar den Hypothesen selbst bestimmte Wahrscheinlichkeiten geben und nicht nur bestimmten Zufallsereignissen. Die Hypothesen sind aber entweder wahr oder falsch und damit sind die entsprechenden von 0 und 1 verschiedenen Wahrscheinlichkeiten für diese Hypothesen subjektiver Natur (oder eben epistemische Wahrscheinlichkeiten) und somit für den klassischen Statistiker abzulehnen. Die Likelihoods sind dagegen im

einfachen Fall einfach nur Behauptungen unserer Theorie, auf die wir uns beim Testen der Theorie natürlich beziehen müssen, sonst wäre eine probabilistische Theorie überhaupt nicht mehr testbar.

Der Bayesianer sollte also zunächst von der Likelihoodanbindung einen Gebrauch machen, denn er wertet dabei nur konsequent seine Informationen aus und stützt sich dabei auf so viele objektive Informationen, wie es nur geht. Erst wenn er damit allein nicht mehr weiterkommt, muss er wieder seine Zuflucht in subjektiveren epistemischen Wahrscheinlichkeiten suchen, um etwa mit dem Problem disjunktiver Hypothesen umzugehen.

Im Update-Faktor des bayesianischen Updatens kommt die Wahrscheinlichkeit P(E) der Daten vor und dort ist der Bayesianer normalerweise froh, die Likelihoodanbindung einsetzen zu können. Das gelingt ihm z.B. dann, wenn er bereits über eine vollständige Menge von paarweise einander ausschließenden Hypothesen $\{H_1,\ldots,H_n\}$ für seine Anwendung verfügt, denn dann gilt:

$$P(E) = \sum_i P(E|H_i)\ P(H_i) = \sum_i P_{Hi}(E)\ P(H_i)$$

Dadurch kann zumindest eine Komponente in der Berechnung von P(E) objektiviert werden. Da es sich bei der Likelihoodanbindung außerdem um ein sehr naheliegendes Verfahren der Bestimmung unserer Likelihoods geht, sollten wir dieses Prinzip übernehmen. Es ist auch tatsächlich oft anwendbar und nicht nur theoretischer Natur, wie das später zu diskutierende Hauptprinzip.

Allerdings werden wir noch sehen (im Zusammenhang mit dem Problem der »old evidence«), dass gerade Bayesianer sich einerseits gern auf dieses Prinzip stützen möchten, aber andererseits es nicht immer zu ihrem Verfahren des Updatens passt, weshalb wir womöglich eine zweite Wahrscheinlichkeitsfunktion benötigen, neben der Funktion für die Glaubensgrade, nämlich eine, die nur die objektiven Zusammenhänge zwischen den Hypothesen und den Daten zum Ausgangspunkt nimmt und dann nur bestimmt, wie stark die Daten die Hypothese stützen, wie hoch also die Wahrscheinlichkeit der Hypothese nach dem Updaten sein sollte, ohne dabei immer schon einzubringen, wie wahrscheinlich die Daten aus anderen Gründen sind (vgl. a. Kap. 5.8.1+5.8.2). Wir

können sagen, dass eine Anfangs-Likelihoodanbindung durchaus zum Bayesianismus passt, aber eine stabile Likelihoodanbindung, die auch bei weiteren Update-Schritten eingehalten wird, leider nicht immer mit dem Bayesianismus kompatibel ist.

Allgemein ist es ein Problem, wie wir Wahrscheinlichkeitskoordinierungsprinzipien mit dem Updaten zusammenbringen können. Relativ unproblematisch ist es zunächst, die Koordinierungsprinzipien für eine Ausgangswahrscheinlichkeit einzusetzen und danach nur noch upzudaten. Aber schon zu Beginn können weitere Informationen E vorliegen, die wir alternativ im Hintergrundwissen der Startwahrscheinlichkeiten ansiedeln können, so dass sie im Prinzip durch Anwendung unserer Koordinierungsprinzipien eigentlich keinen Eingang in unsere Wahrscheinlichkeitseinschätzungen finden, oder wir lassen sie zunächst heraus und updaten dann erst später damit. Beide Wege müssen keineswegs immer zu denselben Ergebnissen führen.

Die Stabilität der Likelihoodanbindung. Das ist kein Problem für den statistischen Syllogismus, da zu seinen Anwendungsbedingungen definitiv dazugehört, dass keine weiteren Informationen über die fragliche Aussage vorliegen. Das Problem sieht man dagegen sogleich am Beispiel der Likelihoodanbindung. Nehmen wir an, wir würden P zunächst mit E' updaten und fragen uns, was dann aus der Likelihood P(E|H) dabei wird: Wenn wir also zunächst updaten, wobei P in P^+ übergeht, und wir erst daran anschließend wieder unsere Likelihoodanbindung einsetzen, dann würde die Likelihoodanbindung besagen:

$$P^+(E|H) = P_H(E) = P(E|H),$$

jedenfalls, wenn auch schon in P das P(E|H) mit Hilfe der Likelihoodanbindung festgelegt wird. Doch damit implizierten wir zugleich eine *bedingte Unabhängigkeit der Daten* E und E' gegeben H, denn es gilt:

$$P^+(E|H) = P^+(E\&H) \,/\, P^+(H) = P(E\&H|E') \,/\, P(H|E') =$$
$$P(E\&H|E')\cdot P(E') \,/\, P(H|E')\cdot P(E') = P(E\&H\&E') \,/\, P(H\&E') =$$
$$P(E|H\&E') = P(E|H)$$

Man sieht also an der letzten Gleichung, dass $P^+(E|H) = P(E|H)$ nur dann gilt, wenn die Daten E und E' probabilistisch unabhängig sind,

gegeben H, d.h., wenn gilt: P(E&E'|H) = P(E|H)·P(E'|H) oder äquivalent dazu eben P(E|H&E') = P(E|H). Das werden wir später auch so notieren: (E⊥E'|H). Nur wenn diese Unabhängigkeit in P enthalten ist, können wir also im Falle unserer Daten E und E' entweder zunächst die Likelihoodanbindung ausnutzen und dann mit E' updaten oder erst mit E' updaten und dann die Likelihoodanbindung erst für E nutzen. Gilt die Unabhängigkeitsbeziehung dagegen nicht, kann der Bayesianer die Likelihoodanbindung nur einmal zu Beginn nutzen und muss anschließend die neuen Likelihoods anhand des üblichen bayesianischen Updatens ermitteln.

In unserem Beispielüberzeugungssystem in Kapitel 5.3 war die Likelihood für D gegeben H zunächst 0,7, aber nach dem Updaten mit E war sie schon auf 0,714 angewachsen. Damit erhalten wir das Problem, dass unser Startpunkt (bei dem unsere Likelihoodanbindung zum Einsatz kommt) eine gewisse Willkür aufweist oder die Likelihoodanbindung sogar nie zum Einsatz kommt, weil wir natürlich meist schon über ein bestimmtes Hintergrundwissen verfügen. Außerdem erscheint es auch seltsam, dass wir in dem Beispiel davon ausgehen, dass eine der beiden Hypothesen wahr ist (Kopf also die objektive Wahrscheinlichkeit 0,3 oder 0,7 hat), dann aber bei Annahme von H die Wahrscheinlichkeit von 0,7 für das Auftreten von Kopf noch überschreiten, nur weil bereits E aufgetreten ist. E sollte zwar für H im Unterschied zu H' sprechen, aber H besagt dann nur, dass das Auftreten von Kopf die Wahrscheinlichkeit 0,7 hat, und es scheint nicht sinnvoll, stattdessen nun 0,714 für Kopf anzusetzen.

Die Likelihoodanbindung scheint jedoch eigentlich auch für den Fall *wiederholten Updatens* weiterhin als Forderung recht plausibel zu sein. Besagt unsere Hypothese, dass die Wahrscheinlichkeit für Kopf gerade 0,4 ist, so sollte unsere subjektive Wahrscheinlichkeit dafür, dass zweimal Kopf kommt, gerade 0,4·0,4 = 0,16 sein. Zumindest die bedingte Wahrscheinlichkeit für ein weiteres Mal Kopf gegeben unsere Hypothese sollte weiterhin 0,4 sein. Das ließe sich etwa so begründen: Nur wenn unsere Glaubensgrade P so gestaltet sind, erfassen sie richtig, was der *Inhalt* unserer Hypothese ist, und nur dann wird die Bestätigung, die unsere Hypothese durch diese speziellen Daten erfährt, angemessen dargestellt. Schließlich ändert sich der Inhalt unserer Hypothese nicht

durch neue Daten und sollte daher auch nach dem Auftreten neuer Daten denselben Wert für das Auftreten von Daten eines bestimmten Typs liefern. Dann erst können wir die Wirkungen wiederholten Updatens für unsere Hypothese auch auf nachvollziehbare Weise berechnen. Nehmen wir dagegen an, dass sich die bedingten Wahrscheinlichkeiten für die Daten E mit immer mehr Kopf-Ergebnissen langsam etwa in Richtung zu höheren Werten hin verschieben (man beachte, dass exakterweise gilt: P(E&E'|H) = P(E|E'&H)·P(E'|H)), dann benötigen wir irgendwoher wieder Wahrscheinlichkeiten für die entsprechenden Konjunktionen. Die sind allerdings wieder einmal rein subjektiv festzulegen. Nehmen wir dagegen an, dass E und E' probabilistisch unabhängig sind gegeben H, so ergibt sich P(E&E'|H) = P(E|H)·P(E'|H) und wir können an der Likelihoodanbindung festhalten.

Was passieren könnte, wenn wir uns überhaupt nicht an die Likelihoodanbindung halten, können wir an unserem Beispielüberzeugungssystem aus Kapitel 5.3 erläutern. Behalten wir in dem Fall etwa alle Glaubensgrade bei, tauschen aber die Inhalte der beiden Hypothesen H und H' gerade aus, so dass nun H besagt, dass die Wahrscheinlichkeit für Kopf 0,3 sei. Trotzdem behalten wir aber den Glaubensgrad P(E|H) = P(D|H) = 0,7. Dann führt das Updaten mit zweimal Kopf zu einer Bestätigung von H (P^{++}(H) = 0,833) und zu einer entsprechenden Abwertung von H', obwohl das unserer intuitiven Einschätzung massiv widerspricht, denn die Daten sprechen eindeutig gegen diese neue Hypothese H. Schon um solche antiintuitiven Resultate zu vermeiden, sind wir auf die Likelihoodanbindung angewiesen und sollten daher die Likelihoods auch möglichst stabil halten.

Um die Likelihoodanbindung in unserem Beispielsystem aus Kapitel 5.3 auch stabil zu gewährleisten (also auch für das zweite Updaten), müssen wir zusätzlich verlangen, dass die Daten E und D probabilistisch unabhängig sind, gegeben unsere Hypothese H und auch gegeben unsere Hypothese H', etwa im Sinne eines bayesschen Netzes mit einem Elternknoten für H bzw. H' und den Kindern E und D, zwischen denen es keine weitere Verbindung gibt. Dann erhalten wir die Forderungen:

(1) P(H) = 0,5 = P(H')
(2) P(E|H) = P(D|H) = 0,7 (anfängliche Likelihoodanbindung)

(3) $P(E|H') = P(D|H') = 0{,}3$ (anfängliche Likelihoodanbindung)
und nun zusätzlich
(4) $P(D|EH) = P(D|H) = 0{,}7$ (stabile Likelihoodanbindung)
(5) $P(D|Eh) = P(D|h) = 0{,}3$ (stabile Likelihoodanbindung)

Wenn wir dann wieder unsere Vollkonjunktionen betrachten, lassen sich die entsprechenden leicht gerundeten Werte nun aufgrund der Forderungen (1) bis (5) definitiv ausrechnen. Sie sind nicht weit entfernt von unseren ersten Einschätzungen:

$P(EDH) = x = 0{,}245$ $P(EDh) = t = 0{,}045$
$P(EdH) = P(eDH) = y = 0{,}105$ $P(Edh) = P(eDh) = u = 0{,}105$
$P(edH) = z = 0{,}045$ $P(edh) = v = 0{,}245$

Das ergibt sich aus den Gleichungen:

Aus (1) folgt: $x+2y+z = 0{,}5 = t+2u+v$
Aus (2) folgt: $x+y = 0{,}35$
Aus (3) folgt: $t+u = 0{,}15$
Aus (4) folgt: $x = (0{,}7/0{,}3)\cdot y = 2{,}33\cdot y$
Aus (5) folgt: $u = (0{,}7/0{,}3)\cdot t = 2{,}33\cdot t$

Dann erhalten wir als neue Werte nun die folgenden. Wir können wieder beispielhaft einige ausrechnen. Man bedenke dazu wiederum, dass z.B. $DE \equiv DEH \vee DEh$ ist etc.:

$P(DE) = x+t = 0{,}29$ (statt 0,3)
$P(EH) = x+y = 0{,}35$ (bleibt nach Voraussetzung)
$P(D|EH) = P(EDH)/P(EH) = x/(x+y) =$
$0{,}245/0{,}35 = 0{,}7$ wie gewünscht (statt 0,714)
$P(E) = x+y+t+u = P(D) = 0{,}5$
Damit erhalten wir die upgedateten Wahrscheinlichkeiten:
$P^+(H) = P(H|E) = P(EH)/P(E) = (x+y)/(x+y+t+u) = 0{,}7$
Damit sinkt die Wahrscheinlichkeit für H': $P^+(H') = 0{,}3$
$P^+(D) = P(D|E) = P(ED)/P(E) = 0{,}29/0{,}5 = 0{,}58$ (statt 0,6)
$P^+(D|H) = P(D|EH) = 0{,}7$ (statt 0,714)
$P^+(D|H') = P(D|Eh) = t/(t+u) = 0{,}3$

$P^+(E) = 1$
$P^{++}(H) = P(H|DE) = P(EDH)/P(DE) = 0{,}245/0{,}29 = 0{,}845$ (statt 0,833)
Damit sinkt die Wahrscheinlichkeit für H': $P^+(H') = 0{,}155$
$P^{++}(E) = P^{++}(D) = 1$

Damit ist die Bestätigung unserer Hypothese H durch das zweite Datum etwas größer geworden als im ersten Beispielsystem, weil wir angenommen haben, dass es unabhängig vom ersten ist gegeben H. Das ist aber nicht unbedingt unerwünscht.

Bayesianer argumentieren allerdings manchmal in einer anderen Richtung, wenn sie etwa innerhalb ihres Ansatzes rekonstruieren möchten, weshalb die wiederholte Ausführung von ein und demselben Experiment keine so guten Bestätigungen einer Hypothese liefert, wie die Daten unterschiedlich gestalteter Experimente. Allerdings ändert sich der Update-Faktor $P(D|H)/P(D)$ in unserem Beispiel auch schon langsam dadurch, dass P(D) durch das Updaten mit E größer wird, weil sich die Gewichte für die Hypothesen langsam verschieben und P(D) gerade die gewichtete Summe der bedingten Wahrscheinlichkeiten $P(D|H_i)$ darstellt, die ihrerseits durch die Likelihoodanbindung stabil gehalten werden. Daher könnten wir durchaus an der Likelihoodanbindung festhalten, und erhielten trotzdem das oben geschilderte Phänomen.

Wenn wir etwa zunächst mit E updaten und dann mit D, so ergäbe sich bei Likelihoodanbindung:

$$P^+(D) = \Sigma_i\, P(D|H_i)\, P^+(H_i)$$

Da sich dabei die $P^+(H_i)$ entsprechend den vorherigen Daten zu den am besten treffenden Hypothesen verändern, werden sich auch die Wahrscheinlichkeiten $P^+(D)$ dem langsam anpassen. Jedenfalls werden wir im Folgenden in der Regel so rechnen, weil wir sonst neue Wahrscheinlichkeiten für den Term $P^+(D|H_i)$ bestimmen müssten.

Das Hauptproblem dürfte wohl weiterhin darin zu sehen sein, dass viele unserer Hypothesen für viele Daten keine genauen Wahrscheinlichkeiten angeben. Besagt die Hypothese etwa, dass eine größere Nachfrage nach einem Konsumgut zu einem höheren Preis für dieses Gut führt, so gibt sie uns (selbst wenn sie quantitativ formuliert wird) normalerweise noch lange keine Wahrscheinlichkeiten für unterschiedliche Preise an.

Um zu konkreten Wahrscheinlichkeiten zu gelangen, müssen Bayesianer diese Spielräume dann doch subjektiv ausfüllen. Das nächste Problem ist die mangelnde Likelihoodanbindung im Rahmen des »old-evidence«-Problems, das wir in Kapitel 5.8.2 aufgreifen werden.

5.3.12 Mehrfaches Updaten und bayesianische Konvergenz

Akzeptieren wir die Likelihoodanbindung, so können wir relativ schnell zu ersten Konvergenzüberlegungen im Rahmen des Bayesianismus kommen. Die sind wichtig für den Ansatz, weil sie das Hauptargument dafür sind, dass die Subjektivität der Startwahrscheinlichkeiten nicht wirklich bedeutsam ist; denn bei einigermaßen eindeutigen Daten nivellieren sich die unterschiedlichen Startwahrscheinlichkeiten verschiedener Personen durch wiederholtes Updaten schnell und führen langfristig zu einheitlichen Bewertungen (vgl. die Konvergenztheoreme dazu in Kap. 5.6.2 und 5.6.3). Das kann man mit Hilfe der Likelihoodanbindung schnell überprüfen und es gibt einfache Programme dazu etwa auf der Seite des »Bayes' Theorem Calculator« (http://statpages.org/bayes.html), mit deren Hilfe wir das weiter ausprobieren können. Aber in entsprechender Weise lassen sich diese Berechnungen natürlich auch (mit höherer Präzision) mit Hilfe eines Tabellenkalkulationsprogramms wiedergeben. Wir benötigen dafür eine Liste von einander ausschließenden aber erschöpfenden Hypothesen $H_1,\ldots,H_n$ und ihren Startwahrscheinlichkeiten mit $\sum P(H_i) = 1$, die jeweils einem bestimmten Datum E eine bestimmte bedingte Wahrscheinlichkeit $P(E|H_i)$ geben. Der Rest lässt sich dann berechnen. Betrachten wir dazu ein kleines Beispiel.

Zunächst wird deutlich, dass die Reihenfolge des Updatens keine Rolle spielt. Sehen wir uns also das Updaten zuerst mit E und dann mit E' an:

$$P^{++}(H) = P^{+}(H|E') = P^{+}(H\&E')/P^{+}(E') =$$
$$P(H\&E'|E) \;/\; P^{+}(E'|E) = P(H\&E'\&E)/P(E'\&E) = P(H|E\&E')$$

Es kommt dasselbe heraus, ob wir nacheinander mit E und dann mit E' updaten oder gleich mit E&E'. Schauen wir uns nun die Daten eines Beispiels an, wobei wir immer wieder mit demselben Datum E (etwa für einmal Kopf) updaten, d.h. *es kommt etwa immer wieder Kopf* und wir

möchten ermitteln, welche der 3 Hypothesen in unserem Fall die richtige ist. Wir starten mit 3 Hypothesen und festen Likelihoods für E:

Ein konkretes Beispiel
$P(H_1) = 0{,}2$ $P(H_2) = 0{,}001$ $P(H_3) = 0{,}799$ und
$P(E|H_1) = 0{,}3$ $P(E|H_2) = 0{,}8$ $P(E|H_3) = 0{,}7$

Wir sehen hier, dass H2 relativ noch die beste Vorhersagen für unser Datum E abgibt, aber was die Startwahrscheinlichkeiten angeht, eindeutig auf dem letzten Platz liegt. Mit »E« ist also z.B. das Auftreten von Kopf gemeint, und die Hypothesen vergeben dafür unterschiedliche bedingte Wahrscheinlichkeiten bzw. Likelihoods, wobei wir aber bereits annehmen, dass eine der drei Likelihoods für Kopf die richtige ist.

Wenn etwa die Hypothese 2 die richtige sein sollte, dann wird sie sich letztlich durchsetzen, obwohl wir ihr zunächst nur eine sehr geringe Plausibilität eingeräumt haben. Im bayesianischen Verfahren wächst die Wahrscheinlichkeit der richtigen Hypothese mit hoher Wahrscheinlichkeit gegen 1. Letztlich konzentriert sich also im Normalfall die Wahrscheinlichkeit auf die richtige Hypothese (wobei wir hier immer davon ausgehen, dass sie in unserer Liste tatsächlich enthalten ist). Wie entwickeln sich nun die Hypothesenwahrscheinlichkeiten, wenn E einfach mehrfach auftritt? Das zeigt die folgende Tabelle:

Anzahl	H_1	H_2	H_3	Anzahl	H_1	H_2	H_3
1	0,097	0,0013	0,9	40	0	0,2	0,8
2	0,044	0,0016	0,95	50	0	0,5	0,5
8	0,0003	0,0036	0,996	80	0	0,98	0,02
10	0,0001	0,0047	0,995	100	0	0,995	0,005
30	0	0,06	0,94	120	0	0,999	0,001

Tabelle 5.4: Ein Beispiel zur bayesianischen Konvergenz für das Datum x-mal Kopf

Man sieht hieran, dass die Hypothese H_1 mit den schlechtesten Vorhersagen relativ schnell falsifiziert wird. Sie sinkt von 20% schon nach zwei Datenupdates auf ca. 4% und ist nach 10 Wiederholungen praktisch bei Null und damit falsifiziert. Hypothese H_2, die eigentlich die besten

Vorhersagen abgibt, hat dagegen sehr große Mühe, zunächst aus dem Keller zu kommen, was vor allem daran liegt, dass sie in H_3 einen sehr starken Konkurrenten hat, der zu Beginn viel besser dasteht mit ca. 80% Wahrscheinlichkeit und nur ein wenig schlechter in seinen Vorhersagen ist als H_2. Daher sieht lange Zeit H_3 wie der klare Sieger aus und verliert erst nach 50 Wiederholungen langsam gegen H_2.

Man könnte hier sagen, dass uns der Bayesianismus lange Zeit in die Irre führt, weil wir H_2 eine so »miese« Ausgangswahrscheinlichkeit gegeben hatten. Der Likelihoodist, der nur die Likelihoods betrachtet (bzw. den Quotienten $P(E|H_2)/P(E|H_3)$, würde das als Wasser auf seine Mühlen ansehen und einwenden, dass wir ohne die weiteren subjektiven Anteile des Bayesianer besser gefahren wären. Dann hätte sich schon viel schneller H_2 als die beste Hypothese erwiesen.

n	1	2	8	15	30	40	50	80	100	120
Quotient:	1,14	1,31	2,91	7,41	54,9	209	794	43587	629788	9099838

Tabelle 5.5: Die Likelihoodquotienten P(E|H2)/P(E|H3)

Bereits ab ca. 15 Wiederholungen wird hier die klare Überlegenheit der Hypothese H_2 gegenüber der Hypothese H_3 deutlich.

Allerdings wird das in beiden Ansätzen wieder deutlich schwieriger, wenn wir es mit *gemischten Ergebnissen* zu tun bekommen. Sollte H_2 korrekt sein, müssten wir im Durchschnitt mit 80% Köpfen, aber in 20% der Fälle auch mit Zahl rechnen. Für diese Verteilung finden wir dann die obigen irreführenden Effekte noch in verstärkter Form:

Anzahl	H_1	H_2	H_3	Anzahl	H_1	H_2	H_3
1	0,13	0,0011	0,87	250	0	0,43	0,56
30	0	0,0027	0,997	300	0	0,74	0,26
50	0	0,0045	0,995	400	0	0,97	0,03
100	0	0,0161	0,98	500	0	0,998	0,002
200	0	0,17	0,82	600	0	0,9998	0,0002

Tabelle 5.6: Ein Beispiel zur bayesianischen Konvergenz
(bei E zu ¬E im Verhältnis 8 zu 2)

Bei diesen gemischten Daten setzt sich H_2 noch viel langsamer gegenüber H_3 durch, doch letztlich erkennen wir auch hier wieder die

Konvergenz der Wahrscheinlichkeit von H_2 gegen 1 bzw. die Falsifikation der anderen Hypothesen. Allerdings benötigen wir hier viel mehr Daten, um die Qualität von H_2 zu erkennen. Die Irreführung ist hartnäckiger geworden. Das gilt in ähnlicher Form für die Likelihoodquotienten:

n	1	2	8	15	30	40	50	80	100	120
Quotient:	1,03	1,05	1,23	1,47	2,16	2,8	3,62	7,83	13,11	21,93

Tabelle 5.7: Likelihoodquotienten P(E|H2)/P(E|H3)
(bei Kopf/Zahl im Verhältnis 8 zu 2)

Statt bei 15 Wiederholungen stellt sich ein signifikanter Unterschied (den manche Likelihoodisten bei ca. 8 erkennen) erst ab ca. 80 Wiederholungen ein. Und dass kann natürlich noch schlechter werden, wenn wir noch Zufallsschwankungen in den Daten berücksichtigen, die womöglich erst bei höheren Anzahlen ausgeglichen werden. (Man vergleiche das auch mit den Interpretationen des Bayes-Faktors in Kapitel 6.)

Doch auch hier ist die angestrebte Konvergenz erkennbar, die sich auch allgemeiner beweisen lässt (s.u.). Die Frage bleibt allerdings (und wird später in einer Glosse verarbeitet), ob sich die erwünschte Konvergenz schnell genug einstellt, wenn wir mit sehr unterschiedlichen Ausgangswahrscheinlichkeiten arbeiten. Der objektive Bayesianer oder der induktive Logiker setzt genau an dieser Stelle an und wird gegen die Vorgehensweise des subjektiven Bayesianers einwenden, dass wir ohne sehr gute Gründe niemals so unterschiedliche Vorher-Wahrscheinlichkeiten vergeben dürfen. Vielmehr sollte ein Indifferenzprinzip (vgl. nächster Abschnitt) zu einer möglichst gleichen Verteilung von jeweils 1/3 als Startwahrscheinlichkeit führen. Dann findet sich selbstverständlich ein viel besseres Konvergenzverhalten selbst bei gemischten Daten. So hat H2 dann bereits nach 100 Wiederholungen eine Wahrscheinlichkeit von ca. 93% (statt 2% bei ungleichen Startbedingungen).

Die Likelihoodanbindung kann allerdings nicht für immer aufrechterhalten werden, denn P(E|H) ist eine epistemische Wahrscheinlichkeit, und sobald uns E erst einmal bekannt geworden ist und wir also P(E) = 1 setzen, wird natürlich auch P(E|H) = 1. Diese Problematik werden wir später wieder aufgreifen.

Noch ein Wort zur Redeweise im Umgang mit Likelihoods, die leider nicht einheitlich ist. Im Folgenden werde ich meist kurz und intuitiv

davon sprechen, dass die Daten E eine Likelihood von r haben gegeben H, wenn gilt: P(E|H) = r; oder wir sagen dazu, das Datum E wäre im Maße r »*likely*« bzw. zu *erwarten* oder auch einfach *wahrscheinlich* gegeben H. Der klassische Statistiker möchte hier nicht unbedingt von Wahrscheinlichkeit sprechen (da sie nur hypothetisch ist und sich nicht als aktuale relative Häufigkeit wiederfinden lässt) und wählt lieber die Likelihood-Redeweise. Jedenfalls soll ab jetzt das Prinzip der Likelihoodanbindung angenommen werden, wonach wir in den Fällen, in denen objektive Likelihoods vorliegen, diese übernehmen.

5.3.13 Das Hauptprinzip

Weit bekannter als die Likelihoodanbindung ist das sogenannte Hauptprinzip (»principal principle«), das vor allem von David Lewis propagiert wurde. Ein Vorläufer davon findet sich in »Millers Prinzip« (das Miller (1966) allerdings als inkonsistent ablehnte). Die Frage ist hier, welchen Glaubensgrad wir in den Fällen annehmen sollten, in denen wir über noch stärkere Informationen als die oben genannten verfügen, aber zugleich möglicherweise gegenläufige Informationen erhalten? Nehmen wir einmal an, wir wüssten schon, dass die *objektive Wahrscheinlichkeit* oder *Chance* (wir werden später auch von *Propensität* sprechen) für ein konkretes Ereignis A in einem bestimmten Zeitraum genau r sei. Über die Problematik solcher physikalischen Einzelfallwahrscheinlichkeiten werden wir später noch diskutieren. Nehmen wir sie an dieser Stelle einfach einmal als gegeben hin.

Als Beispiel könnten wir an den radioaktiven Zerfall denken. Ein bestimmtes Atom hat in einem vorgegebenen Zeitraum eine bestimmte Wahrscheinlichkeit zu zerfallen. In dem Fall verfügen wir über gute Gründe für die Annahme, dass es sich dabei tatsächlich um einen genuin indeterministischen Vorgang handelt, so dass die Chance für einen Zerfall eine objektive physikalische Größe darstellt. Außerdem liegen hier nicht nur die trivialen Wahrscheinlichkeiten 0 und 1 vor.

Da es etwas umständlich ist, immer über den radioaktiven Zerfall zu sprechen, soll uns als Beispiel ein Münzwurf genügen, und wir nehmen einfach an, dass dieser gleichfalls genuin indeterministisch sei. Wird die Münze aus größerer Höhe geworfen, finden viele Zusammenstöße mit

Luftmolekülen statt. Diese Ereignisse finden zum einen auf der Mikro-Ebene statt und unterliegen somit den Regeln der Quantenmechanik, zum anderen haben wir es hier mit großen unkalkulierbaren Anzahlen von Zusammenstößen zu tun, weshalb die Indeterminiertheitsannahme jedenfalls nicht ganz unplausibel ist. Nehmen wir also an, die objektive Chance für Kopf (K) beim nächsten Wurf sei 0,4. Die Münze ist also wieder einmal nicht fair, sondern vermutlich in irgendeiner Weise asymmetrisch. Die Art des Werfens sei dagegen fair und bevorzuge keine Seite.

Wenn wir nun nichts anderes wissen, als dass gilt ch(K) = 0,4 (Chance von K sei 0,4), so sollten wir auch unsere subjektive Wahrscheinlichkeit P(K) = 0,4 wählen. Das nennt man in allgemeinerer Form gerade Millers Prinzip, und es wirkt recht plausibel.

Millers Prinzip: $P(A|ch(A)=r) = r$

Es besagt schlicht und einfach: Wenn wir schon annehmen, dass A eine objektive Wahrscheinlichkeit von r besitzt, wahr zu sein, so sollte unsere subjektive Wahrscheinlichkeit das widerspiegeln. Das würde in der bisherigen Form allerdings noch nichts darüber sagen, ob wir über weiteres Wissen über A verfügen.

Um damit realistischere Situationen erfassen zu können, hat David Lewis es zu dem sogenannten Hauptprinzip ausgebaut. Dabei wird angenommen, dass wir noch eine ganze Menge an weiteren Informationen E haben dürfen, die sogar das Auftreten von E betreffen können und die sogar in eine ganz andere Richtung weisen dürfen. Dann sollen wir trotzdem nach dem Hauptprinzip weiterhin an r als subjektiver Wahrscheinlichkeit festhalten. Solche zusätzlichen Informationen, für die das Hauptprinzip gilt, werden als die *zulässigen Informationen* relativ zu A bezeichnet.

Das Hauptprinzip: $P(A|ch(A)=r \,\&\, E) = r$

Die Idee beim Münzwurf ist etwa: Wenn wir definitiv wüssten, dass die Chance für Kopf beim nächsten Wurfe genau 0,4 ist, dann würde dieses Wissen viele andere Informationen, die wir sonst noch über

den Münzwurf haben könnten, aus dem Felde schlagen. Etwa die Information, dass in den bisherigen Würfen eine Quote von 80% Köpfen mit dieser Münze erzielt wurde.

Man muss sich allerdings erst einmal vorstellen, wir verfügten tatsächlich über dieses besondere Wissen für die tatsächliche Chance für Kopf bei unserem nächsten Wurf. Das fällt uns nicht so ganz leicht, da dieses Wissen im Normalfall nicht vorliegen wird. Am ehesten könnten wir solche Werte aus bestimmten Theorien erhalten, die wir akzeptieren und deren Vorhersagen bestimmter Wahrscheinlichkeiten wir deshalb ebenfalls akzeptieren würden. Nehmen wir hier einmal der Anschaulichkeit halber an, der liebe Gott würde uns mit der wahren Wahrscheinlichkeit für Kopf versorgen. Er würde sagen: »Es stimmt schon, dass der Werfer bisher eine sehr gute Trefferquote von 0,8 für Kopf hatte, aber ich garantiere, dass die reale Chance für Kopf beim nächsten Wurf nur 0,4 ist.« Welche Wetten auf Kopf sollten wir dann akzeptieren? Nicht mehr als Quoten von 4:6, denn entscheidend bleibt doch wohl die reale Wahrscheinlichkeit für Kopf und nicht, was der Werfer früher erreicht hat. Das sind doch bestenfalls Indikatoren, aus denen wir die wahre Wahrscheinlichkeit ablesen möchten.

Wenn uns der liebe Gott oder zuverlässige wissenschaftliche Theorien nicht helfen, sind wir natürlich immer ganz auf die früheren Quoten angewiesen und verfügen eben nicht über die Informationen über die tatsächlichen objektiven Chancen. Doch in unserem fiktiven Fall haben wir diese spezielle Information, die das alles aus dem Felde schlägt. Das scheint klar zu sein, sobald wir uns mit dieser recht ungewöhnlichen Situation einmal vertraut gemacht haben.

Es gibt aber auch *nicht-zulässige Informationen*. Gott könnte etwa sagen: »Die objektive Chance für A ist zurzeit 0,4, aber ich kann bereits in die Zukunft sehen und weiß, worauf die Münze fallen wird. Sie wird nämlich tatsächlich auf Kopf fallen.« Solche Art hellseherischer Informationen sind unzulässig. Sie könnten natürlich unsere bisherigen Informationen über die tatsächlichen Chancen übertrumpfen. Um das Prinzip besser zu verstehen und schließlich zum Verständnis der beteiligten Begriffe von Wahrscheinlichkeit beizutragen, müssten wir nun genauer spezifizieren, welche Informationen zulässig sind. Dafür gibt es keine einfache Abgrenzung, aber doch einige klare Fälle. Die

Idee ist, dass übliche Informationen über unsere Welt, wie wir sie eben normalerweise erlangen können, durchaus zulässig sind. Dazu gehören (nach Lewis 1980) zum einen alle Informationen über Ereignisse vor dem Zeitpunkt t, zu dem die Chancen bestehen, und zum anderen all unsere Theorien zum Zeitpunkt t über gesetzesartige Zusammenhänge, die auch zukünftige Chancen betreffen können.

Doch selbst wenn diese Prinzipien plausibel erscheinen, können wir uns natürlich trotzdem fragen, warum wir sie akzeptieren sollten, und das muss etwas mit der Trefferquote zu tun haben, die wir erzielen, wenn wir uns danach richten, im Vergleich dazu, wenn wir andere Verfahrensweisen nutzen. Einige Autoren (s. dazu Strevens 1999) haben argumentiert, dass objektive Wahrscheinlichkeiten einen guten Führer durch unser Leben abgeben. Die Idee ist die folgende: Wenn ch(A) = r ist, so dürfen wir zumindest auf lange Sicht erwarten, dass A mit der relativen Häufigkeit r eintritt. Diese Kenntnis kann uns helfen, wichtige Entscheidungen zu treffen. Führt etwa ein BWL-Studium mit einer objektiven Wahrscheinlichkeit von 90% für mich zu einem hohen Einkommen, so könnte das ein guter Grund für mich sein, das BWL-Studium aufzunehmen, denn in den meisten Fällen wird sich das Studium schließlich auszahlen. Hätte es dagegen nur eine Quote von 10%, wäre der Grund zumindest deutlich schwächer. Wir müssten das jeweils mit den Quoten der alternativen Studiengänge vergleichen, einen hochbezahlten Job zu erreichen. (Man beachte dabei: An dieser Stelle erschließen wir aus den objektiven Wahrscheinlichkeiten bestimmte Häufigkeiten und daraus wieder epistemische Wahrscheinlichkeiten, die unser Handeln anleiten. Für den zweiten Schritt benötigen wir schon wieder den statistischen Syllogismus.)

Doch wie gut ist die Leitfunktion objektiver Wahrscheinlichkeiten tatsächlich? Strevens (1999) kommt schließlich zu einem skeptischen Ergebnis. Das liegt vor allem daran, dass die Aussage B ≡ [ch(Das BWL-Studium führt für mich zu hohem Einkommen.) = 0,9] selbst so wenig greifbar ist. Wir verstehen sie nicht richtig (vgl. Kap. 5.4) und behelfen uns etwa mit Formulierungen wie: »auf lange Sicht treten die relativen Häufigkeiten x auf«. Hätten wir einfach die Aussage A ≡ »Das BWL-Studium führt immer zu hohem Einkommen«, so wären wir auf der sicheren Seite und könnten daraus direkt unsere Konsequenzen ziehen.

Im Falle der Aussage B geht das nicht. Fritz könnte uns fragen, warum er nun BWL studieren sollte. B sagt doch nur etwas über große Anzahlen und auch für die gibt es eigentlich wieder nur eine Wahrscheinlichkeit an, dass sich das BWL-Studium für die meisten auszahlt. Doch wie hilft ihm das für seinen eigenen Fall weiter? Er steht nur *einmal* vor dieser Entscheidung und ist nur eine Person. Soll er also erwarten (=subjektive Wahrscheinlichkeit dafür ist hoch), dass sein BWL-Studium sich auszahlt? Das besagt gerade das Hauptprinzip.

Genau genommen ist unsere ganze Argumentation an dieser Stelle zirkulär. Wir möchten das Prinzip rechtfertigen, dass wir bei einer hohen objektiven Wahrscheinlichkeit für A nun A auch erwarten sollten. Eine hohe Wahrscheinlichkeit für A führt aber nur zu der Konsequenz, dass mit *hoher objektiver Wahrscheinlichkeit* in den meisten Fällen auch A auftreten wird. Sollten wir daher A *erwarten*? Das war eigentlich unsere Ausgangsfrage. Wir sind hier tatsächlich auf das Hauptprinzip angewiesen, um die Frage positiv beantworten zu können. Dann bleibt fast nur der Ausweg, den David Lewis vorschlägt, dass zumindest eine bestimmte Konsequenz dieser Art bereits zum Begriff der objektiven physikalischen Wahrscheinlichkeit dazugehört. Wir denken uns objektive Wahrscheinlichkeit so, dass sie u.a. einen Leitfaden für unsere Entscheidungen abgibt.

Darauf müssen wir in Kapitel 5.4 noch einmal zurückkommen. So wird selbst für den Probabilisten oder spezieller den subjektiven Bayesianer die Frage bedeutsam, was wir unter einer objektiven physikalischen Wahrscheinlichkeit verstehen sollten. Leider haben Wahrscheinlichkeitsaussagen keine Konsequenzen, die nicht selbst wieder mit dem Vorbehalt der Wahrscheinlichkeit versehen sind. Insbesondere geben Wahrscheinlichkeitsaussagen keine klaren Vorhersagen ab, sondern wiederum nur *wahrscheinliche* Vorhersagen. Ohne zu verstehen, was mit objektiver Wahrscheinlichkeit gemeint ist, sind wir dann noch nicht wirklich weitergekommen. Strevens (1999) vergleicht dieses Problem mit dem ursprünglichen Humeschen Induktionsproblem und vertritt hier den Skeptiker.

Strevens (1999) hat mit diesen Überlegungen strikt betrachtet Recht, aber wir arbeiten bereits auf eine bestimmte intuitive Weise mit dem Begriff der Wahrscheinlichkeit, auf die wir uns hier stützen dürfen.

Danach kommt uns eine Frage wie die folgende schon sehr seltsam vor:

> (F) Natürlich habe ich eine objektive Wahrscheinlichkeit von 99% eine Querschnittslähmung davonzutragen, wenn ich von dieser Höhe ins Wasser springe, aber weshalb sollte das ein Grund sein, es nicht zu tun?

Nehmen wir an, es sei klar, dass der Fragende die Querschnittslähmung als genauso unerfreulich betrachtet wie wir. Wenn er dann trotzdem nicht versteht, wieso hier ein (prima-facie) Grund vorliegt, den Sprung zu unterlassen, so scheint er nicht richtig zu verstehen, was wir mit *objektiver Wahrscheinlichkeit* meinen. (Das scheint analog zum Fall des statistischen Syllogismus zu sein, und schließlich führen normalerweise die entsprechenden objektiven Wahrscheinlichkeiten auch zu den relativen Häufigkeiten, die wir anschließend im Rahmen des statistischen Syllogismus entsprechend nutzen würden, nur dass der nicht so stark ist, weitere Informationen aus dem Felde zu schlagen.)

Wir nehmen jedenfalls eine gewisse analytische Verbindung zwischen objektiver Wahrscheinlichkeit und Erwartbarkeit an, die in Frage (F) verletzt wird. Deshalb erscheint sie uns so seltsam. Natürlich wäre sie noch seltsamer, wenn die Querschnittslähmung mit Sicherheit eintreten würde. Es bleibt ein gewisser Unterschied zwischen Wahrscheinlichkeitsaussagen und sicheren Aussagen zurück, doch die hier behaupteten Zusammenhänge sind ebenso wenig von der Hand zu weisen.

Allerdings ist das Hauptprinzip – wie schon erwähnt – eher theoretischer Natur, weil wir normalerweise keine zuverlässigen Informationen über objektive Einzelfallwahrscheinlichkeiten besitzen. Es gibt für uns nur eine Zielvorstellung dafür ab, was wir mit entsprechenden Informationen anfangen könnten, und liefert uns gute Gründe entsprechende Informationen anhand von statistischen Daten und Theorien zumindest zu schätzen, weil sie unser Leben anleiten können. Es erläutert damit die Bedeutung des Konzepts der objektiven physikalischen Wahrscheinlichkeit und bietet so zusätzliche Gründe, Theorien mit objektiven Wahrscheinlichkeiten anzustreben.

Häufiger werden wir jedoch auf den statistischen Syllogismus zurückgreifen können, der allerdings ähnliche Rechtfertigungsprobleme

mit sich bringt. Man könnte das sogar so ausdrücken: Wir nutzen die relativen Häufigkeiten als unseren besten objektiven Hinweis auf eine bestimmte objektive Wahrscheinlichkeit für den Einzelfall, jedenfalls solange keine anderen Informationen vorliegen. Dann nutzen wir diese objektive Einzelfallwahrscheinlichkeit im Sinne des Hauptprinzips, um damit unsere Erwartungen festzulegen. Doch das Hauptprinzip ist an dieser Stelle noch deutlich stärker, denn es lässt zu, dass wir zusätzlich noch über viele weitere Informationen verfügen. Dann ist der statistische Syllogismus aber nicht mehr anwendbar.

Die Stärke der Chancen-Information bringt den Bayesianer aber auch in Schwierigkeiten, denn eigentlich sollen wir doch mit relevanten Informationen jeweils *updaten* und diese nicht einfach links liegen lassen. Doch die Frage ist dann, wann wir mit dem Updaten beginnen und wie lange wir unsere Chancen-Information schlicht dazu nutzen, um damit andere Informationen als irrelevant zurückzuweisen. Eigentlich ist das Hauptprinzip nur dazu gedacht, bestimmte Startwahrscheinlichkeiten festzulegen, mit denen wir dann das bayesianische Update-Spiel beginnen können. Betrachten wir dazu noch einmal das Beispiel des möglichen Dschungelfiebers von Herrn Vorsichtig und denken uns dazu zwei unterschiedliche Situationen. Nur soll diesmal Herr Vorsichtig schon vor seinem Urlaub beurteilt werden und der Test ist ein Test für die Empfänglichkeit von Dschungelfieber. Ansonsten seien alle Werte dieselben.

Die Daten der Berechnung
$P(H) = 1/1000$ $P(\neg H) = 999/1000$
$P(T^+|\neg H) = 3/100$ $P(T^+|H) = 99/100$
Resultat: $P(H|T^+) = 0{,}032$

In Situation S1 wird die Wahrscheinlichkeit für Herrn Vorsichtig diesmal nicht anhand des statistischen Syllogismus, sondern anhand des Hauptprinzips festgelegt. Nehmen wir also an, er hätte die objektive Wahrscheinlichkeit von 1/1000 dafür, Dschungelfieber im Urlaub zu entwickeln. Das nehmen wir als Ausgangspunkt für weitere Überlegungen. Kommt dann das positive Testergebnis (für erhöhte Anfälligkeit für Dschungelfieber) hinzu, müssen wir eigentlich damit updaten und

erhalten als neue Wahrscheinlichkeit für Herrn Vorsichtig im Urlaub an Dschungelfieber zu erkranken nun 0,032.

Situation S2 ist ähnlich, aber wir erhalten das positive Testergebnis zugleich mit der Information über seine objektive Wahrscheinlichkeit von 1/1000 für Dschungelfieber. Das positive Testergebnis ist sicher eine zulässige Information im Sinne des Hauptprinzips. Sie wird also im Sinne des Hauptprinzips durch die objektive Wahrscheinlichkeit aus dem Felde geschlagen. Statt updaten zu müssen, müssen wir dann bei der Wahrscheinlichkeit 0,001 für Dschungelfieber stehenbleiben.

Das heißt, je nachdem, wann und in welcher Situation uns bestimmte Informationen erreichen, werden sie entweder mit Hilfe der Konditionalisierungsregel zum Updaten verwendet oder durch das Hauptprinzip einfach übertrumpft. Hauptprinzip und klassischer Bayesianismus verfahren eben ganz unterschiedlich mit Informationen und passen daher nicht wirklich zusammen. In den beiden Situationen erhalten wir jedenfalls unterschiedliche Endwahrscheinlichkeiten, obwohl wir in beiden Situationen insgesamt über dieselben Informationen verfügen. Der Bayesianismus passt eben besser zu Prinzipien, die für den Fall gedacht sind, dass keine weiteren Informationen vorliegen, wie das im statistischen Syllogismus der Fall war oder wie das auch für die Indifferenzprinzipien z.T. der Fall ist. Bayesianer können sich allerdings damit trösten, dass es sich beim Hauptprinzip eher um ein theoretisches Prinzip handelt, da wir in der Praxis kaum über so starke Informationen verfügen, wie sie im Hauptprinzip vorausgesetzt werden.

Das Paradox der schlafenden Schönen. Es gibt auch paradoxe Situationen, in denen wir nicht genau wissen, ob wir das Hauptprinzip anwenden können oder nicht. Die wohl prominenteste ist die der »sleeping beauty«, die von Adam Elga (2000) entworfen wurde. Nehmen wir an, Anna (die schlafende Schöne) nimmt an einem Experiment teil, bei dem sie am Sonntag in den Schlaf versetzt wird. Dann wird eine faire Münze geworfen und bei Kopf wird Anna nur am Montag kurz geweckt und schläft dann bis Donnerstag. Bei Zahl wird Anna am Montag kurz geweckt und so wieder in den Schlaf versetzt, dass sie sich daran nicht erinnern kann. Anna kennt alle Bedingungen des Experiments. Bei jedem Aufwecken, wird Anna nun gefragt, für wie wahrscheinlich sie es hält,

dass der Münzwurf Kopf ergeben hat. Was sollte sie vernünftigerweise antworten? Was sollte ihr Glaubensgrad sein?

Wir können dem Hauptprinzip folgen und argumentieren: Anna weiß, dass die Münze fair ist, also sollte sie antworten 1/2, zumal sie keine wirklich neuen Informationen über die Münze erhalten hat. Außerdem würden die durch das Hauptprinzip vermutlich aus dem Felde geschlagen. Die meisten Philosophen, die darüber geschrieben haben, scheinen allerdings anderer Meinung zu sein und antworten stattdessen: 1/3. (Aber einige wie Roger White (2006) verteidigen auch überzeugend die 1/2-Antwort.)

Die 1/3-Befürworter argumentieren: Es gibt drei Situationen, in denen Anna geweckt wird (Kopf-Montag, Zahl-Montag, Zahl Dienstag). Anna hat keine Hinweise darauf, in welcher davon sie sich befindet. In nur einem davon liegt auch Kopf vor. Also sollte Anna davon ausgehen, dass in 2/3 der Fälle gerade Zahl aufgetreten ist und daran sollten sie ihre aktuellen epistemischen Wahrscheinlichkeiten orientieren. Solche Fälle belegen, dass es nicht immer leicht ist zu erkennen, ob das Hauptprinzip anzuwenden ist. Eventuell ist die Information, die Anna erhält, nämlich die über das Experiment plus der Information, dass sie tatsächlich wach ist, doch ein Hinweis darauf, dass de facto wohl eher Zahl gekommen ist. Dann könnte es sich nach dem Hauptprinzip um eine unzulässige Information handeln und es wäre demnach nicht mehr anwendbar. Doch solchen extremen Situationen kann ich hier nicht weiter nachgehen.

Schön wäre es auch, wenn es ein sinnvolles einfaches Prinzip in der anderen Richtung gäbe, nach dem wir von bestimmten epistemischen Wahrscheinlichkeiten auf objektive physikalische Wahrscheinlichkeiten schließen könnten, doch das ist normalerweise kaum möglich. Ein solches Prinzip hätte darüber zu entscheiden, wann unsere Plausibilitätsgrade gut genug gestützt wären, so dass wir gute Gründe für die Annahme hätten, wir könnten damit einigermaßen zuverlässig objektive Wahrscheinlichkeiten schätzen. Das setzt eine anspruchsvolle Metabewertung unserer epistemischen Wahrscheinlichkeiten voraus, über die wir im Normalfall leider nicht verfügen. Um auf objektive Zusammenhänge schließen zu können, benötigen wir vielmehr *objektive Daten*, etwa um darauf Hypothesen über kausale Tendenzen im Sinne von Chancen zu stützen, und wir werden dazu im letzten Kapitel etwa

die Differenzmethode kennenlernen, die uns typischerweise die erforderlichen Daten liefert, um relativ direkt auf kausale Zusammenhänge zu schließen.

5.3.14 Das Reflexionsprinzip

Auch das letzte Prinzip in dieser Rubrik hat eher einen theoretischen Charakter und wird mehr der Vollständigkeit halber erwähnt, und weil es etwas über die Qualität von Dutch-Book-Argumenten aussagen kann. Es handelt sich eigentlich nicht um ein Wahrscheinlichkeitskoordinierungsprinzip, weil es nur Glaubensgrade miteinander verbindet, aber da es ebenfalls weitere Einschränkungen für die Glaubensfunktionen bieten kann, erwähne ich es kurz an dieser Stelle. Gemeint ist das sogenannte *Reflexionsprinzip*, das auf van Fraassen (1984) zurückgeht und dort durch ein diachronisches Dutch-Book-Argument begründet wird. Wenn wir schon heute wüssten, dass wir in Zukunft einen bestimmten Glaubensgrad für A haben werden $P^+(A) = r$, dann sollten wir gemäß diesem Prinzip schon heute den entsprechenden Glaubensgrad übernehmen:

Reflexionsprinzip: $P(A|P^+(A)=r) = r$

Doch das Prinzip ist keineswegs unproblematisch. Es könnte passieren, dass meine späteren Glaubensgrade daran kranken, dass ich inzwischen bestimmte Daten vergessen habe oder dass ich senil geworden bin oder – noch einfacher – dass ich schlicht und einfach irreführenden Daten aufsitzen werde, obwohl ich diese verantwortungsvoll ausgewertet habe. Es ist etwa ein glaubwürdiger Alibizeuge aufgetaucht, der uns aber tatsächlich anlügt. Zumindest die ersten Probleme mit dem Reflexionsprinzip können vielleicht dadurch ausgeräumt werden, dass wir es abschwächen und nur noch solche zukünftigen Glaubensgrade zu unserem heutigen Maßstab erheben, die auf rationale Weise entstanden sind. Die Definition von »rational« ist dabei natürlich wieder mit all den hier diskutierten Problemen belastet. Das letztgenannte Problem erledigt sich nicht so einfach. Jedenfalls ist es nicht zweifelsfrei rational, heute schon Glaubensgrade anzunehmen, die durch Daten entstehen werden,

die ich aus heutiger Sicht nicht kenne und daher nicht bewerten kann. Wir können uns nicht sicher sein, dass unsere zukünftigen epistemischen Wahrscheinlichkeiten wahrheitsnäher sind als unsere heutigen.

Mein Problem ist aber ein noch anderes: Wenn ich heute in meinen Glaubensgraden nicht mit meinen zukünftigen übereinstimme, kann wieder ein diachronisches Dutch-Book gegen mich konstruiert werden – gleichgültig, wie vernünftig meine zukünftigen Glaubensgrade sein werden. Nehmen wir an, meine Glaubensgrade für A seien die folgenden:

(1) $P(A) = 0{,}4$
(2) $P^{+}(A) = 0{,}6$

Dann muss ich heute bzw. in der Zukunft die folgenden beiden Wetten jeweils um einen Topf von 10 Euro akzeptieren:

(1) Wette 1 (heute): 6 Euro auf ¬A
(2) Wette 2 (in Zukunft): 6 Euro auf A

Wiederum ist klar, dass ich 12 Euro bezahle, aber nur 10 Euro erhalten werde und damit in jedem Fall mit einem Verlust zurückbleibe. Stellen für uns solche diachronischen Dutch-Books (diachronisch, weil die Wetten zu verschiedenen Zeitpunkten stattfinden) aber einen guten Grund dar, unsere heutigen Glaubensgrade schon den morgigen anzupassen, so müssten wir für das einfache Reflexionsprinzip eintreten, das wir oben aufgeführt haben, ohne zusätzliche Rationalitätsforderungen. Das Reflexionsprinzip ist aber – wie die Beispiele zeigen – nicht sehr plausibel und könnte daher auf andere Weise ausgebeutet werden.

Nehmen wir an, ich bin mir heute aufgrund vieler guter Gründe sehr sicher (Wahrscheinlichkeit 99%), dass ein bestimmtes Wunder W nicht eintreten wird, bzw. haben nur eine 1% Wahrscheinlichkeit dafür, dass das Wunder eintritt. Allerdings weiß ich nun, dass ich gezwungen werde, mich einer bestimmten Form von Gehirnwäsche innerhalb einer Sekte zu unterziehen und man kann mir praktisch garantieren, dass ich danach den Glaubensgrad 95% haben werde, dass das Wunder W eintritt. Sollte ich den dann heute schon übernehmen, um die sonst auftretende diachronische Inkohärenz zu vermeiden und damit einem Dutch-Book-Argument zu entgehen? Das scheint erkenntnistheoretisch kaum sinnvoll

zu sein. Vielleicht kann man es durch diese Überlegung praktisch rechtfertigen, aber mit meiner heutigen epistemischen Situation hat es nicht viel zu tun (vgl. dazu auch Christensen 1991).

Es hieße außerdem, dass ich nun heute hohe Summen auf das Auftreten des Wunders gegen kleine Beträge setzen müsste, obwohl doch alle meine bisherigen Kenntnisse das Wunder als praktisch unmöglich erscheinen lassen. Nach meinem heutigen Kenntnisstand würde ich durch solche Wetten bereits heute viel Geld verlieren. Also bleibe ich heute bei meinen jetzigen Überzeugungen und gehe heute einfach keine Wetten gegen das Wunder ein, weil ich schon weiß, dass man die morgen ausschlachten könnte. Meine beste Reaktion auf die Information, dass ich morgen an das Wunder glauben werde, ist also nicht, heute auch schon daran zu glauben, damit ich möglichst kohärent bleibe, sondern einfach nur Enthaltsamkeit zu üben, was Wetten gegen das Wunder angeht, und mir das besonders für die Zukunft fest vorzunehmen.

Das Dutch-Book-Argument »beweist« hier zu viel. Folgen wir ihm, sollten wir unsere morgigen Glaubensgrade schon heute übernehmen, was aber nicht wirklich vernünftig wäre. Es ist jedenfalls nicht zwingend, dass wir heute bereits etwas rationalerweise akzeptieren sollten, was uns morgen als plausibel erscheint, aber eben aufgrund unserer jetzigen Datenlage heute keineswegs bereits wahrscheinlich ist. Vernünftig scheint es eher zu sein, das heute zu akzeptieren, wofür heute unsere Daten sprechen und die alleinige Auskunft, dass wir morgen etwas anders akzeptieren werden, kann kaum als gute Begründung für eine bestimmte Überzeugung herhalten. Dieses Resultat schwächt zugleich die Überzeugungskraft von diachronischen Dutch-Book-Argumenten allgemein. Insbesondere wird auch die Konditionalisierungsregel durch entsprechende diachronische Dutch-Book-Argumente begründet. Die bisherige Debatte belegt nun aber, dass wir diesen Überlegungen nicht zu viel Gewicht geben dürfen.

5.3.15 Das Prinzip der epistemischen Gleichbehandlung und weitere Indifferenzprinzipien

Das bekannteste Verfahren, um zu ersten Wahrscheinlichkeiten zu gelangen, finden wir in den Indifferenzprinzipien, die wir oft anwenden,

ohne uns dessen bewusst zu sein. Wenn ich frage: »Wie hoch ist die Wahrscheinlichkeit mit zwei Würfeln eine Augensumme echt kleiner 5 zu würfeln?«, so werden Sie rechnen: Es gibt die Paare (1,1), (1,2), (2,1) und (2,2), die die Bedingung erfüllen, also erhalten wir: 4/36 = 1/9. Das ist meistens auch genauso beabsichtigt. Aber damit haben wir schon vorausgesetzt, dass beide Würfel fair sind, oder zumindest, dass es für uns rational ist, eine Wahrscheinlichkeit von 1/6 für jede Augenzahl der beiden Würfel anzunehmen und anzunehmen, dass sich die Augenzahlen der beiden Würfel unabhängig voneinander einstellen.

Die erste Annahme wollen wir nun weiterverfolgen. Ich könnte etwa fragen: »Woher weißt Du, dass alle Ergebnisse dieselbe Wahrscheinlichkeit haben?« Darauf ist eine naheliegende Antwort, dass wir bisher *keinen Grund* dafür haben, dass irgendeine Zahl bevorzugt wird, und dann müsste unser bester Tipp doch wohl sein, von einer Gleichverteilung auszugehen. Eine ähnliche Argumentation hatten wir schon im Falle des statistischen Syllogismus kennengelernt. Ich finde diese Argumentation auch durchaus überzeugend, wenn es uns darum geht, zu epistemischen Wahrscheinlichkeiten zu gelangen. Sie stützt sich wieder auf das *epistemische Gleichbehandlungsprinzip*. Leider gibt es eine Reihe von Paradoxien, die das Indifferenzprinzip und seine Verwandten in Verruf gebracht haben, so dass es heute oft abgelehnt wird.

Indifferenzprinzip (IP): Gibt es eine Reihe von Möglichkeiten $m_1,\ldots,$ m_k, die einander ausschließen, aber zusammen erschöpfend sind (eine davon wird also eintreten), und wir haben keinen Grund anzunehmen, dass eine davon besser begründet bzw. wahrscheinlicher ist als eine andere (bzw. schlicht: eine zu bevorzugen), dann sollten wir eine Gleichverteilung darauf annehmen: $P(m_1) = \ldots = P(m_k) = 1/k$.

Das Prinzip stellt zunächst eine Forderung für unsere subjektiven epistemischen Wahrscheinlichkeiten dar. Wir könnten darüber hinaus versuchen, eine objektive Variante des Indifferenzprinzips zu formulieren, die sich darauf stützen müsste, dass es keine objektiven, relevanten

Unterschiede zwischen den Möglichkeiten gibt. Das hieße also im Falle des Würfels, dass er *objektive Symmetrien* aufweist, die keine der Seiten hervorstechen lassen. Das sollte jedenfalls einen positiven Grund dafür bieten, von gleichen Wahrscheinlichkeiten auszugehen.

Doch eine solche objektive Variante verlangt nach stärkerem Hintergrundwissen als (IP) selbst. Wir müssten schon wissen, welche Eigenschaften für das Ergebnis relevant sind, und für die einen Gleichstand verlangen. Ist eine der Augenzahlen etwa gelb und sind die anderen rot aufgedruckt, müssten wir schon wissen oder zumindest begründen können, dass das keinen Einfluss auf das Ergebnis hat, um sagen zu können, dass alle Seiten in relevanten Hinsichten symmetrisch sind. Stattdessen verlangt (IP) nur, dass wir über *keinen Grund* für die Annahme verfügen, dass rot oder gelb einen Vorzug darstellen. Mit (IP) versuchen wir also gerade aus der Abwesenheit von Wissen erste Schlüsse zu ziehen. Das soll hier mein Thema sein. Das grundlegendere Prinzip ist aber das der epistemischen Gleichbehandlung, für das ich daher zunächst argumentieren möchte (vgl. dazu Kap. 5.2).

(EG) Epistemisches Gleichbehandlungsprinzip: Wenn in unserem Überzeugungssystem zwei Aussagen A und B gleich stark begründet sind (oder sogar dieselben Gründe für A wie für B sprechen), so sollten wir A und B epistemisch gleich behandeln (bzw. bewerten).

Es ist natürlich sehr schwierig, für ein so grundlegendes Prinzip epistemischer Rationalität weitere Argumente anzugeben, aber wir können zumindest untersuchen, welche Folgen eine Verletzung oder Aufgabe des Prinzips haben würde. Dabei werden wir auch auf die praktische Konsequenzen schauen müssen, da sich dort erst Folgen zeigen, über deren Bewertung wir klare Vorstellungen haben. Roger White (2010) argumentiert in einer ähnlichen Form und hat noch eine bestimmte Notation eingeführt, die ich übernehmen möchte. Er spricht von symmetrischer Evidenz für A und B und notiert das: $A \approx B$. Für ihn lautet dann das Indifferenzprinzip kurz:

(POI) Das Indifferenzprinzip nach White: $A \approx B \Rightarrow P(A) = P(B)$

Das Gleichbehandlungsprinzip ist dabei so zu verstehen, dass Sicht des epistemischen Subjekts – also nach seiner Einschätzung der Dinge – die Gründe für A und für B jeweils insgesamt betrachtet gleich gut sind. Hat er dazu keine Meinung oder scheinen ihm die Gründe für A doch etwas zu überwiegen, greift das Prinzip einfach nicht. Die Situation kann auftreten, indem viele Gründe für (oder gegen) A und B vorliegen, aber es soll genauso der Grenzfall dazu zählen, dass wir weder für A noch für B gute Gründe haben. (Bei Gregory Novack (2010) finden wir übrigens noch eine allgemeinere Formulierung des Indifferenzprinzips sowie eine tiefergehende Analyse der (formalen) Zusammenhänge mit anderen Prinzipien und eine weitergehende Verteidigung des Prinzips.)

Wenn wir das (EG) zunächst auf *klassische kategorische Überzeugungen* anwenden, sieht es ganz unproblematisch aus und würde wohl kaum von jemandem bestritten. Wenn keine guten Gründe für oder gegen A und B vorliegen, gehören sie eindeutig in den neutralen Bereich. Finden wir Gründe für A, die uns gut genug erscheinen, dass wir A akzeptieren (ablehnen) sollten, dann sollten wir auch B akzeptieren, weil unsere Gründe für B genauso stark sind. Wie ließe sich hier eine unterschiedliche Behandlung von A und B rechtfertigen, wenn doch unser eigenes Urteil lautet, dass wir genauso starke Indizien für die Wahrheit von A wie für die von B haben?

Denken wir einmal an mögliche Anwendungen und Konsequenzen. Beipiel 1: Kommissar X hat genau zwei Verdächtige a und b und ist sich sicher, dass einer von beiden der Täter ist. Gegen beide liegen dieselben Indizien vor, beide haben etwa Fingerabdrücke am Tatort hinterlassen, hatten ein gleichstarkes Motiv für die Tat etc. Trotzdem hält Kommissar X nun a für den Täter (A) und b nicht (¬B). Auf Befragen kann er dafür keine Gründe nennen, noch nicht einmal ein kriminalistisches Bauchgefühl, das wir bei einem erfahrenen Ermittler sehr wohl unter die Wahrheitsindikatoren rechnen könnten. Dann handelt Kommissar X in irgendeinem Sinn falsch – soviel scheint mir offensichtlich zu sein. Er ist aber nicht nur *moralisch unfair* gegenüber a, sondern hier liegt auch ein *epistemischer Fehler* vor, eine Form epistemischer Irrationalität,

die erst die ebenfalls moralisch verfehlte Ungleichbehandlung von a und b verursacht. Wir würden von X erwarten, dass er uns zumindest irgendwelche epistemischen Unterschiede nennen könnte für sein Urteil, sonst ist es erkenntnistheoretisch inakzeptabel und damit auch moralisch falsch.

Beispiel 2: Ein Wissenschaftler W hat zwei Theorien t1 und t2, die einander ausschließen, und er ist sich sicher, dass eine davon wahr ist. Alle Daten, über die er in diesem Zusammenhang verfügt, werden nach seiner Meinung von t1 genausogut erklärt wie von t2. Auch alle anderen seiner Bewertungsmaßstäbe für Theorien zeigen einen völligen Gleichstand zwischen t1 und t2. Trotzdem akzeptiert W t1, lehnt t2 aber ab (oder belässt sie jedenfalls im neutralen Bereich). Damit verstößt er m.E. eindeutig gegen Standards der wissenschaftlichen Redlichkeit. Die würde in so einem Fall klar eine Neutralität zwischen den beiden Theorien verlangen. Auch dieser Verstoß ist vor allem ein Verstoß gegen epistemische Standards wie den der epistemischen Gleichbehandlung. Die Kollegen von W würden zu Recht nach irgendeinem Aspekt oder Datum verlangen, das eher für t1 als für t2 spricht, zu dem es für t2 keine Gegenstück gibt. Kann W das nicht liefern, hätte er sich als Wissenschaftler unseriös verhalten. Diese Standards sollten wir ebenso für andere Bereiche außerhalb der Wissenschaft akzeptieren – zumindest als epistemische Zielvorstellung.

Gehen wir nun einen Schritt weiter und nehmen an, X und W würden ihre Bewertungen durch *komparative Urteile* ausdrücken, wonach eine Aussage *plausibler* als eine andere wäre. Wie könnten sie dann in den Beispielen vorgehen? Sie könnten natürlich sagen, dass A und B unvergleichbar wären. Das müssten wir wohl akzeptieren. Aber wenn sie einen Vergleich vornähmen und dabei käme nicht heraus, dass sie als äquivalent zu beurteilen wären, sondern A wäre ihrer Ansicht nach deutlich plausibler als B, dann kämen wieder unsere obigen Vorwürfe zum Tragen.

Natürlich sollten wir diese Vorstellungen und Standards in entsprechender Weise schließlich auf den probabilistischen Fall übertragen. Das führt das (EG) zum (POI). Wir können die obigen Beispiele in ganz ähnlicher Form diskutieren. Selbst wenn W ein Probabilist wäre, könnte er sich natürlich weigern für diese Theorien Glaubensgrade zu

vergeben. Wäre er dazu aber bereit und dabei käme heraus, dass für ihn P(t1) > P(t2) sei, würden wir ihn wiederum nach einer Erklärung für diesen Unterschied fragen. Kann er keine liefern, sollten wir wieder sagen, dass W sich wissenschaftlich unredlich verhalten hat, denn er verletzt wiederum zu seiner Pflicht einer neutralen, unvoreingenommenen Betrachtung der Theorien. Die Glaubensgrade bieten zwar feinere Abstufungen als die klassischen Einstufungen, aber die Grundproblematik bleibt dieselbe: *Kein Unterschiede in der Bewertung, ohne entsprechende Unterschiede in den Begründungen*. Diese wissenschaftlichen Normen sollten wir wieder als Vorbild für einen rationalen Umgang mit unseren Überzeugungssystemen ansehen. Wenn wir davon abweichen, sind wir in jedem Fall weniger rational, als wenn wir die Norm (EG) einhalten. Man sollte nicht vergessen, dass (EG) uns auch intuitiv als sehr plausible *Rationalitätsforderung* erscheint.

In der Praxis gibt es sicher immer wieder Tricks, um für sich oder andere die eine Theorie doch als die bessere erscheinen zu lassen. Wir können die Indizien für die Theorien einfach etwas anders gewichten. Doch das ist nicht mein Thema. Es geht mir nur um die Zielvorstellung. Wenn sich ein epistemisches Subjekt dazu bekennt, dass t1 und t2 gleichgut begründet sind, dann handelt es jedenfalls *epistemisch irrational* wenn es die ein Theorie der anderen vorzieht.

Außerdem erweisen sich die Konsequenzen des (EG) wie etwa der statistische Syllogismus als ebenfalls sinnvolle Prinzipien, auf die wir immer wieder angewiesen sind und die wir im Normalfall ohne größere Bedenken einsetzen. Das sollten wir dann genauso mit dem klassischen Indifferenzprinzip halten. Soweit sieht die Sache eindeutig aus. Gäbe es nicht bestimmte Probleme mit den Prinzipien, könnten sie vermutlich die Mehrheit der Wissenschaftler und Philosophen auf ihre Seite bringen. Doch was können wir zu den Problemen sagen?

Einwände gegen (EG). Ein erster Kritikpunkt der Skeptiker besagt: Wie kann man aus *Nichtwissen* nun mit (IP) *Wissen* machen? Nämlich das Wissen um bestimmte Wahrscheinlichkeiten für die einzelnen Möglichkeiten. Doch das verkennt den obigen Schluss. Er soll nur unser Unwissen in bestimmte Zahlen übersetzen. Da unser Unwissen symmetrisch bzgl. der Möglichkeiten ist, besagt das Prinzip nur, dass unsere Glaubensgrade das ebenso sein sollten. Mehr nicht. Wir sagen

nichts über die objektiven Wahrscheinlichkeiten. Das wäre zu viel verlangt. Es ist sogar erlaubt, zu sagen, wir vergeben in bestimmten Fällen keine Wahrscheinlichkeiten, weil wir zumindest für diese Fälle das nicht für sinnvoll halten. (EG) besagt nur, dass wir gleiche Glaubensgrade vergeben müssen, wenn wir überhaupt welche vergeben.

Trotzdem scheint das ein guter Ausgangspunkt gerade für bayesianische Überlegungen zu sein. Wir gelangen so zu Startwahrscheinlichkeiten, die nicht so extrem sind, dass eine Veränderung durch reale Daten nicht mehr zu erwarten ist. Das kann der subjektive Bayesianer sonst nicht ausschließen. Auf dieses Problem komme ich später noch zurück. Die induktiven Logiker haben das (IP) sehr stark gemacht und als Ausgangspunkt all ihrer Überlegungen hergenommen. Wenn unser bisheriges Hintergrundwissen neutral bzgl. verschiedener Möglichkeiten ist, so sollte sich diese Neutralität genauso in unseren Glaubensgraden widerspiegeln. Jede Abweichung von der Neutralität wäre demnach begründungspflichtig.

Leider lässt sich diese Neutralität nicht immer so einfach aufrechterhalten. Zunächst bestimmt sich die Anzahl der Optionen anhand unserer sprachlichen Darstellung dieser Möglichkeiten. Schon dabei können Probleme auftreten. Wie das passiert, mag ein einfaches Beispiel erläutern. Nehmen wir an, wir haben weiße, blaue und rote Kugeln in einem unbekannten Mischungsverhältnis in einer Urne. Wie groß ist in diesem Fall die Wahrscheinlichkeit, eine weiße Kugel zu ziehen? Nach (IP) müssten wir sagen: 1/3. Aber wir hätten die Situation auch anders beschreiben können. Vielleicht hätten wir sogar nur die Farbprädikate *weiß* und *bunt* zur Verfügung gehabt. Dann hätten wir gesagt: Es sind weiße und bunte Kugeln in der Urne, aber wir kennen das Mischungsverhältnis nicht. Dann hätte man mit (IP) sagen müssen, dass die Wahrscheinlichkeit für eine weiße Kugel 1/2 gewesen wäre. Die Anwendung führt also je nach Beschreibung oder Sprache zu unterschiedlichen Ausgangswahrscheinlichkeiten.

Nach Williamson (2010) ist das aber auch zu erwarten, denn jede solche Beschreibung enthält bestimmte Informationen über die Situation und unterschiedliche Beschreibungen können unterschiedliche Informationen enthalten, die dann auch zu anderen Wahrscheinlichkeitseinschätzungen führen sollten. Ob das in unserem Fall den Unter-

schied erklärt, ist sicher nicht so leicht zu entscheiden. Aber die Idee ist sicher überlegenswert, dass Beschreibungen keineswegs harmlos und neutral sind, sondern bereits erste Informationen enthalten. Haben wir keine weiteren Informationen können die Beschreibungen vielleicht schon den Tag entscheiden.

Auch nach Franklin (2001) müssen wir wegen der Problemfälle nicht gleich verzweifeln, sondern sollten weiter an solchen Beispielen arbeiten. Wir können in diesem Fall nach den grundlegenderen Beschreibungen Ausschau halten und die Anwendung von (IP) auf diese beschränken (hier also auf den ersten Fall) oder wir müssen uns darauf einlassen, verschiedene Darstellungen als gleichwertig zu betrachten und etwa eine Mischung daraus zulassen oder sogar mit einer Menge von zulässigen Wahrscheinlichkeiten weiterzuarbeiten. Oder manchmal müssen wir sogar zugeben, dass (IP) hier einfach nicht anwendbar ist. Niemand kann uns zu der verrückten Behauptung zwingen, (IP) müsste auf jedes Beispiel anwendbar sein und würde immer zu sinnvollen Ergebnissen führen. Sind wir davon erst einmal abgerückt, können wir nach weiteren Anforderungen an (IP) suchen, die zu sinnvollen Einschätzungen führen sollen. Insbesondere sollten wir solche Fälle ausschließen, in denen gleichermaßen grundlegende Beschreibungen zu unterschiedlichen Einschätzungen durch (IP) führen. Überhaupt hängt alles von einer sorgfältigen Beschreibung der jeweiligen Situationen ab. Die genannten Prinzipien werden oft zu leichtfertig ins Spiel gebracht.

Das könnte uns z.B. für wissenschaftliche Hypothesen oder ähnliche Fälle passieren. So könnte jemand fragen, ob wir eine neue quantenmechanische Theorie Q mit ganz neuen Gesetzen für wahr oder falsch halten. Da wir bisher noch nicht wissen, wie erfolgreich die Theorie Q angewandt werden kann, sollten wir seiner Meinung nach (IP) einsetzen und müssten dann sagen: Wir haben zwei Möglichkeiten (wahr oder falsch). Also ist die Ausgangswahrscheinlichkeit für die Theorie 1/2. Doch das erscheint uns nicht besonders plausibel und ist vermutlich zu hoch angesetzt.

Für wissenschaftliche Theorien sollte die Anwendung eher wie folgt aussehen: Wenn wir für einen bestimmten Bereich n Theorien t1, …, tn finden, die einander ausschließen, von denen wir aber annehmen, dass

eine wahr ist und außerdem gilt: t1 ≈ … ≈ tn, dann sollten wir P(t1) = … = P(tn) = 1/n ansetzen.

Wir könnten sonst auch ganz anders rechnen: Die Theorie Q stellt Behauptungen über die potentiell unendlich große Zahl an beobachtbaren Sachverhalten b1, b2,… auf und behauptet, dass genau diese auftreten und nicht ihre Negationen. Dann könnten wir die komplexen zugrundeliegenden Sachverhalte etwa als unendliche Vollkonjunktionen ±b1 & ±b2 &… beschreiben. Im Idealfall sagt uns die Theorie Q also, dass genau einer dieser komplexen Sachverhalte (nämlich b1 & b2 & …) realisiert ist. Doch jeder dieser Sachverhalte kann zunächst falsch oder richtig sein. Also ist die Anzahl an Möglichkeiten durch die Menge aller Folgen ±b1, ±b2,… gegeben. Die ist aber sogar überabzählbar. Daher sollte bei einer Gleichverteilung jede Folge die Wahrscheinlichkeit 0 erhalten. Nur eine der Folgen besagt, dass unsere Theorie wahr ist, also sollten wir der Theorie die Ausgangswahrscheinlichkeit 0 geben, was bekanntlich für den Bayesianer fatal wäre, da sich die Theorie selbst durch wiederholtes Updaten davon nicht mehr erholen kann.

Ähnlich sah schon Poppers Überlegung aus. Der Probabilist sollte an dieser Stelle vermutlich dafür eintreten, dass wir immer nur Theorien über de facto endlich viele Objekte betrachten, wenn wir das Indifferenzprinzip zum Einsatz bringen möchten. Für abzählbar unendlich viele Alternativen gibt es keine Gleichverteilung mehr und für überabzählbar viele Optionen, die etwa parametrisiert sind, können wiederum spezielle Probleme auftreten, wenn wir die Parameter transformieren, wie sich gleich zeigen wird.

Schauen wir dazu auf ein Beispiel von van Fraassen (1989). Nehmen wir an, eine Fabrik stellt Holzwürfel mit einer *Kantenlänge* zwischen 0 und 1 Meter her. Dann liegen die *Flächen* der Würfel zwischen 0 und 1 Quadratmeter. Wir denken uns im Folgenden irgendwelche Einheiten hinzu. Gehen wir davon aus, wir wüssten nicht mehr über die Produktion und fragen uns nun, welche Ausgangswahrscheinlichkeit wir für bestimmte Würfelgrößen annehmen sollten. Fragen wir etwa, wie hoch die Wahrscheinlichkeit p ist, dass die Würfel eine Kantenlänge zwischen 0 und 0,5 haben werden. Zunächst bietet es sich an, auf dem Intervall [0;1] eine Gleichverteilung anzusetzen (eine konstante Dichtefunktion mit Wert 1 auf dem Intervall) und dann den beiden

Intervallen I1 = [0;0,5] und I2 = (0,5;1] dieselbe Wahrscheinlichkeit p=1/2 dafür zu geben, dass unser Würfel eine Kantenlänge aus diesem Intervall hat. Wir erhalten dann zunächst die Behauptung (P1), um damit das (IP) anwenden zu können:

(P1) »Der Würfel hat eine Kantenlänge in I1.« ≈
»Der Würfel hat eine Kantenlänge in I2.«

Ähnlich können wir aber auch für die Flächen vorgehen (und schließlich noch für die Volumina). Nehmen wir auf den Flächen wieder eine Gleichverteilung an, könnten wir in analoger Form für die Würfelflächen F ansetzen und erhalten so die Forderung (P2):

(P2) $0 \leq F \leq 0{,}25 \approx 0{,}25 < F \leq 0{,}5 \approx 0{,}5 < F \leq 0{,}75 \approx 0{,}75 < F \leq 1$

Wenn wir nun das (IP) einmal im Sinne von (P1) und einmal im Sinne von (P2) anwenden erhalten wir offensichtlich widersprüchliche Resultate, denn eine Kantenlänge bis 0,5 entspricht gerade der Fläche bis 0,25.

Nach Roger White (2010, 164 ff.) sind aber die beiden Annahmen (P1) und (P2) bereits an sich widersprüchlich, auch ohne dass wir dafür das Indifferenzprinzip einbringen müssten. Er begründet dies noch anhand weiterer Überlegungen. Das ist zumindest ein sehr guter Hinweis dafür, dass wir hier nicht leichtfertig das Indifferenzprinzip anwenden sollten, weil einfach die Voraussetzungen dafür nicht geklärt sind. Wir wissen nämlich nicht, welche Aussagen wir tatsächlich als gleich gut bestätigt ansehen sollen. Dafür widersprechen sich (P1) und (P2), denn die Aussage »Der Würfel hat eine Kantenlänge in I1« ist äquivalent zu der Aussage »Der Würfel hat eine Fläche zwischen 0 und 0,25 cm«. Diese Aussage wird aber einmal als gleichbestätigt zu einer Gegenaussage und einmal zu drei Gegenaussagen angesehen. Das kann nicht zugleich richtig sein. Also sind die Voraussetzungen für die Anwendung von (IP) nicht wirklich erfüllt. Wir müssten erst eine Beschreibungsebene finden, die invariant unter den entsprechenden Parametertransformationen ist, um das (IP) zum Einsatz bringen zu können. Das ist zumindest der Weg, den einige Philosophen beschritten haben (vgl. Childers 2013, Kap. 5+6). Wenn uns das aber nicht gelingt, ist das Indifferenzprinzip schlicht nicht anwendbar. Dabei zeigt sich wieder einmal, dass unsere

Beschreibungen der Situation bereits bestimmte Informationen über die Situation einschmuggeln. Novack (2010) gibt dazu noch eingehendere formale Analysen des Beispiels, die aber hier den Rahmen sprengen würden.

Schauen wir noch kurz auf das *Wasser-Wein-Paradox*, das schon bei von Mises (1957, 77) diskutiert wurde. Es wird meist als das hartnäckigste Paradox dieser Art angesehen. Darin geht es um eine Flüssigkeit zusammengemischt aus Wein und Wasser. Als einzige Information erfahren wir, dass von keiner der Flüssigkeiten mehr als das dreifache der anderen im Gemisch enthalten ist. Die Frage ist dann, wie hoch die Wahrscheinlichkeit ist, dass vom Wein nicht mehr als doppelt so viel wie Wasser in der Flüssigkeit ist: Was ist also P(Wein/Wasser ≤ 2)? Dazu wird der folgende Ansatz vorgeschlagen. Es ist klar, dass das Mischungsverhältnis Wein-zu-Wasser zwischen 1/3 und 3 liegt. Nehmen wir dann eine Gleichverteilung auf diesem Intervall von Verhältniszahlen an, dann ergibt sich die folgende Rechnung, wobei die Wahrscheinlichkeit einfach als Anteil am gesamten Intervall betrachtet wird:

$$\mathrm{P}(\text{Wein/Wasser} \leq 2) = \left(2 - \tfrac{1}{3}\right) / \left(3 - \tfrac{1}{3}\right) = \tfrac{5}{8}$$

Das basiert allerdings darauf, dass wir alle Mischungsverhältnisse, die erlaubt sind, auf diese Weise sinnvoll für den Einsatz im (IP) parametrisiert haben. Daran kommen Zweifel auf, sobald wir die umgekehrten Mischungsverhältnisse betrachten (auch hier gilt wieder: $1/3 \leq$ Wasser / Wein ≤ 3) und dann für die entsprechende Aussage die Wahrscheinlichkeit nach demselben Verfahren berechnen:

$$\mathrm{P}(\text{Wasser/Wein} \geq \tfrac{1}{2}) = \left(3 - \tfrac{1}{2}\right) / \left(3 - \tfrac{1}{3}\right) = \tfrac{15}{16}$$

Hier gibt es wiederum die Möglichkeit nach einer sinnvollen invarianten Beschreibung zu suchen, oder wir müssen eben davon Abstand nehmen, das Indifferenzprinzip anzuwenden. Mikkelson (2004) schlägt den ersten Weg ein. Er stellt die hinzugefügten Flüssigkeiten über ihre Mengenangaben in einem vorgegeben Zylinder dar und gewinnt so eine invariante Darstellung gegenüber der Wasser—Wein zu Wein—Wasser Transformation, die uns die Schwierigkeiten eingebracht hat. Dann lässt sich das (PI) eindeutig anwenden. Ob diese Lösung nun Bestand haben

wird, muss die weitere Debatte erweisen. In jedem Fall müssen wir die Voraussetzungen des (IP) gründlich prüfen, wonach zwei Aussagen definitiv gleich gut begründet sein müssen. Wenn das nicht der Fall ist, können wir das (IP) nicht anwenden. Solange wir das beachten, haben wir eine naheliegende Antwort auf die Paradoxien.

Außerdem sollten wir immer wieder den Hinweis von Jon Williamson beachten, dass in jeder Beschreibung einer Situation bereits bestimmte *Informationen über die Situation* zu finden sind. Eine Beschreibung kann etwa nahelegen oder sogar implizieren, dass wir bestimmte Möglichkeiten als gleich gut begründet ansehen. Dann können unterschiedliche Beschreibungen zu unterschiedlichen Einschätzungen anhand des Indifferenzprinzips führen. Das ist auch zu erwarten und kein Fehler des Prinzips. Akzeptieren wir aber beide Beschreibungen dann als doch gleichwertig, müssen wir zugeben, dass das Indifferenzprinzip nicht mehr anwendbar ist.

Mein Fazit ist daher: Das Prinzip der epistemischen Gleichbehandlung ist ein sehr plausibles Prinzip der epistemischen Rationalität. Um es korrekt anwenden zu können, darf es aber keine inkohärenten Beschreibungen einer Situation geben, die jeweils unterschiedliche Aussagen als gleich gut begründet auszeichnen. Sind die aber in unserem Hintergrundwissen nicht vorhanden, sollten wir dem (EG) folgen und erhalten dann auch das Indifferenzprinzip und den statistischen Syllogismus, auf den wir angewiesen sind, damit wir aus unseren Daten überhaupt sinnvolle Schlussfolgerungen ziehen können.

Das MaxEnt-Prinzip. Es gibt auch noch modernere und allgemeinere Varianten des Indifferenzprinzips, die u.a. von Jon Williamson als Grundlage unseres epistemischen Schließens angesehen werden. Da ist vor allem das Prinzip der *maximalen Entropie* zu nennen. Wenn wir z.B. bestimmte Nebenbedingungen haben, wie etwa die, dass eine Wahrscheinlichkeit mindestens doppelt so groß wie eine andere ist (es sind natürlich auch ganz andere Arten von Vorgaben für die Wahrscheinlichkeiten möglich), dann möchte man z.B. wissen, welche der dann noch zulässigen Verteilungen diejenige ist, die einer Gleichverteilung am nächsten kommt. Nehmen wir also an, wir haben nur zwei Möglichkeiten mit zwei Wahrscheinlichkeiten p_1 und p_2 und es gilt $p_1 \geq 2 \cdot p_1$, dann sind noch viele Verteilungen zulässig. Doch $p_1 = 2/3$ und $p_2 = 1/3$ ist dabei

sicherlich die gleichmäßigste Verteilung, die noch möglich ist. Heißt unsere Nebenbedingung allerdings $p_1 > 2 \cdot p_1$, dann finden wir schon keine solche Lösung mehr, sondern nur noch Annäherungen daran bzw. Grenzwahrscheinlichkeiten zu dem Problem. Haben wir n Möglichkeiten und müssen darauf n Wahrscheinlichkeiten festlegen, die aber einen Nebenbedingung N erfüllen müssen, so sollten wir die Entropie unserer Verteilung im Bereich der Nebenbedingung N maximieren:

Entropie: $H(\mathbf{p}) = H(p_1,\ldots,p_n) = -\sum_i p_i \cdot \log(p_i)$

Da der Logarithmus der Wahrscheinlichkeiten negativ ist, wird die rechte Seite mit einem Minuszeichen versehen, um das folgende Prinzip als Form der Maximierung beschreiben zu können. H(p) nimmt bei einer Gleichverteilung sein Maximum an. Einige der Eigenschaften der Informationsentropie finden sich etwa bei MacKay (2005, 32 ff., 308) aufgelistet. Sind weitere Nebenbedingungen N zu erfüllen, sollte nach dem Prinzip der maximalen Entropie nach der Verteilung gesucht werden, die N erfüllt und für die dabei H(p) maximal wird.

Prinzip der maximalen Entropie (MaxEnt): Wähle die Verteilung **p** aus einer Menge von Verteilungen aus, die die Nebenbedingungen N erfüllen, für die H(**p**) maximal wird.

Das ist in diesem Fall die Verteilung mit der niedrigsten Information in der Wahrscheinlichkeitsfunktion, die nach N noch zulässig ist. So erhalten wir eine Verallgemeinerung von (IP), die z.B. der Physiker E.T. Jaynes (2003) oder der Wissenschaftstheoretiker Jon Williamson (2010) (vgl. a. Landes & Williamson 2013) als zentralen Startpunkt weiterer Überlegungen der Bewertung wissenschaftlicher Theorien betrachten. Williamson plädiert etwa dafür, unsere Daten immer als Nebenbedingungen zu betrachten und dann die Glaubensgrade für alle anderen Annahmen und Theorien zu vergeben, die durch MaxEnt bestimmt werden. So sollte dann auch mit neuen Daten upgedatet werden. Wir erhalten eine neue Nebenbedingung und wenden danach MaxEnt an. Das fällt in den einfachen Fällen des Updatens (in denen wir einfach eine bestimmte Beobachtungsüberzeugungen E akzeptieren) auch mit dem Updaten anhand der bayesianischen Konditionalisierungsregel

zusammen, wenn wir mit einer Gleichverteilung starten. Das entspricht der Idee, nur soweit von der Gleichverteilung abzuweichen, wie die Daten es erzwingen. Ein weiteres Abrücken von der Gleichverteilung zugunsten einer bestimmten Theorie würde dann schon wieder die Unvoreingenommenheitsforderung verletzen.

5.3.16 Bedingte Wahrscheinlichkeiten als basal

Wir haben konditionale Wahrscheinlichkeiten bisher immer durch einfache Wahrscheinlichkeiten definiert:

Bedingte Wahrscheinlichkeit: P(A|B) := P(A&B) / P(B)

Diese Definition setzt aber voraus, dass P(B) > 0 gilt. Doch selbst in den Fällen, in denen P(B) = 0 ist, können wir oft noch sinnvoll von bedingten Wahrscheinlichkeiten sprechen. Wenn z.B. das Ereignis E, dass eine gerade Zahl gewürfelt wird, aus irgendwelchen Gründen im Moment die Wahrscheinlichkeit 0 hat (z.B. weil die Würfel bereits gefallen sind), dann können wir doch trotzdem die Frage stellen, welche Wahrscheinlichkeit die 4 gehabt hätte, wenn eine gerade Zahl gefallen wäre. Hajek (2003) bietet andere Beispiele, in denen wir gezwungenermaßen eine Wahrscheinlichkeit von 0 für B erhalten. Was ist z.B. die Wahrscheinlichkeit dafür, dass ein zufällig herausgegriffener Punkt auf der Erde in der westlichen Hemisphäre liegt, wenn wir schon wissen, dass er auf dem Äquator liegt? Naheliegende Antwort 1/2, obwohl die Wahrscheinlichkeit für das Liegen auf dem Äquator selbst wiederum 0 ist, da es sich um eine sogenannte Nullmenge innerhalb der Kugeloberfläche handelt. Um diese Fälle korrekt beschreiben zu können, scheint es sinnvoll zu sein, unsere ganze Betrachtung der Wahrscheinlichkeitsanwendungen mit bedingten Wahrscheinlichkeiten zu starten und diese dann als basal zu betrachten.

Außerdem können wir in praktischen Anwendungen versuchen, die Anzahl der zu schätzenden Wahrscheinlichkeiten deutlich zu reduzieren, indem wir uns auf wenige relevante Wahrscheinlichkeiten beschränken, die wir tatsächlich für Berechnungen zu einem bestimmten Problem benötigen. So wissen wir etwa, dass die bedingte Wahrscheinlichkeit dafür, dass es in Moskau nass ist, gegeben, es hat dort geregnet, sehr hoch ist, selbst wenn wir uns mit den klimatischen Bedingungen in Moskau

nicht wirklich auskennen (vgl. Hajek 2003). Das bedeutet etwa, dass es uns nicht gelingen mag, auf vernünftige Weise zu schätzen, wie hoch die Wahrscheinlichkeit dafür ist, dass es in Moskau regnet und es dort nass ist. Dann können wir unsere konditionale Analyse der bedingten Wahrscheinlichkeit nicht wirklich einsetzen, und es erscheint in derartigen Fällen viel naheliegender, die bedingten Wahrscheinlichkeiten als die grundlegenden zu betrachten. Diese lassen sich auf einfache Weise axiomatisieren:

Bedingte Wahrscheinlichkeiten als basal
Es gilt für alle A, B und C aus unserer Sprache L:
(A1) P(·|A) ist eine Wahrscheinlichkeitsfunktion im ersten Argument (für alle A)
(A2) P(A|A) = 1
(A3) P(B&C|A) = P(B|A&C)·P(C|A)

Es wurden darüber hinaus einige andere Axiomensysteme vorgeschlagen, aber dieses hat sich schon fast als recht einfacher Standard etabliert (vgl. Earman 1992, Appendix 1). Popper war einer der Ersten, die sich in den Anhängen (etwa *IV und *V) zur *Logik der Forschung* ausführlicher mit derartigen Systemen auseinandergesetzt haben.

Man erhält natürlich eine unbedingte Wahrscheinlichkeitsfunktion wieder zurück, indem wir als Bedingung eine feste Aussage üblicherweise eine Tautologie t wählen:

Unbedingte Wahrscheinlichkeit: P*(A) = P(A|t)

Außerdem gilt dann natürlich:

P(A|B) = P*(A&B) / P*(B)

So erhalten wir als Basis nun die konditionalen Wahrscheinlichkeiten, die in bestimmten Kontexten spezielle inhaltliche Interpretationen erfahren können, wie etwa im Rahmen einer induktiven Logik. Dort steht P(A|B) dafür, wie die Aussage A durch die Aussage B induktiv gestützt wird, und ist ein Maß der Bestätigung von A durch B. Auch wenn Bayesianer ihre Relativierung der Plausibilitätseinschätzungen auf ein bestimmtes

Hintergrundwissen explizit machen möchten, haben sie es eigentlich immer mit bedingten Wahrscheinlichkeiten zu tun und könnten die sehr wohl als grundlegend betrachten. Das scheint daher eine durchaus interessante Alternative zur klassischen Vorgehensweise zu sein.

5.4 Objektive Wahrscheinlichkeiten und Propensitäten

Neben den subjektiven oder epistemischen Wahrscheinlichkeiten, die Auskunft darüber geben, wie sicher wir uns bestimmter Aussagen bereits sind oder wie plausibel bestimmte Aussagen im Lichte unseres Hintergrundwissens sind, wird man üblicherweise auch eine Konzeption *objektiver Wahrscheinlichkeit* einsetzen. Im Hauptprinzip hatten wir schon gefordert, dass sich die epistemischen Wahrscheinlichkeiten an den objektiven Chancen zu orientieren haben. Außerdem kommen objektive Wahrscheinlichkeitsaussagen in vielen wissenschaftlichen Theorien vor. Sie behaupten, dass bestimmte Ereignisse mit einer von uns und unserem Kenntnisstand unabhängigen Tendenz auftreten. Diese Tendenz ist allerdings meist erst in längeren Folgen von gleichartigen Ereignissen als *relative Häufigkeit* erkennbar, mit der ein bestimmtes Resultat auftritt. Wenn ein bestimmtes Atom A eine gewisse Tendenz etwa der Stärke 0,7 besitzt, in einem bestimmten Zeitraum T zu zerfallen und wir beobachten viele zu A gleichartige Atome über den Zeitraum T, dann werden wir feststellen, dass ca. 70% diese Atome zerfallen.

Wichtig ist immer, sich vor Augen zu führen, dass es uns hierbei nicht um die formale Theorie der Wahrscheinlichkeit geht, sondern um die *Anwendungen des Wahrscheinlichkeitsbegriffs* auf die Wirklichkeit, in denen wir versuchen, bestimmte Aspekte der Wirklichkeit durch Wahrscheinlichkeiten zu modellieren. Man spricht auch von *Interpretationen der Wahrscheinlichkeit*, wenn wir thematisieren, wie wir solche empirischen Wahrscheinlichkeitsaussagen verstehen können. In derartigen Aussagen beziehen wir uns mit »Wahrscheinlichkeit« nicht nur auf das formale Konzept, sondern möchten damit reale physikalische Eigenschaften bezeichnen. Doch wie können wir solche Aussagen verstehen? Das ist natürlich wesentlich auch für das induktive Schließen im Hinblick auf solche probabilistischen Behauptungen bzw. Theorien.

5.4.1 Objektive physikalische Wahrscheinlichkeit

Statt von Atomen, die zerfallen, oder anderen komplexen Systemen werde ich allerdings meist nur von geworfenen Münzen als Zufallsexperiment sprechen, die in einer bestimmten Situation (oder auch einem Typ von Situation) geworfen werden und dabei Kopf oder Zahl ergeben. Das soll mein Beispiel für ein indeterministisches System sein, weil das etwas einfacher und übersichtlicher ist als etwa quantenmechanische Beispiele. Wir werden später noch darüber diskutieren, ob es überhaupt genuin indeterministische Systeme gibt. Zunächst gehe ich hier einfach davon aus und wähle oft das Münzbeispiel dafür. Wem das nicht plausibel erscheint, der muss dieses eben durch eigene Beispiele etwa aus dem quantenmechanischen Bereich ersetzen. Hat meine Münze eine Wahrscheinlichkeit von 0,6 dafür, Kopf zu ergeben (es ist also keine faire Münze), so erwarte ich, dass bei 100 Würfen ca. 60-mal Kopf auftreten wird. Das ist eine bekannte erste Erläuterung eines objektiven physikalischen Wahrscheinlichkeitsbegriffs. Doch was behauptet man hier genau, wenn man einem solchen System eine objektive Wahrscheinlichkeit von 0,6 für ein bestimmtes Ergebnis zuschreibt? Untersuchen wir also die Behauptung:

(*) Die objektive physikalische Wahrscheinlichkeit für Kopf ist 0,6

Das mathematische Konzept der Wahrscheinlichkeit wird durch die genannten einfachen Axiome beschrieben, aber damit ist noch nichts darüber gesagt, welche empirische Behauptung wir mit der Aussage (*) aufstellen. Das zu klären ist Aufgabe einer *Interpretation* von objektiver Wahrscheinlichkeit.

Wir versuchen typischerweise mit Hilfe des Konzepts physikalischer Wahrscheinlichkeit Phänomene zu beschreiben und sie womöglich sogar zu erklären, bei denen wir *lokal ein regelloses (zufälliges) Verhalten* vorfinden, für die sich aber bei längeren Versuchsreihen *stabile relative Häufigkeiten* einstellen. Zumindest für einige derartige Prozesse etwa im Rahmen der Quantenmechanik haben wir zusätzliche gute Gründe zu der Annahme, dass sie genuin indeterministisch sind. Die Frage, ob ein Atom A in einem vorgegebenen Zeitraum T zerfällt, deutet demnach nicht nur auf eine Wissenslücke im Hinblick auf die Ursachen

des Zerfalls hin, sondern dieser Vorgang ist durch alle seine Ursachen nicht determiniert. Dafür sprechen zumindest diejenigen Überlegungen aus der Quantenmechanik, die Gründe gegen die Existenz verborgener Parameter darstellen.

Wir finden also z.B. gewisse Situationen, in denen es durchaus sein kann, dass ein Atom in einem bestimmten Zeitraum eine *starke* Zerfallstendenz hat; dann sollten wir eher darauf wetten, dass es in T zerfällt, als dass es durchhält. Die solchen Vorgängen zugrunde liegende dispositionale Eigenschaft hat Popper als *Propensität* bezeichnet. Das ist in unserem Beispiel eine *Disposition zu zerfallen* von einer bestimmten Stärke, die gerade durch unsere physikalische Wahrscheinlichkeit als Maß für das Zerfallen gekennzeichnet wird. Darauf kommen wir gleich wieder zurück. Jedenfalls dürfen wir solange davon ausgehen, dass es genuin indeterministische Vorgänge gibt, wie wir die Quantenmechanik akzeptieren und verborgene Parameter weiterhin ablehnen.

In anderen Bereichen ist es dagegen eher unklar, ob wir dort genuin indeterministische Prozesse vorfinden, oder ob sie uns nur indeterministisch erscheinen, weil wir ihre Ausgangsbedingungen nicht genau genug kennen bzw. nicht einmal genau genug bestimmen *können*, oder ob die beteiligten Vorgänge einfach so komplex sind, dass wir sie nur noch als indeterministische Vorgänge beschreiben können (vgl. Strevens 2003). Wir sollten zumindest darauf vorbereitet sein, auch objektive Wahrscheinlichkeitsaussagen richtig verstehen zu können, und es ist nicht ausgeschlossen, dass wir sie ebenso in anderen Bereichen außerhalb der Quantenmechanik benötigen, zumal Quanteneffekte durchaus makroskopische Auswirkungen haben können, wie jeder Messvorgang demonstriert. Wenn wir im Rahmen der klassischen Mechanik bestimmte Idealisierungen wie die der Kontinuität des Raumes aufgeben, kann es selbst dort indeterministische Systeme geben (vgl. Bartelborth 1994).

5.4.2 Wahrscheinlichkeit und relative Häufigkeiten

Die naheliegendste und prominenteste Idee zur Explikation von objektiver physikalischer Wahrscheinlichkeit ist natürlich, die genannte Verbindung zur relativen Häufigkeit auszunutzen. Wie groß ist danach die Wahrscheinlichkeit für das Auftreten einer Eigenschaft A (also z.B. »Kopf

oben«) für ein bestimmtes Zufallsexperiment (z.B. einen Münzwurf)? Sie ist gleich der relativen Häufigkeit $h_n(A) = r/n$ mit der A in einer langen Reihe n gleichartiger Zufallsexperimente (vom Typ S) auftritt, wenn r die Anzahl der As in der Reihe darstellt:

1. Reduktionsversuch (endliche relative Häufigkeit)
$P(A) = h_n(A)$ für eine große natürliche Zahl n

Dieser erste Reduktionsversuch für Wahrscheinlichkeitsaussagen scheint sogar strikten empiristischen Ansprüchen zu genügen, denn im Definiens steht eine im Prinzip leicht beobachtbare Größe. Allerdings werden sogleich drei erste Schwierigkeiten sichtbar: 1. Wie groß muss denn n sein, damit wir $h_n(A)$ als Wahrscheinlichkeit deuten dürfen? Sehr kleine Zahlen wie n=1 genügen sicher nicht, denn bei n=1 gäbe es nur die trivialen Wahrscheinlichkeiten 0 und 1. Die Anzahl der Versuche n sollte also schon recht groß sein, aber eine genauere Zahl können wir nicht nennen. 2. Außerdem finden wir nur rationale Zahlen (also Brüche) als Wahrscheinlichkeiten, obwohl wir in unserem mathematischen Modell sonst typischerweise reelle Zahlen erlauben (und das sind sehr viel mehr Zahlen). Die reellen Zahlen können wir jedoch zunächst einfach als mathematische Idealisierung ansehen, während die tatsächlichen empirischen Werte eben nur rational sind. 3. Wir stoßen bereits hier auf ein virulentes Problem der Wahrscheinlichkeitseinschätzungen, nämlich die Frage, wann wir es mit *gleichartigen Zufallsexperimenten* bzw. dem *Wiederholen eines Experiments* zu tun haben. Das wird uns noch als sogenanntes Problem der *Referenzklasse* verfolgen. Jedenfalls vergeben wir diese Wahrscheinlichkeiten nicht für ganz heterogene Versuchsreihen, sondern erwarten intuitiv vielmehr, dass A in diesen Experimenten immer mit derselben Wahrscheinlichkeit auftritt und diese sich dann bei größeren Zahlen entsprechend manifestiert. Beliebigen Reihen von Zufallsexperimenten ihre relativen Häufigkeiten, mit der ein bestimmtes Resultat auftritt, als Wahrscheinlichkeit zuzuordnen, ergibt dagegen nur wenig Sinn. Darauf kommen wir wieder zurück. Die gerade formulierte intuitive Auszeichnung der gleichartigen Zufallsexperimente würde allerdings zirkulär wirken und wäre so nicht einsetzbar. Daher ist das Referenzklassenproblem noch nicht gelöst, sondern nur beschrieben worden.

Neben diesen Unklarheiten der Definition 1 gibt es leider noch eine ganze Reihe von grundsätzlicheren Problemen, die dieser Definition im Wege stehen. Das erste nennt Rosenthal (2004) das »*Lückenproblem*«. Es bleibt eine Lücke zwischen Definiendum und Definiens. Nehmen wir an, wir werfen 1000-mal unsere Münze und erhalten 496-mal Kopf. Nähmen wir unsere Definition ernst, müssten wir nun behaupten, dass die Münze *mit Sicherheit* gefälscht sei, denn die Wahrscheinlichkeit P(Kopf) = 496/1000 ist nicht genau 1/2. Doch tatsächlich gehört zu Zufallsprozessen immer ein gewisser Spielraum und immer eine Schwankung der Ergebnisse. Unser Resultat passt daher durchaus gut zu einer fairen Münze. Wir erwarten überhaupt nicht, dass bei einer fairen Münze (d.h. P(Kopf) = 1/2) immer genau in der Hälfte der Fälle Kopf kommt. Für ungerades n ist das sogar unmöglich. Aufgrund der Lücke zwischen Definiens und Definiendum kann es sich in unserem ersten Versuch nicht um eine echte Definition handeln. Wir werden auf solche relativen Häufigkeiten erst zurückgreifen dürfen, wenn es uns nur noch um eine (approximative) Messung der Wahrscheinlichkeit geht und nicht mehr um eine definitorische Reduktion.

Endliche relative Häufigkeiten können also durchaus von der zugrundeliegenden Wahrscheinlichkeit abweichen. Das gehört sogar zwingend zu unserem Konzept von Wahrscheinlichkeit dazu. Größere Abweichungen haben allerdings kleinere Wahrscheinlichkeiten aufzutreten, doch das hilft uns für eine Definition von Wahrscheinlichkeit zunächst nicht weiter. Es wird aber natürlich ein wichtiger Gedanke für die Messung von objektiven Wahrscheinlichkeiten sein.

Ein zweites Problem ist damit verwandt. Die relativen Häufigkeiten schwanken stark, wenn wir weiter werfen. Wir suchen andererseits nach einer festen Wahrscheinlichkeit für unseren Münzwürfe und nicht nach einer Zahl, die auf einen bestimmten Zeitpunkt oder eine konkrete Versuchsreihe zu relativieren wäre. Wahrscheinlichkeit würde ein höchst instabiles Gut, das je nach Anzahl meiner Versuche hin und her schwankt. Die relativen Häufigkeiten stellen daher keine geeignete Explikation von objektiver Wahrscheinlichkeit dar.

Drittens bleibt immer die Frage offen, ob wir auf diese Weise auch Wahrscheinlichkeiten für Einzelfälle vergeben können. Wir sagen doch: Beim nächsten Wurf mit diesem Würfel ist die Wahrscheinlichkeit 1/6,

dass wir eine 5 erzielen. Dann suchen wir nach einer Einzelfallwahrscheinlichkeit, doch wie bestimmt man die? Etwa so: Wir suchen nach einer *Referenzklasse* von Würfen zu dem nächsten Wurf, deren relative Häufigkeit des Auftretens der 5 dann unsere Wahrscheinlichkeit bestimmt, dass beim nächsten Wurf eine 5 fällt. Besteht die Referenzklasse aber nun aus allen Würfen mit demselben Würfel oder auch solchen mit ähnlichen Würfeln? Nehmen wir nur die Würfe bis heute oder auch nur einen Teil davon? (Etwa die, bei denen nicht so ein starker Seitenwind geweht hat.) Müssen wir die zukünftigen Würfe ebenfalls mit berücksichtigen? Wie soll das jedoch gehen? Tue ich das nicht, schwankt die Wahrscheinlichkeit aber wiederum stark. Und was ist mit dem Fall, in dem wir einen einzigartigen Würfel haben und ihn nach einem Versuch gleich wieder zerstören? Hatte der überhaupt keine Wahrscheinlichkeit dafür, dass sich eine 5 ergibt, oder war es einfach nur nicht möglich, über die relative Häufigkeit die tatsächliche physikalische Wahrscheinlichkeit dieses einen Wurfs zu bestimmen?

Um eine Lösung des Referenzklassenproblems kommen wir eigentlich nie herum in der Anwendung der Wahrscheinlichkeitsrechnung. Doch wie könnte die aussehen? Jedes Zufallsereignis können wir als ein Element unterschiedlicher Versuchsreihen betrachten. Welche davon bestimmt die Wahrscheinlichkeit für unser spezielles Zufallsereignis, d.h., welches ist die *passende* Referenzklasse für das Ereignis? Oder man könnte das auch so formulieren: Wann dürfen wir davon sprechen, dass es sich bei bestimmten Ereignissen um (gleichartige) Wiederholungen unseres ursprünglichen Ereignisses handelt? Wir werden sehen, dass der Vertreter der Propensitäten dieses Problem vermeidet, indem er von der Tendenz im Einzelfall spricht, dass ein bestimmtes Ergebnis auftritt. Deren Stärke sei unsere objektive Wahrscheinlichkeit. Erst wenn wir diese Größe bestimmen wollen (also beim Messverfahren), muss auch der Propensitätenvertreter sich dem Referenzklassenproblem zuwenden.

Um zumindest die Relativität auf eine bestimmte endliche Folge von Würfen aufzuheben, hören wir von Mathematikern meist: Wahrscheinlichkeit ist die relative Häufigkeit *auf lange Sicht*. Und mit »langer Sicht« meinen sie dann oft eine unendliche Folge. Die Wahrscheinlichkeit für Kopf wäre also die relative Häufigkeit von Kopf, die sich ergeben würde,

wenn wir unendlich oft unsere Münze in gleichartigen Versuchen werfen würden. Dazu gehört demnach die Definition:

2. Reduktionsversuch (unendliche relative Häufigkeit)
$P(A) = \lim_{n\to\infty} h_n(A)$

Dieser Ansatz setzt wiederum eine ganze Menge voraus. Er nimmt an, es gäbe unendlich viele irgendwie bestimmte, allerdings fiktive oder hypothetische Folgenglieder, die wir für unsere Definition heranziehen dürfen. Damit stoßen wir schon auf das erste Problem dieses Ansatzes. Obwohl dieser Vorschlag aus dem empiristischen Lager kam, kann ein überzeugter Empirist damit nicht wirklich glücklich sein. Die fiktiven Folgenglieder können wir nicht beobachten und die Mathematik sagt uns zusätzlich, dass wir in einer unendlichen Folge jede endliche Teilfolge durch eine andere ersetzen könnten und dabei trotzdem denselben Grenzwert behalten. Das heißt, bei der zweiten Definition spielen die endlich vielen Resultate unserer Zufallsexperimente, die wir tatsächlich beobachten können, für den Grenzwert eigentlich keine Rolle. Sie geben uns aus rein mathematischer Sicht nicht den geringsten Hinweis auf den Grenzwert im Unendlichen. Jedenfalls dann nicht, wenn wir nicht spezielle weitere Annahmen über die endlichen Ergebnisse oder Einzelfälle in der Folge hinzunehmen, wie es der Propensitätenvertreter unternimmt. Für ihn manifestiert sich in jedem Wurf der Münze eine spezielle Tendenz zu Kopf und daher geben die ersten Folgenglieder bereits erste Informationen über diese Tendenz. In der rein formalen Definition der Empiristen, die ohne die Bezugnahme auf solche Propensitäten auskommen muss, haben endlich viele Folgenglieder aber überhaupt keine Aussagekraft mehr, denn es zählt nur der Grenzwert der relativen Häufigkeiten im Unendlichen. Wir müssen dann behaupten, dass P(A)=r besagt: Wenn wir das Experiment unendlich oft wiederholen würden, würde sich als Grenzwert der relativen Häufigkeit von A der Wert r einstellen. Alan Hajek (2009, 215) nennt solche kontrafaktischen Konditionale zu Recht »utterly bizarre«, denn die mögliche Welt, in der der Vordersatz wahr ist und wir etwa unendlich oft eine Münze werfen, hat nun wirklich keine großen Ähnlichkeiten mehr mit unserer tatsächlichen Welt. Wie sollen wir wissen, was in solchen entfernten Welten passieren würde? Diese Herangehensweise kann also einen

Empiristen nicht wirklich erfreuen, denn damit sagen uns alle tatsächlich beobachtbaren Ereignisse nichts mehr über die Wahrscheinlichkeit.

Außerdem setzt man voraus, dass die Folge der relativen Häufigkeiten tatsächlich konvergieren wird. Das ist keineswegs zwangsläufig der Fall. Um aus unseren endlich vielen Beobachtungen etwas über den Grenzwert sagen zu können, müssen wir unsere Definition so ergänzen, dass wir Grund zu der Annahme hätten, dass sich in den ersten Folgengliedern und frühen relativen Häufigkeiten bereits der Grenzwert zeigen würde. Das wird aber erst in unserem Propensitätenansatz (s.u.) der Fall sein. Der verlangt, dass die kausal relevanten Umstände des Versuchs in all den betrachteten Zufallsexperimenten immer gleich sein müssen. Das ist eine zusätzliche Annahme, die so etwas besagt wie: Die Wahrscheinlichkeit für das Auftreten von A sollte in allen Experimenten unserer Folge die gleiche sein. Das würde die Idee der kausalen Relevanz explizieren. Eine solche Zusatzbedingung ließe unsere Definition allerdings zirkulär werden.

Die Forderung nach Konvergenz hat der Mathematiker Richard Edler von Mises (1928) einfach seiner Konzeption von sogenannten *Kollektiven* hinzugefügt. Ein solches Kollektiv nach von Mises ist eine unendliche Folge von Ereignissen (E_i), bei der die relative Häufigkeit, mit der die Ereignisse die Eigenschaft A aufweisen, im Grenzwert konvergiert. Außerdem verlangt von Mises noch, dass für jede Teilfolge, die durch einen einfachen Algorithmus ausgewählt wird (der nicht auf die jeweiligen Resultate in der Folge Bezug nimmt), sich derselbe Wert als Grenzwert der relativen Häufigkeiten ergibt. Das soll sicherstellen, dass es sich bei den Kollektiven tatsächlich nur um *Zufallsfolgen* handelt und nicht etwa um solche Folgen wie: »Kopf, Zahl, Kopf, Zahl, Kopf,…«, die keinen wirklichen Zufallscharakter haben. Dieser Aspekt ist für uns allerdings nicht so wesentlich, so dass wir uns nicht weiter mit den Problemen dieses Zufallsaxioms beschäftigen müssen, sondern eher mit denen des Konvergenzaxioms.

Für von Mises war jedenfalls klar, dass der Begriff der Wahrscheinlichkeit immer auf ein ganz bestimmtes Kollektiv zu relativieren ist und Aussagen über Einzelfallwahrscheinlichkeiten demnach sinnlos sind. Dann sind jedoch ebenfalls alle Aussagen über die Wahrscheinlichkeit, mit der endliche Folgen auftreten, sinnlos, denn endliche Folgen sind

schließlich wiederum Einzelfälle etwas komplexerer Experimente. Für von Mises ist die Wahrscheinlichkeit hingegen immer nur eine Eigenschaft unendlicher Kollektive. Damit haben wir den Kontakt zu endlichen relativen Häufigkeiten verloren und erhalten genau genommen keine *empirische* Interpretation von Wahrscheinlichkeit mehr, sondern nur eine rein mathematische Ergänzung der mathematischen Theorie der Wahrscheinlichkeit. Über endliche Teilfolgen der Kollektive können wir keine weiteren Aussagen treffen. Sie auszutauschen ändert schließlich den Grenzwert der relativen Häufigkeiten nicht, und sie sind daher in der Konzeption von von Mises schlicht irrelevant.

In vielen Anwendungen der klassischen Statistik geht es um Wiederholungen eines bestimmten Experiments oder Experimenttyps. Dabei setzen wir meistens voraus, dass wir es in den einzelnen Wiederholungen mit voneinander unabhängigen und identisch verteilten Zufallsexperimenten (»independend and identically distributed« IID) zu tun haben. Erst die Wiederholungen von Ereignissen mit derselben Wahrscheinlichkeit geben uns Hinweise auf diese Wahrscheinlichkeit und nutzen diese Wahrscheinlichkeit, um bestimmte Vorhersagen zu erzeugen. Die Voraussetzung für diese Anwendungen der klassischen Statistik ist dann aber die Sinnhaftigkeit von Einzelfallwahrscheinlichkeiten, die der Frequentist gerade bestreitet. Er sollte sich daher dem Propensitätenansatz zuwenden.

Erst der Propensitätenvertreter schreibt jedem einzelnen Versuch eine Wahrscheinlichkeit zu und damit natürlich gleichfalls den endlichen Folgen gebildet aus solchen Versuchen. Die Konvergenz der relativen Häufigkeiten im Unendlichen (zumindest mit der Wahrscheinlichkeit 1) ergibt sich dabei von selbst (im Gesetz der großen Zahlen), spielt aber keine so bedeutsame Rolle mehr im Propensitätenansatz.

Für von Mises bleiben andererseits viele weitere Probleme übrig, von denen ich nur einige erwähnen möchte (für weitere vgl. Hajek 2010). Eine unangenehme Eigenschaft dieser unendlichen Folgen ist z.B., dass eine Umordnung der fiktiven Resultate uns jeweils zu ganz neuen relativen Häufigkeiten führen kann. Nehmen wir etwa die folgenden Ergebnisse jeweils durch 0 (Zahl) und 1 (Kopf) wiedergegeben:

Beispielfolge 1: a_1=1, a_2=0, a_3=1, a_4=0 usw.
Umordnungsfolge 2: a_1,a_3,a_2,a_5,a_7,a_9,a_4,...
also: 1,1,0,1,1,1,0,1,1,1,1,0,1,1,1,1,1,0....

Dabei wird in Folge 2 die Anzahl der vorgezogenen ungeraden Folgenglieder immer um 1 erhöht, bis wieder ein gerades Folgenglied aus der Folge 1 eingesetzt wird. Man sieht, dass die Grenzwerte sich unterscheiden: Folge 1 konvergiert gegen 1/2 (für die relativen Häufigkeiten) und Folge 2 gegen 1, obwohl jedes Element aus Folge 1 auch in Folge 2 auftritt. Auch wenn diese spezielle Folge gegen das Zufallsaxiom verstößt, erkennt man daran das Prinzip des Umordnens, das zu neuen Grenzwerten führen kann. Die *Reihenfolge* der fiktiven Folgenglieder spielt hier also eine wichtige Rolle. Doch wie soll man für fiktive Zufallsresultate eine Reihenfolge festlegen? Höchstens durch eine fiktive Zeit, zu der sie aufgetreten sind. Das ist jedenfalls alles andere als empiristisch akzeptabel.

Schlimmer ist aber wohl noch, dass selbst die *Lücke* zwischen Wahrscheinlichkeit und relativer Häufigkeit verbleibt. Das können wir uns am leichtesten klar machen, wenn wir die folgenden beiden Ergebnisfolgen für eine faire Münze betrachten, bei denen die 1 wieder für Kopf steht und die 0 für Zahl:

Folge A: 1,0,0,0,1,1,0,0,1,1
Folge B: 1,1,1,1,1,1,1,1,1,1

Welche Folge stellt das wahrscheinlichere Ergebnis dar? Beide sind gleichwahrscheinlich und haben jeweils die Wahrscheinlichkeit 2^{10} = 1/1024. Das Entsprechende gilt auch für längere Folgen dieser Art. Obwohl Folgen vom Typ A repräsentativer aussehen für eine faire Münze als die vom Typ B, hat jede konkrete Folge vom Typ B immer dieselbe Wahrscheinlichkeit wie eine gleichlange Folge vom Typ A. Damit steuern wir auf eine Seltsamkeit solcher unendlichen Mengen zu. Die Menge aller solchen unendlichen Folgen ist bereits überabzählbar. Wenn wir ihren Elementen noch eine Wahrscheinlichkeit zuordnen möchten und dabei gemäß dem Phänomen unserer zwei Folgen davon ausgehen, dass alle konkreten Folgen dieselbe Wahrscheinlichkeit erhalten sollen, dann bleibt nur die Wahrscheinlichkeit 0 für die einzelnen Folgen. Wir müssen

auf dem Raum dieser Folgen schließlich mit einer Wahrscheinlichkeitsdichte arbeiten, um auch komplexere Fragestellungen behandeln zu können.

Die angekündigte Seltsamkeit ist die Folgende: Wenn wir davon ausgehen, dass wir jedenfalls definitiv *eine* solche unendliche Folge würfeln, dann erhalten wir als Ergebnis eine unendliche Folge, die auf jeden Fall die Wahrscheinlichkeit 0 hat. Wir erhalten also *zwangsläufig* ein Ergebnis mit Wahrscheinlichkeit 0. Das zeigt, dass Wahrscheinlichkeit gleich 0 in diesen Fällen nicht bedeutet, dass ein Ereignis unmöglich ist.

Insbesondere können wir selbst mit einer *fairen* Münze noch die Folge 1,1,1,1,1,… werfen. Diese Folge ist genauso wahrscheinlich wie jede andere einzelne Folge. Sie hat zwar die Wahrscheinlichkeit 0 wie die anderen Folgen auch, aber sie kann eben trotzdem auftreten. Sie ist sozusagen gleichberechtigt zu allen anderen konkreten Folgen. Damit haben wir wieder unsere Lücke. Die relative Häufigkeit kann somit nicht zur *Definition* der Wahrscheinlichkeit dienen, Kopf zu werfen, denn in unserem Beispiel ist diese Wahrscheinlichkeit 1/2, aber die relative Häufigkeit in der B-Folge bleibt immer 1 und konvergiert so ebenfalls gegen 1. Das gilt übrigens auch für jede Teilfolge, die wir daraus auswählen, und damit ist die Folge sogar zufällig im Sinne von von Mises.

Das Besondere an unserer Folge B ist, dass es nur eine solche Folge gibt und nur wenige ähnliche Folgen mit wenigen Nullen, aber viel mehr Folgen, für die die relative Häufigkeit in der Nähe von 1/2 liegt. Daher ist es sehr viel wahrscheinlicher, Folgen mit einer relativen Häufigkeit nahe 1/2 zu finden, als Folgen, die die relative Häufigkeit 1 aufweisen. Doch für eine definitorische Festlegung ist das zu wenig. Das Gesetz der großen Zahlen besagt zwar, dass die Wahrscheinlichkeit für eine Folge mit dem »falschen« Grenzwert Null ist, aber das bedeutet leider nicht, dass sie nicht auftreten können. Dann hilft es uns für unser Problem nicht weiter. Außerdem setzt das Gesetz voraus, dass in jedem Einzelversuch dieselbe Wahrscheinlichkeit für das Auftreten eines Ereignisses A vorliegt, was nur im Rahmen der Propensitätenkonzeption angenommen werden darf.

Zusammenfassend ergibt sich: Insbesondere Empiristen können eigentlich nicht glücklich mit der ganzen Konzeption sein. Fiktive Folgenglieder sind nicht beobachtbar und damit sind die hier betrachteten hypothetischen relativen Häufigkeiten ebenso wenig beobachtbar. Ins-

besondere dürfte das für unendliche Folgen deutlich sein, denn alle tatsächlichen Beobachtungen betreffen nur endlich viele Ereignisse vom Typ eines Münzwurfs. Haben wir tatsächlich nichts anderes zur Verfügung als die rein mathematischen Bestimmungen der Kollektive von von Mises, wird es gleich noch viel schlimmer, denn jede endliche Teilfolge einer unendlichen Folge F sagt uns nichts über den Grenzwert der Folge F. Wir könnten also insbesondere jeden endlichen Anfangsabschnitt einer solchen Folge durch einen beliebigen anderen Anfangsabschnitt ersetzen und erhielten trotzdem denselben Grenzwert der relativen Häufigkeiten. Das bedeutet: Wenn es sich nur um von Misesche Kollektive handelt, geben uns tatsächlich beobachtete relative Häufigkeiten nicht mehr den geringsten Hinweis auf die relativen Häufigkeiten auf lange Sicht und damit keinen Hinweis mehr auf die zugrunde liegenden Wahrscheinlichkeiten. Die bleiben empirisch völlig unzugänglich. Die endlichen relativen Häufigkeiten sind noch nicht einmal mit hoher Wahrscheinlichkeit approximative Hinweise auf den gesuchten Grenzwert. Das ist verheerend für diesen Ansatz.

Außerdem sind selbst Wahrscheinlichkeiten für längere endliche Folgen dann sinnlos, wenn es keine Einzelfallwahrscheinlichkeiten gibt, denn jede längere endliche Folge solcher Ereignisse – also z.B. 1000 Münzwürfe mit derselben Münze – ist selbst ein Einzelfall eines komplexeren Zufallsexperiments. Das könnte etwa aus einzelnen Ereignissen bestehen, die immer Experimentserien von 1000 Münzwürfen sind und damit natürlich einen etwas größeren Ereignisraum haben. Jedenfalls würde es dann zugleich sinnlos zu sagen: Die Wahrscheinlichkeit bei dieser einen Experimentserie, mehr als 950-mal Kopf zu erhalten, ist sehr klein. Wenn aber selbst solche Aussagen sinnlos werden, dann ist dieses Konzept von Wahrscheinlichkeit überhaupt nicht mehr in der empirischen Welt anwendbar und damit genau genommen keine Interpretation von Wahrscheinlichkeit, sondern nur noch eine innermathematische Erläuterung des Konzepts.

Letztlich werden wir also nicht umhin kommen, doch wieder über Einzelfallwahrscheinlichkeiten zu sprechen, wenn wir überhaupt einen empirischen Wahrscheinlichkeitsbegriff mit realen Anwendungen erhalten wollen. Dabei müssen wir uns auf die jeweiligen Erzeugungsbedingungen der Resultate beziehen. Das ist bereits in den Definitionen von

Kolmogorov und von von Mises an einigen Stellen angelegt. Doch es war vor allem Popper, der daraus eine Interpretation von Wahrscheinlichkeit entwickelt hat, die als Propensitätenkonzeption weiter unten erläutert werden wird. Es gibt eine Reihe weiterer Probleme dieser Häufigkeitskonzeption, die man etwa bei Hajek (2010) findet, aber die genannten sollten genügen, die Konzeption ad acta zu legen.

5.4.3 Die klassische Auffassung der Wahrscheinlichkeit

Ein klassischer Ansatz ist der von Laplace, wonach die Wahrscheinlichkeit für ein Ereignis einfach in der Anzahl der günstigen Fälle geteilt durch die Anzahl der möglichen Fälle zu sehen ist:

Laplacesche Definition
P(A) = (Anzahl der günstigen Fälle) / (Anzahl aller gleichmöglichen Fälle)

In bestimmten Fällen ist das gut nachvollziehbar wie etwa beim fairen Würfel. Bei Würfen mit zwei fairen Würfeln sind insgesamt 36 verschiedene Konstellationen denkbar, die wir alle für gleichmöglich halten. Wie hoch ist dann die Wahrscheinlichkeit dafür, insgesamt drei Augen zu erhalten? Dazu müssen wir nur noch die günstigen Fälle zählen: <1,2> und <2,1>. Also ist die gesuchte Wahrscheinlichkeit gerade 1/18.

Doch die ersten Probleme mit dieser Auffassung setzen bei unserem Verständnis von »gleichmöglich« an. Wie sollen wir diesen Ausdruck verstehen, ohne ihn einfach als »gleichwahrscheinlich« zu lesen? Würde unsere Definition aber »gleichwahrscheinlich« enthalten, erschiene sie zirkulär, denn wir möchten schließlich erst erläutern, was wir mit bestimmten Wahrscheinlichkeitsaussagen meinen.

Ein zweites Problem ist dann, dass die Definition nur auf solche Fälle anwendbar ist, bei der wir eine eindeutige Menge von gleichmöglichen Fällen haben. Auf einen nicht-fairen Würfel ist die laplacesche Konzeption bereits nicht mehr anwendbar. Im Hintergrund des Ansatzes finden wir das *Indifferenzprinzip* (s.o.), wonach wir die Wahrscheinlichkeit auf alle Fälle in gleicher Weise aufteilen sollten, für die wir keine Gründe haben, die eine oder die andere Möglichkeit vorzuziehen. Das ist auch eine Grundidee der logischen Wahrscheinlichkeitsansätze.

Genau genommen ist das die epistemische Lesart des Indifferenzprinzips, die uns hier nicht interessiert. Wir können uns daneben eine objektive physikalische Lesart vorstellen. Danach sind die Fälle gleich zu behandeln, in denen eine *objektive Symmetrie* in den Bedingungen vorliegt, wonach die Möglichkeiten (bzw. hier die Seiten des Würfels) also in allen relevanten Bedingungen übereinstimmen. Es geht dann für den Würfel nicht darum, dass wir kein Wissen darüber haben, ob eine Seite anders ist als die anderen, sondern darum, dass die Seiten objektiv einander ähnlich sind bzw. objektive Symmetrien des Würfels vorliegen. Dazu benötigen wir bereits eine Menge an Wissen über die Situation, während das epistemische Indifferenzprinzip, um das es uns meistens geht, gerade aus dem Nichtvorliegen von Informationen Schlussfolgerungen zieht. Leider ist diese klassische Konzeption in beiden Lesarten nur sehr eingeschränkt anwendbar und wir sind uns nicht immer darüber im Klaren, dass wir dabei auf das Indifferenzprinzip angewiesen sind, von dem wir schon wissen, dass seine Anwendung uns in Paradoxien führen kann.

Die Grundidee dieser Probleme ließ sich bereits an dem sehr einfachen Urnen-Beispiel mit roten, blauen und weißen Kugeln erkennen. Wir können die Urne als bunte und weiße Kugeln enthaltend oder als blaue, rote und weiße Kugeln enthaltend beschreiben und erhalten unterschiedliche Wahrscheinlichkeitswerte. Beide Anwendungen scheinen aber durchaus zu rechtfertigen zu sein. Dann wird die Anwendung des Indifferenzprinzips letztlich eine beschreibungsabhängige Angelegenheit, die bestenfalls für epistemische Wahrscheinlichkeiten geeignet erscheint. Natürlich können wir uns in diesem Fall noch überlegen, dass womöglich die erste Beschreibung bei unserem momentanen Kenntnisstand die grundlegendere ist und sie daher zu bevorzugen sei. Oder wir können zumindest sagen, dass es nicht allzu viel Spielraum bei der Zuweisung von Wahrscheinlichkeiten gibt; allerdings müssen wir bestimmte Mehrdeutigkeiten akzeptieren. Insbesondere für die Unterteilung von kontinuierlichen Größen traten entsprechende Mehrdeutigkeiten in der Anwendung des Indifferenzprinzips auf (s.o.). Also liefert das Indifferenzprinzip keine eindeutigen Wahrscheinlichkeitsverteilungen für solche Fälle mehr. Vertreter einer modernen induktiven Logik würden allerdings zugestehen, dass wir nicht in allen Fällen immer eindeutige

Wahrscheinlichkeiten erhalten müssen. Es genügen manchmal schon mehrdeutige Verteilungen, um damit weitere Schlüsse ziehen zu können oder diese mit weiteren Daten upzudaten. Das sei immer noch besser, als bloß über rein subjektive Ausgangswahrscheinlichkeiten zu verfügen, wie es bei den Bayesianern der Fall sei. Allerdings haben wir es an dieser Stelle wieder mit epistemischen Wahrscheinlichkeiten zu tun und nicht mehr mit objektiven physikalischen, nach denen wir hier eigentlich suchen. Die klassische Wahrscheinlichkeitskonzeption hilft uns daher an dieser Stelle nicht viel weiter.

5.4.4 Reduktion der physikalischen Wahrscheinlichkeit auf die subjektive

Insbesondere David Lewis (1980) hat eine recht trickreiche Reduktion der objektiven physikalischen Einzelfallwahrscheinlichkeit, die er *Chance* nennt, auf subjektive Wahrscheinlichkeiten versucht. So sagt er zu Beginn seines Aufsatzes, dass sein sogenanntes Hauptprinzip (das wir bereits kurz eingeführt haben) alles wäre, was man über Chancen sagen kann; erst später drückt er sich dann wieder etwas zurückhaltender aus.

Hauptprinzip (»Lewis' principal principle«)
$P(A|ch(A)=r \,\&\, E) = r$ ist rational für zulässige Informationen E

Das Hauptprinzip besagt, dass es rational ist, sich in seiner subjektiven Wahrscheinlichkeit P(A) für A an den Chancen für A zu orientieren, wenn ansonsten nur »normale« zulässige Informationen vorliegen. Damit sind vor allem Informationen über die Vergangenheit gemeint, und es werden z.B. solche durch einen perfekten Hellseher ausgeschlossen. David Lewis gibt noch weitere Hinweise dazu, welche Informationen zulässig sind (s.o.), und es finden sich weitere Debatten zu diesem Punkt, die für uns an dieser Stelle jedoch nicht unbedingt bedeutsam sind.

Weiß ich etwa über eine Münze, dass die objektive Chance für Kopf 0,6 ist, so sollte ich das zugleich als subjektive Wahrscheinlichkeit wählen, selbst wenn die Münze in der Vergangenheit viel häufiger auf Zahl gefallen ist. Solche Informationen werden durch die Information über die tatsächliche Chance übertrumpft. Eine Information über Chancen bzw. über objektive physikalische Wahrscheinlichkeiten ist also etwas,

das in der Lage ist, andere Informationen aus dem Felde zu schlagen. Das zeigt wiederum, wie grundlegend diese objektiven Wahrscheinlichkeiten sind, und es scheint eine recht naheliegende Forderung zu sein, dass sich unsere subjektiven Wahrscheinlichkeiten daran orientieren sollten. Doch das ist hier nicht unser Thema, sondern die Frage, inwiefern uns das Hauptprinzip eine geeignete Charakterisierung der Chancen liefert. Leider kann das Hauptprinzip das aus verschiedenen Gründen gerade nicht leisten.

Zunächst einmal haben wir gute Gründe für die Annahme, dass *Glaubensgrad* ebenfalls ein Grundbegriff ist bzw. einen theoretischen Term darstellt (vgl. Eriksson & Hajek 2007), für den es jedenfalls sehr schwer ist, Gründe dafür zu finden, dass er sich an den Wahrscheinlichkeitsaxiomen orientieren sollte. Das hat unsere bisherige Debatte schon gezeigt. Daher kann ich den Optimismus von Rosenthal (2004) auch nicht teilen, dass wir mit dem Hauptprinzip sogleich schnell die Wahrscheinlichkeitsaxiome für das Konzept der Chancen erhielten. Das Hauptprinzip dient auch nicht als Brückenprinzip für einen theoretischen Term, da in diesem Prinzip nur zwei theoretische Terme miteinander verbunden werden, statt dass eine Verbindung zwischen theoretischen Termen und Beobachtungsausdrücken hergestellt würde. Denn selbst, wenn wir Eriksson & Hajek (2007) nicht darin folgen sollten, dass Glaubensgrade Grundbegriffe darstellen, so dürfte doch klar sein, dass die *rationalen* Glaubensgrade nicht messbar sind, sondern nur die tatsächlichen.

Das Hauptprinzip lässt auch völlig offen, was Chancen sind und ob es solche Größen mit normativer Kraft überhaupt gibt (vgl. Rosenthal 2004). Man hat vielmehr den Eindruck, dass hier die falsche Reihenfolge der Explikation gewählt wird, da die Chancen basaler sind als die subjektiven Wahrscheinlichkeiten, die sich an den Chancen orientieren sollten, sollten die Chancen als Erstes expliziert werden. Außerdem wird auch der für Chancen zentrale Zusammenhang zu relativen Häufigkeiten hier nicht thematisiert. Insgesamt scheint daher das Hauptprinzip nicht den entscheidenden Fortschritt für die Explikation von objektiven physikalischen Wahrscheinlichkeiten darzustellen. Es beschreibt uns zwar eine wichtige Aufgabe für Chancen, aber es kann nicht viel Erhellendes darüber sagen, was es ist, das diese Aufgabe wahrnimmt.

In der umfangreichen modernen Debatte um das Hauptprinzip geht es dagegen oft um die Frage, ob es sich mit der empiristischen Forderung nach der *Humeschen Supervenienz* vereinbaren lässt (vgl. z.B. Lewis 1994, Schaffer 2003, Schwarz 2014). Diese Forderung nach einer möglichst einfachen Ontologie ohne die Annahme notwendiger Beziehungen in der Natur (Gesetze, Kausalität und objektive Wahrscheinlichkeiten supervenieren demnach auf den kategorischen (nicht-dispositionalen) Eigenschaften, die an bestimmten Raum-Zeit-Punkten instantiiert sind) ist aber selbst recht umstritten (vgl. Esfeld 2010a). In einer Humeschen Welt gibt es nur zufällige Regularitäten und unsere Gesetzesaussagen dienen dazu, diese in möglichst systematischer Art und Weise zu beschreiben. Dabei können auch probabilistische Gesetze auftreten, die uns dann die Chancen für das Hauptprinzip liefern.

Eine moderne Ausarbeitung dieser Konzeption von Chancen im Sinne von Lewis findet sich bei Hoefer (2007). Hoefer geht über Lewis hinaus. Für Hoefer sind es nicht nur Naturgesetze, die uns Chancen liefern, sondern auch *stochastische nomologische Maschinen* wie etwa Roulettegeräte. Wenn wir auch für sie unsere Regularitäten in möglichst systematischer Weise beschreiben möchten, können wir dafür weitere Chancen einsetzen, die uns dann sogar weitere Einzelfallwahrscheinlichkeiten liefern. Doch m.E. setzt die Annahme solcher stochastischen nomologischen Maschinen bereits die Annahme von Propensitäten voraus, die in den einzelnen Instanzen der Maschinen immer wieder wirksam sind, sonst ist es eher ein Rätsel, warum wir erwarten sollten, dass sich die bisher gezeigten relativen Häufigkeiten – etwa bei bestimmten Rouletterädern – in der Zukunft in ähnlicher Weise zeigen sollten. Doch nur, wenn wir Grund zu der Annahme einer entsprechenden Konstanz haben, ist das Hauptprinzip eine sinnvolle Rationalitätsforderung. An dieser Stelle drohen daher wiederum empiristischen Beschränkungen ein angemesseneres Verständnis unserer wissenschaftlichen Grundbegriffe zu behindern, aber ich werde dem hier nicht weiter nachgehen.

5.4.5 Chaotische Systeme

Für bestimmte deterministische Systeme, die uns aber indeterministisch erscheinen (etwa aufgrund ihrer Komplexität oder weil wir die

Anfangsbedingungen nicht beliebig genau bestimmen können), gibt es noch eine interessante Interpretation, die Rosenthal (2004, 2012) die *Inhaltsauffassung* oder die »range conception« nennt. Die Grundidee ist recht anschaulich. Wir stellen dafür jeden Zustand eines Systems durch einen Punkt in seinem Zustandsraum Z dar. Für ein abgeschlossenes n-Teilchen-System benötigen wir für jedes Teilchen etwa 3 Dimensionen für seinen Ort und 3 Dimensionen für seinen Impuls. Damit wird Z eine Teilmenge des 6n-dimensionalen reellen Raumes. Für eine Münze, die wir werfen, wird der Zustandsraum M für den Anfangszustand etwa durch die Höhe über dem Tisch, bei der wir sie loslassen, sowie dem Impuls und dem Drehimpuls angegeben und womöglich einigen weiteren Randbedingungen. Da wir davon ausgehen, dass als Endzustand nur die beiden Zustände {Kopf, Zahl} relevant sind, teilen wir nun den Ausgangsraum Z in zwei Teilmengen auf, nämlich die Zustände K, die zu Kopf führen und die Zustände Z, die zu Zahl führen: $M = K \cup Z$. Sind diese beiden Mengen in der Grundmenge M gut vermengt, so können wir bei normaler begrenzter Kenntnis des Ausgangszustandes nicht mehr vorhersagen, ob *Kopf* oder *Zahl* auftreten wird.

Im Idealfall liegen K und Z so zerstreut in M, dass in jeder offenen Umgebung eines Punktes aus K noch mindestens einer aus Z zu finden ist und umgekehrt. (Genauer sollten wir fordern, dass in jeder offenen Umgebung eines Punktes aus K noch mindestens eine Menge von Punkten aus Z mit einem Lebesgue-Maß echt größer Null zu finden ist und umgekehrt.) Dann erscheint uns das System als komplett indeterminiert, wenn wir den jeweiligen Anfangszustand nicht mit unendlicher Genauigkeit kennen, was physikalisch unmöglich zu sein scheint. Um für ein solches System noch Vorhersagen abgeben zu können, sind wir auf das Konzept der Wahrscheinlichkeit daher zwingend angewiesen. Wir benötigen etwa ein Inhaltsmaß μ (z.B. das Lebesgue-Maß) auf dem Raum M und vergeben die Wahrscheinlichkeiten für das Auftreten von Kopf und Zahl nach ihren Inhaltsanteilen am Ausgangsraum M. Sollten diese darüber hinaus noch relativ stabil für die meisten Intervalle in unserem Raum M sein, können wir diese Anteile als Wahrscheinlichkeiten auffassen: Wissen wir etwa, dass der Ausgangszustand in einem (m-dimensionalen) Intervall $I \subseteq M$ liegt, dann gilt:

$$P_I(\text{Kopf}) = \mu(K \cap I)/\mu(I) \approx \mu(K)/\mu(M) = P(\text{Kopf})$$

Michael Strevens (2003) untersucht die Anwendbarkeit und Erklärungskraft dieser Wahrscheinlichkeitsinterpretation und nennt die Konstanzforderung: *Mikrokonstanz*. Wenn die vorliegt, wenn also für alle (nicht zu kleinen) Intervalle I im Bereich M dieser Wert approximativ derselbe ist, so können wir sagen, dass er die objektive Wahrscheinlichkeit dafür liefert, dass wir Kopf erhalten. Ergibt sich etwa ein Wert von 0,6, so bedeutet das, dass eben 60% der Zustände am Anfangsraum zu Kopf führen und das ist eine objektive Auskunft über unser System und keine Auskunft nur über unsere Unkenntnis. Außerdem gewinnen wir damit ein aussagekräftiges Konzept für unsere Prognosen. Im Durchschnitt werden wir in 60% unserer Würfe das Ergebnis *Kopf* erzielen.

Ein solches System können wir äußerlich überhaupt nicht von einem genuin indeterministischen System unterscheiden. Strevens (2003) weist darauf hin, dass wir manche komplexen Systeme so beschreiben *müssen*. Einen besseren Zugang haben wir zu solchen Systemen oft nicht. Es bleibt dann nur die Frage offen, ob ein bestimmtes System bloß chaotisch ist, jedoch eigentlich ein deterministisches System darstellt, oder ob es nicht sogar genuin indeterministisch ist. Das werden wir vielfach nicht wirklich entscheiden können. Die indeterministischen Einflüsse können etwa durch quantenmechanische Effekte hervorgerufen werden, die sich für die Münze etwa aus Stößen mit bestimmten Atomen der Luft ergeben können. Zum anderen ist nicht ausgemacht, dass wir es außerhalb der Quantenmechanik mit einer deterministischen Welt zu tun haben. Die meisten unserer physikalischen Basistheorien sind letztlich nicht deterministisch (vgl. Earman 1986). Das gilt umso mehr, wenn wir beachten, dass es sich bei den zugrundegelegten kontinuierlichen Räumen genau genommen um eine Idealisierung handelt. Wir haben keine guten Gründe für die Annahme, dass unser physikalischer Raum tatsächlich kontinuierlich statt diskret mit sehr kleinen diskreten Abständen ist. Doch erst die Annahme beliebig genauer Anfangsbedingungen führt zu einer teilweise deterministischen nomischen Struktur, die uns etwa in der newtonschen Mechanik den Eindruck vermittelt, sie beschreibe eine deterministische Welt.

Für viele Systeme sollten wir daher am besten Agnostiker bleiben in Fragen des Determinismus. Da es zumindest einige vermutlich indeterministische Systeme gibt, müssen wir in unserem Interpretationsprojekt auf jeden Fall auch für genuin ontologisch indeterministische Systeme Vorsorge treffen und dafür eine Interpretation entwickeln, die m.E. vor allem in einem Propensitätenansatz im Sinne Poppers zu suchen ist.

Wir sind auch nicht zwingend darauf angewiesen anzunehmen, dass alle Anfangszustände genau gleichwahrscheinlich sind. Die sogenannte Methode der willkürlichen Funktionen gestattet es uns, beliebige aber noch relativ glatte Dichtefunktionen δ auf M einzuführen und wir erhalten das Resultat, dass wir bei Vorliegen der Mikrokonstanz auch für mit δ gewichteten Bereiche wieder eine Konstanz für die folgenden Werte für alle entsprechenden Intervalle I (s.o.) erhalten:

$$P_I(\text{Kopf}) = \int_{K \cap I} \delta(x)\, d\mu(x) \approx P(\text{Kopf})$$

Hier werden also durch δ Gewichtungen zugelassen, die besagen, dass bestimmte Anfangsbedingungen durchaus öfter auftreten können als andere. Sind die Unterschiede dabei aber nicht zu extrem, sorgt die Mikrokonstanz dafür, dass sich an den wie oben definierten Wahrscheinlichkeiten nichts ändert.

Uns erscheinen zumindest viele Lebensbereich recht zufällig und es wäre schon eine gewagte Annahme, dass das, was uns im Laufe mehrerer Jahre im Leben zustößt (etwa im Straßenverkehr), zu Beginn schon determiniert war oder sogar bereits sehr lange Zeit vor unserer Geburt determiniert war. Die Konzeption eines weitgehenden Determinismus widerspricht (auch in der Makrowelt) vielfach unserer Lebenserfahrung und wird in der Philosophie manchmal zu schnell als gut begründet angenommen. Trotzdem wären wir selbst in einer deterministischen Welt auf Wahrscheinlichkeitsaussagen angewiesen und dafür kann die Inhaltsauffassung eine Erläuterung anbieten, wie wir die interpretieren können. Da sie ganz wesentlich auf bestimmte objektive Eigenschaften des Raumes M und damit des betrachteten Systems Bezug nimmt, können wir sie ebenfalls als objektive Konzeption von Wahrscheinlichkeit bezeichnen.

5.4.6 Poppersche Propensitäten als theoretischer Term

Für die Anwendung der klassischen Statistik und der Idee aus zahlreichen Wiederholungen eines Experiments auf bestimmte involvierte Wahrscheinlichkeiten schließen zu können, sind wir auf die Annahme von Einzelfallwahrscheinlichkeiten angewiesen, die in jedem der Experimente identisch sind. Daher benötigt wir zwingend eine Interpretation von Wahrscheinlichkeit, die mit diesen Einzelfallwahrscheinlichkeiten kompatibel ist. Die findet sich vor allem in dem folgenden Ansatz.

Dabei gehen wir der Einfachheit nun davon aus, dass wir mit unseren Wahrscheinlichkeitsaussagen Systeme beschreiben, die genuin *indeterministisch* sind. Da die Häufigkeitenkonzeption versagt hat und die Reduktionsversuche auf rein subjektive Wahrscheinlichkeiten ebenso wenig erfolgreich waren, führte Popper ein neues Konzept ein: *Propensität*. Mit Propensität bezeichnet er eine *dispositionale Eigenschaft* bestimmter Systeme, mit einer gewissen Stärke ein bestimmtes Ereignis bzw. ein bestimmtes Eigenschaftsvorkommnis hervorzubringen. In unserem Münzwurfbeispiel besitzt das System bestehend aus der Münze und dem Werfer sowie weiteren Umgebungsbedingungen etwa eine Tendenz der Stärke 0,5 bei entsprechenden Würfen dafür, dass Kopf oben zu liegen kommt. Damit ist die *Verwirklichungstendenz* für Kopf und die für Zahl (bzw. kein Kopf) hier gleich groß. Dabei handelt es sich um Tendenzen für den Einzelfall, aber messbar werden sie natürlich erst durch Wiederholung des Vorgangs, denn erst dadurch lassen sich diese Tendenzen anhand der relativen Häufigkeiten quantitativ bestimmen.

Dieser Zusammenhang hat Gillies (2000) und Rosenthal (2004) wohl dazu verführt anzunehmen, Popper könnte eine »long run« Propensität im Sinn gehabt haben, doch die Propensität soll uns nach Popper vor allem helfen, die probabilistischen Einzelfallaussagen etwa im Rahmen der Quantenmechanik zu verstehen. Nur für die *Messung* von Propensitäten sind wir dann auf längere Versuchsreihen angewiesen. Die (zugegeben etwas vage) Bedeutung dieser Quantität einer solchen Propensität wäre approximativ durch die relativen Häufigkeiten längerer Versuchsreihen gegeben.

Besser sind Propensitäten als *theoretische Terme* zu verstehen, die bestimmte Axiome und bestimmte Brückenprinzipien zu erfüllen haben

und vor allem daraus ihre Bedeutung beziehen. Da auch die subjektiven Wahrscheinlichkeiten bzw. Glaubensgrade sich ebenfalls nur als theoretische Terme darstellen ließen, ist das für die noch grundlegenderen Propensitäten eigentlich keine besondere Überraschung mehr. Sie sind Eigenschaften, die ganzen Systemen zukommen können und dann auf den möglicherweise noch grundlegenderen Eigenschaften dieser Systeme supervenieren – oder sie sind selbst bereits basal. Die jeweiligen Realisierungen solcher Propensitäten können jedoch sehr unterschiedlich aussehen, was Eagle (2004) kritisch anmerkt. Aber dasselbe Problem kennen wir bereits aus dem Beispiel der *Fitness* in der Biologie. Was jeweils mit Fitness gemeint ist, variiert sehr stark von Fall zu Fall. Es können lange Hälse sein oder ein gutes Immunsystem, um auch Aas fressen zu können, oder spezielle Vorrichtungen und Fertigkeiten zum Klettern oder bestimmte Färbungen der Außenhaut etc. Trotzdem bringt das theoretische Konzept der Fitness all diese unterschiedlichen Eigenschaften unter ein evolutionäres nomisches Muster, mit dessen Hilfe wir das Auftreten dieser Eigenschaften erklären und besser verstehen können.

Ähnlich müssen wir Propensitäten verstehen. Sie lassen uns neue nomische Muster erkennen. Dabei treten bestimmte Eigenschaften lokal zwar völlig regellos auf, aber auf längere Sicht zeigt sich eine große Stabilität der relativen Häufigkeiten, mit deren Hilfe wir bestimmte Ergebnisse prognostizieren können. Realisiert werden die Propensitäten dabei durch unterschiedliche Systeme, das sollte aber kein entscheidender Einwand für einen theoretischen Term sein. Es zählt vielmehr seine Vereinheitlichungs- und Erklärungskraft. Die Frage ist allerdings, welche Brückenprinzipien und eventuell weiteren methodologischen Regeln uns das Konzept der Propensität tatsächlich explizieren helfen, und es nach Möglichkeit sogar gestatten, es zumindest in einigen Anwendungen wenigstens approximativ zu *messen*.

Zunächst können wir feststellen, dass uns allgemein Dispositionen aus dem Alltag und der Wissenschaft recht vertraut sind. Allerdings weisen die popperschen Verwirklichungstendenzen doch bestimmte Eigenheiten auf. Eine Disposition wie *zerbrechlich zu sein* können wir etwa durch ein kontrafaktisches Konditional im Sinne einer konditionalen Analyse (KA) erläutern:

(KA) »x ist zerbrechlich« bedeutet in etwa:
»Würde man x auf einen Steinboden werfen, würde x zerbrechen.«

Propensitäten sind dagegen nur *probabilistische* Dispositionen, die nicht in jedem Fall bestimmte Ergebnisse zeitigen, sobald die Manifestationsbedingungen der Disposition vorliegen, sondern eben nur in manchen Fällen. Allerdings gilt genau genommen etwas Ähnliches ebenfalls für normale Dispositionen. Selbst eine an sich zerbrechliche Flasche zerspringt nicht immer zuverlässig beim Auftreffen auf einen harten Boden. Der Mutige wird nicht in jeder Situation wieder seinen Mut beweisen etc. Neu ist eher, dass wir die Stärke der Disposition nun quantifizieren und sogar sehr schwache Dispositionen (solche mit kleinen Wahrscheinlichkeiten) zulassen. Außerdem muss Popper darauf bestehen, dass zumindest bestimmte Dispositionen wie die Propensität keiner reduktiven konditionalen Analyse zugänglich sind. Dann liefert (KA) zwar gute Indikatoren für das Vorliegen einer solchen Disposition, ist aber keineswegs bedeutungsgleich oder extensionsgleich mit dem Vorliegen der Disposition. (KA) stellt nur eine Indikatoraussage dar, sowie das Testergebnis in einem IQ-Test einen Indikator für das Vorliegen von Intelligenz darstellt, aber keineswegs damit zu identifizieren ist. Stärkere Forderungen nach Operationalisierung theoretischer Terme haben selbst Empiristen wie Carnap schon zurückgewiesen (s.u.). Allerdings weist die konditionale Analyse selbst auch viele Probleme auf, so dass wir sie nicht einmal für normale Dispositionen verbindlich verlangen sollen (vgl. Bartelborth 2007).

Neben den Wahrscheinlichkeitsaxiomen als den *Grundgesetzen* für einen Propensitätsbegriff benötigen wir noch *Brückenprinzipien*, um den Begriff weiter zu charakterisieren. Allerdings dürfen wir uns von diesen Prinzipien nicht zu viel erwarten. Es sind zunächst einmal gesetzesartige Aussagen, die sowohl theoretische Begriffe wie Beobachtungsbegriffe beinhalten. Viel mehr verlangen auch Empiristen wie Carnap nicht mehr von Brückenprinzipien. Auf Reduktionsbehauptungen wird für theoretische Terme inzwischen verzichtet.

> There is a temptation at times to think that the set of rules provides a means for defining theoretical terms, whereas just the opposite is

> really true. A theoretical term can never be explicitly defined on the basis of observable terms, [...]. (Carnap 1995, 234).
>
> There is always the possibility of adding new rules, thereby increasing the amount of interpretation specified for the theoretical terms; but no matter how much this is increased, the interpretation is never final. (Carnap 1995, 238)

Auch darf man sich diese Regeln nicht unbedingt als präzise Gleichungen denken, denn die Beobachtungsterme müssen keineswegs immer scharfe quantitative Größen sein. Aus vielen Beispielen aus der Wissenschaftsgeschichte hat etwa Ernest Nagel das gelernt:

> The haziness that surrounds such correspondence rules is inevitable, since experimental ideas do not have the sharp contours that theoretical notions possess. This is the primary reason why it is not possible to formalize with much precision the rules (or habits) for establishing a correspondence between theoretical and experimental ideas. (Nagel 1979, 100)

Wir werden aber über die Forderungen der Empiristen hinausgehen und werden noch nach weiteren methodologischen Regeln suchen, die es uns erlauben, die theoretischen Terme zumindest in manchen Anwendungen approximativ zu bestimmen.

Schauen wir uns das Verfahren für das theoretische Konzept des elektrischen Feldes E einmal exemplarisch dazu an. Die Maxwellschen Gleichungen stellen die Grundaxiome für E zur Verfügung und die Lorentzkraftgleichung $F = qE$ liefert einen Zusammenhang für Partikel mit Ladung q, die sich im Feld E befinden, zu einer beobachtbaren Größe, nämlich der Kraft, die das Feld E auf einen Partikel mit Ladung q ausübt. Unser Verfahren zur Messung von E sieht dann etwa so aus, dass wir einen Probepartikel mit einer recht kleinen Ladung q in das elektrische Feld einbringen und die Kraft auf dieses Partikel messen, um dann anhand des Lorentzkraftgesetzes auf die Feldstärke an dieser einen Stelle schließen zu können. Dabei treten allerdings eine ganze Reihe von Problemen auf. Zunächst wirken noch andere Kräfte auf das Partikel wie etwa die Gravitationskräfte aller anderen Objekte, die wir nicht

abschirmen können. Gravierender ist allerdings, dass das Einbringen des Probeteilchens in das Feld bereits das Feld verändert. Wir müssen hier mit kleinen Ladungen arbeiten, die dann nur kleine Änderungen des Feldes bedeuten. Deren Umfang können wir allerdings wiederum nur anhand der Maxwellschen Gleichungen abschätzen, die wir u.a. mit diesem Verfahren empirisch testen wollen.

Des weiteren bestimmen wir das Feld auf diese Weise nur an endlich vielen einzelnen Stellen: Es ist aber auf einem Kontinuum definiert. Außerdem funktioniert das Verfahren nicht mehr für sich schnell ändernde Felder. Schließlich haben wir weitere ganz spezielle Probleme, die elektrischen Felder innerhalb von Materie zu messen. Um ein solches Feld anhand endlich vieler Messungen weiter bestimmen zu können und für das ganze Kontinuum festzulegen, müssen wir bestimmte Stetigkeiten annehmen und annehmen, dass unsere Feldgleichungen auch für die weiteren Raumstellen gelten, und können diese dann eventuell berechnen. Schon an diesem Beispiel sieht man, dass das Messen von theoretischen Größen alles andere als ein Kinderspiel ist und natürlich mit einigen Idealisierungen und Approximationen verbunden ist.

Noch problematischer wird es, wenn wir die Ladung q ebenfalls zu den theoretischen Termen zählen, deren Werte nicht schon unabhängig von der Elektrodynamik zu bestimmen wären. Dann liefert uns das Lorentzkraftgesetz zwar immer noch ein Brückenprinzip, ist aber weit davon entfernt, zu konkreten Messverfahren zu führen. Diesem Problem können wir hier nicht weiter nachgehen, es illustriert aber noch einmal die Problematik von theoretischen Termen. Wir dürfen auch für seriöse theoretische Terme nicht erwarten, dass es einfache Messverfahren für sie gibt.

Ein weiteres Problem finden wir in den *Messfehlern*, wie sie etwa Freedman et al. (1983) beschreiben. Das amerikanische Eichamt hat über Jahre hinweg immer wieder das Probegewicht NB 10 mit einem Nominalgewicht von 10g gemessen und dabei besonders genaue Messungen für diese recht einfache Größe durchgeführt. Trotzdem streuen die Messergebnisse immer in einem gewissen Ausmaß, und es finden sich sogar genuine Ausreißer. Das zeigen die Veröffentlichungen dieser Messungen, die als ein Vorbild für typische Messergebnisse gelten

können. Mehr darf man für die Messung einer quantitativen Größe eben nicht erwarten. Wir erhalten ein Intervall, in dem die meisten Messwerte enthalten sind, und das uns daher als Messergebnis gilt. Möchten wir uns dann noch auf einen Wert konzentrieren, wird man etwa den arithmetischen Mittelwert dazu hernehmen. Ähnlich werden wir auch für die Messung von Propensitäten vorgehen können. Es gilt aber immer die Regel:

Messwert = wahrer Wert + Messfehler (bzw. »Rauschen«)

Diese Messfehler erhalten wir schon im Falle einfacher Größen, und natürlich sind sie umso mehr im Falle theoretischer Größen zu erwarten, für die wir eher größere Probleme und Vagheiten bei der Messung zu gewärtigen haben, was wir im Falle der elektrischen Felder erkennen konnten.

Dafür benötigen wir als Erstes ein geeignetes Brückenprinzip. Häufig wird als mögliches Prinzip das *Gesetz der großen Zahlen* genannt, denn schließlich besagt es, dass die relativen Häufigkeiten sich im Grenzwert den Wahrscheinlichkeiten annähern, und stellt damit einen Zusammenhang zwischen Wahrscheinlichkeiten und relativen Häufigkeiten her. Doch es taugt nicht wirklich als Brückengesetz, weil die Grenzwerte der relativen Häufigkeiten selbst nicht beobachtbar sind, und ein Brückengesetz immer theoretische Größen und *beobachtbare Größen* zusammen enthalten sollte. Wir sollten uns also besser an eine Vorstufe des Gesetzes der großen Zahlen halten, nämlich an die *Tschebyscheffsche Ungleichung*.

Zu diesem Zwecke haben wir allerdings zunächst genauer zu überlegen, wie wir die Propensitäten konzipieren. Sie sollen Einzelfalltendenzen sein, die für einzelne konkrete Situationen σ eine bestimmte Verwirklichungstendenz $P_\sigma(a)$ für ein bestimmtes Resultat a vom Typ A angeben. Um sie mit einer bestimmten relativen Häufigkeit in Verbindung zu bringen, benötigen wir dann einen Situationstyp S(σ), der die relevante Situation beschreibt (bzw. die Referenzklasse zu σ angibt). Damit ist ein Typ von Situation gemeint, der praktisch alle für das Auftreten von A in σ relevanten Faktoren umfasst. In unserem Beispiel des Münzwurfs wäre A etwa das Resultat *Kopf* und a demnach das

konkrete Ereignis des Auftretens von Kopf in einer konkreten Situation σ, in der wir den Münzwurf betrachten. S(σ) wäre dann der dazu gehörige Typ von Situation (unsere Referenzklasse), der alle für das Auftreten von Kopf in unserem Anwendungsfall relevanten Merkmale der Situation enthält. So erst gelangen wir zu einer Reihe von Situationen, in denen wir eine bestimmte Zufallsvariable $Z_{n,S(\sigma)}(A)$, die die relative Häufigkeit von A-Ereignissen in der Folge unserer wiederholten Zufallsexperimente anzeigt, als Resultat von identisch verteilten unabhängigen Ereignissen betrachten können, die der Statistiker immer schnell annimmt, um seine mathematischen Theoreme wie das Gesetz der großen Zahlen oder die Tschebyscheffsche Ungleichung ableiten zu können.

5.4.7 Der Einwand von Gillies und eine transzendentale Annahme

Ob es so etwas wie den Situationstyp S(σ) aber überhaupt gibt, ist praktisch unsere erste Frage. Gillies (2000) bezweifelt das. Seiner Meinung nach kommt als Situationstyp nur ein gesamter Weltzustand vor a in Frage und bei Popper gibt es ebenfalls einige Äußerungen in dieser Richtung. Dann wären wir allerdings in großen Schwierigkeiten, denn wir könnten die Situation σ nicht wirklich wiederholen und erhielten damit auch keine entsprechenden relativen Häufigkeiten mehr, mit deren Hilfe wir $P_\sigma(a)$ zumindest in einigen Fällen approximativ bestimmen könnten. Unser Konzept von Propensität würde damit nach Gillies im starken Sinne *metaphysisch*, da wir es niemals messen könnten. Zumindest würde es kaum als praktisch anwendbares Konzept durchgehen und bliebe wohl nur eine theoretische Abstraktion.

Gillies (2000) argumentiert für seine These anhand eines klassischen Beispiels. Was ist die objektive Wahrscheinlichkeit für einen 40-jährigen Mann in seinem 41-ten Jahr zu sterben? In dem Jahr können ihm unüberschaubar viele Dinge zustoßen, die wir kaum nachbilden können. Er kann in einen Verkehrsunfall verwickelt werden, es kann ihm ein Blumentopf auf den Kopf fallen, er kann einem Verbrechen zum Opfer fallen oder einer Krankheit erliegen etc. Wie sollen wir seine spezielle Wahrscheinlichkeit in seiner Situation bestimmen? Wie könnten wir genau seine Situation mehrfach wiederholen? Das ist sicher nicht mehr praktisch ausführbar.

Dazu sind m.E. zwei Dinge zu sagen. Erstens müssen theoretische Terme nicht in all ihren Anwendungen bestimmbar sein. Denken wir daran, dass wir für das kontinuierliche elektrische Feld auch nur endlich viele ausgewählte Punkte für eine Messung heranziehen können. Aber es dürfte zweitens vor allem klar sein, dass es unterschiedliche Typen von Situationen gibt. Für einfache Spielsituationen (etwa beim Roulett) oder die Frage, ob ein Atom in einer bestimmten (nicht allzu langen) Zeit zerfallen wird, können wir viel eher die Situationen nachstellen und dürfen sogar die Hoffnung haben, dass wir dabei nur endlich viele Faktoren berücksichtigen müssen. Dazu unterscheide ich Situationen mit einem *lokalen Einflussbereich für A* und solche mit einem *globalen bzw. diffusen Einflussbereich* wie in Gillies Sterbebeispiel. Auf die Letzteren müssen wir das Konzept der Propensität nicht unbedingt anwenden. Dort hätte es jedenfalls nur theoretische Bedeutung.

Doch im Falle nur lokaler Einflussbereiche sieht die Sache ganz anders aus. Für viele solcher Situationen nehmen wir in der Wissenschaft und im Alltag mit großem Erfolg an, die Menge an relevanten Faktoren wäre endlich und relativ klein, so dass wir sie *approximativ* wiederholen können und so brauchbare Einschätzungen der Wahrscheinlichkeiten erhalten. Im Prinzip nehmen wir an, dass die Menge der wirklich relevanten Einflussfaktoren klein und sogar recht überschaubar bleibt. Viele unserer Praktiken in der Wissenschaft wie die Randomisierung würden sonst nicht funktionieren. Wir hätten auch bei anderen Gelegenheiten kaum eine Chance, Kausalzusammenhänge zu ermitteln, wenn es von möglichen einflussreichen Faktoren und Kofaktoren nur so wimmeln würde. Wie sollten uns dann bestimmte Zusammenhänge zwischen Einzelfaktoren überhaupt noch auffallen? Wir setzen in der Wissenschaft geradezu als *transzendentale Annahme* voraus, dass wir in vielen Fällen nur einen lokalen Einflussbereich haben, für den sich bestimmte Messverfahren anbieten, und jedenfalls schnell stabile relative Häufigkeiten auftreten (vgl. Bartelborth 2011). Daher gelingt es uns in solchen Fällen meist, den Situationstyp $S(\sigma)$ zu bestimmen und oft, weitere Instanzen zu finden, so dass sich eine Auswertung der relativen Häufigkeiten vornehmen lässt. Davon müssen wir zumindest in vielen Anwendungen ausgehen, wenn wir weiterhin hoffen wollen, erfolgreich Wissenschaft betreiben zu können und die Zusammenhänge der Natur

überhaupt enträtseln zu können. Für einfache Beispiele gesteht Gillies uns das sogar zu und mehr sollte man für einen theoretischen Term eben nicht erwarten.

Bestimmte idealisierende Annahmen sind dabei zulässig und gewisse Approximationen unerlässlich. Denken wir an die Prognosen des Gewichts oder Gravitationskraft anhand unserer Gravitationstheorien. Dabei wissen wir zwar, dass wir im Prinzip auch die gravitativen Einflüsse von Milliarden weit entfernter Himmelsobjekte berücksichtigen müssten (er ist auch nicht abschirmbar), doch das ist in der Praxis unmöglich und unnötig, denn in den meisten Anwendungen ist deren Einfluss sehr gering und kann daher vernachlässigt werden. Solche »kleinen Faktoren« dürfen wir auch bei der Bestimmung von $S(\sigma)$ vernachlässigen.

5.4.8 Ein Brückenprinzip und die Regeln der Propensitäts-Messung

Als Brückenprinzip für die Propensität (BPP) haben wir bereits oben die sogenannte Tschebyscheffsche Ungleichung genannt, die eine Verbindung zwischen den angeführten Größen angibt und damit den Boden für erste Messverfahren der Propensität in bestimmten Anwendungsfällen bereitet.

(BPP) Für alle A, $n \in \mathrm{N}$, $\varepsilon > 0$: $\mathrm{P}_{\mathrm{meta}}(|Z_{n,\mathrm{S}(\sigma)}(A) - \mathrm{P}_\sigma(a)| \leq \varepsilon) > 1 - 1/4n\varepsilon^2$,

dabei ist $Z_{\mathrm{n,S}(\sigma)}(\mathrm{A})$ eine Zufallsvariable, die die relative Häufigkeit der A-Ergebnisse in einer Versuchsreihe von n Versuchen mit nahezu gleichverteilten und voneinander unabhängigen Zufallsvariablen X_i beschreibt, die die einzelnen Resultate der Versuche festhalten und in $Z_{\mathrm{n,S}(\sigma)}(\mathrm{A}) = 1/\mathrm{n}\ \Sigma_i\ X_i$ zusammenführen (vgl. zum Folgenden Bartelborth 2011).

Einer echten Reduktion der Propensität durch (BPP) steht der Einsatz der speziellen Metawahrscheinlichkeit $\mathrm{P}_{\mathrm{meta}}$ entgegen, die in unserem Prinzip enthalten ist, die besagt, dass eine kleine Abweichung von Propensität und relativen Häufigkeiten selbst nur mit einer bestimmten Wahrscheinlichkeit zu erwarten ist. Diese Metawahrscheinlichkeit müssen wir zunächst besser verstehen. Aber wir dürfen eben nicht wieder in den Fehler verfallen, dabei eine Reduktion von Propensitäten zu erwarten. Selbst Carnap sah letztlich diese Forderung für theoretische

Terme als falsch an, wie sein obiges Zitat sehr deutlich zeigt. Um nun eine Aussage wie $P_{meta}(|Z-p|\leq\varepsilon) = P_{meta}(\varphi) > 0{,}99$ in (BPP) zu interpretieren, stehen uns unterschiedliche Möglichkeiten offen.

Zunächst können wir $P_{meta}(\varphi)$ ebenfalls als gewöhnliche *Propensität* auffassen. Dann liefert (BPP) immer noch eine Verbindung zwischen Propensitäten und relativen Häufigkeiten, aber eben keine reduzierende. Trotzdem liefern hier die Axiome und das Brückenprinzip eine partielle Bedeutung für Propensitäten. Allerdings werden durch (BPP) allein so noch nicht einmal einzelne Messungen von Propensitäten möglich. Dazu müssten wir verstärkt auf weitere methodologische Regeln setzen.

In einer zweiten Lesart können wir $P_{meta}(\varphi)$ als *epistemische Wahrscheinlichkeit* auffassen, die angibt, in welchem Ausmaß wir Grund zu der Annahme haben, dass φ auftritt. Das geht wieder einen Schritt weiter in die Richtung einer Reduktion der Propensitäten. Da wir ohnehin nicht jede Wahrscheinlichkeitsaussage als eine Aussage über objektive physikalische Wahrscheinlichkeiten deuten können, sind wir auch so schon darauf angewiesen, solche epistemischen Wahrscheinlichkeiten anzuerkennen und einzusetzen. Wenn wir sie für wohldefiniert und gut verständlich halten, können wir uns ohne Bedenken an dieser Stelle in (BPP) auf sie stützen, um damit unser Verständnis von Propensitäten zu befördern.

Klassische Statistiker werden allerdings nach einer dritten und möglichst objektiven Interpretation von $P_{meta}(\varphi)$ Ausschau halten und wären im Gegenzug vielleicht bereit, dafür eine *vagere* Interpretation zu akzeptieren, die dafür $P_{meta}(\varphi)$ näher an nicht-probabilistische Dispositionen heranbringt. Wenn $P_{meta}(\varphi)$ etwa größere Werte wie 0,95 oder 0,99 annimmt, können wir φ als *Normalfall* ansehen, d.h., wir gehen in der Praxis schlicht davon aus an, dass φ gilt. Nach Nagel (s. Zitat oben) sind solche Vagheiten für eine »Übersetzung« theoretischer Terme in die Alltagssprache zu erwarten. Für die Frage der empirischen Überprüfbarkeit bedeutet diese Interpretation, dass wir Daten, die gegen φ sprechen, ebenso als Daten betrachten dürfen, die gegen $P_{meta}(\varphi)$ sprechen. Ein Vorteil von (BPP) ist, dass wir solche vagen Interpretationen der Meta-Wahrscheinlichkeit nur für recht große Wahrscheinlichkeitswerte benötigen und die ergänzenden methodologischen Regeln daher auch nur für diesen Fall gelten müssen.

Wir könnten uns an dieser Stelle also auch auf eine Art von starkem statistischen Syllogismus stützen und damit auf φ schließen. Das ist dann schon relativ ähnlich zu den Fällen von normalen Dispositionen, in denen wir auch nicht in allen Fällen erwarten dürfen, dass sie sich manifestieren, sondern nur in der großen Mehrzahl der Fälle. Trotzdem werden wir den Objekten weiterhin etwa die Disposition der Zerbrechlichkeit zuschreiben, selbst wenn sie nicht bei jedem Fall zu Boden zerspringen. Ähnliche Schwellenwertregeln finden wir in der anderen Richtung in den Regeln der klassischen Signifikanztests, in denen konventionell bestimmte Irrtumswahrscheinlichkeiten festgelegt sind, die wir in probabilistischen Falsifikationen zu akzeptieren gedenken (vgl. dazu Kapitel 6).

Das Brückenprinzip (BPP) erlaubt es uns darüber hinaus, einfache Konfidenzintervalle für die gesuchte Wahrscheinlichkeit p festzulegen. Wenn wir als Konfidenzniveau etwa 99% wählen, wissen wir, dass in ca. 99% aller Fälle unser Konfidenzintervall I den gesuchten Wert p überdecken wird. Damit können wir I als eine approximative Messung von p betrachten, die uns also in ca. 99% der Fälle eine approximativ korrekte Messung bietet und nur in 1% eine grob fehlerbehaftete Messung. Dabei sollte wiederum beachtet werden, dass wir die relative Häufigkeit Z nur als einen Indikator für die gesuchte Wahrscheinlichkeit p ansehen und keinesfalls als definierend für p betrachten. Natürlich müssen wir immer Messfehler erlauben, die zu Abweichungen zwischen Z und p führen, die in Ausnahmefällen sogar so groß sein können, dass p eben nicht in I liegt. Doch das ist ganz normal für alle Messverfahren quantitativer Größen, wie wir oben bereits gesehen haben.

Unsere ergänzende *methodologische Regel* könnte also ungefähr folgendermaßen lauten: Wähle ein Konfidenzintervall I mit Hilfe von (BPP) um die gemessene relative Häufigkeit herum als approximative Messung von p und wähle dazu das Konfidenzniveau, das gemäß den jeweiligen Zwecken als angemessen erscheint.

Zur Illustration können wir uns ein konkretes Zahlenbeispiel ansehen: Nehmen wir an, wir werfen eine Münze 1000-mal und erhalten dabei 600-mal Kopf. Zum Konfidenzniveau von 99% erhalten wir dann mit Hilfe von (BPP) die folgende Ungleichung: $1–(1/4000\varepsilon^2) \geq 0{,}99$. Daraus können wir schließen: $\varepsilon \approx 0{,}16$, und erhalten also das Intervall I = [0,44;0,76]. Für

10000 Würfe mit 6000-mal Kopf würden wir das kleinere Intervall I = [0.55;0.65]) erhalten. So könnten wir unsere Messung weiter verbessern. Immerhin erhalten wir damit ein Intervall, das unseren gesuchten Messwert p in ca. 99% der Fälle enthält. Unsere Meta-Wahrscheinlichkeit $P_{meta}(\varphi)$ erfährt so eine Deutung als Unschärfemaß für unsere Messung von p und sagt uns insbesondere, wie häufig wir im Durchschnitt mit Ausreißern zu rechnen haben, d.h., in wie vielen Fällen unser Intervall den Wert p noch nicht enthält.

Das sieht ganz ähnlich aus wie für Messungen anderer recht »harmloser« (also empiristisch unverdächtiger) Größen wie etwa des Gewichts. Denken wir dazu wieder an die Messungen des Gewichts NB 10 durch das amerikanische Eichamt (s.o.). Auch diese streuen um den Nennwert von 10 Gramm und es treten dabei sogar gelegentlich Ausreißer auf, obwohl die Messungen höchsten Qualitätsanforderungen genügen. Oft nimmt man an, dass diese Messwerte ungefähr im Sinne einer Normalverteilung um den wahren Wert streuen, und kann dann etwa erwarten, dass in einem Intervall von drei Standardabweichungen um den Mittelwert der Messung herum der wahre Wert in ca. 99% der Fälle zu finden ist. Damit haben wir eine vernünftige Interpretation von objektiver physikalischer Wahrscheinlichkeit auch für den Einzelfall gefunden, die zumindest in den Situationen mit lokalem Einflussbereich durchaus sinnvoll erscheint und sich zumindest im Prinzip approximativ messen lässt.

Jedenfalls sind die Probleme dabei nicht unbedingt von ganz anderer Art oder größerem Ausmaß als für andere theoretische Größen. Trotzdem bleiben natürlich noch viele Fragen offen. Anthony Eagles (2004) hat 21 davon zusammengestellt, auf die ich in Bartelborth (2011) antworte. Da wir auf Propensitäten m.E. sowohl in der Quantenmechanik aber möglicherweise auch in anderen Bereichen angewiesen sind und selbst Bayesianer sich im Hauptprinzip darauf stützen, sollten wir jedenfalls nicht zu schnell die Flinte ins Korn werfen und aufgrund empiristischer Skrupel gleich ganz auf einen objektiven physikalischen Begriff von Wahrscheinlichkeit verzichten. Den Weg, den wir dabei beschreiten können, habe ich skizziert, so dass wir uns im Folgenden auf ein entsprechendes Konzept von Propensität stützen dürfen, ohne sogleich ein schlechtes Gewissen haben zu müssen.

5.5 Der klassische Bayesianismus

5.5.1 Grundlegende Verfahren

Klassische Bayesianer sind in erster Linie Probabilisten, die für alle Aussagen Wahrscheinlichkeiten vergeben. Diese Wahrscheinlichkeiten geben unseren Grad an, in dem wir an etwas glauben, man könnte sie als *Plausibilitätsgrade* bezeichnen, die darstellen, inwiefern unser übriges Hintergrundwissen die in Frage stehenden Aussagen plausibel erscheinen lässt. Wie man zu Beginn zu Glaubensgraden oder Wahrscheinlichkeiten gelangt, darüber sagt der klassische Bayesianismus aber nicht sehr viel. Das sind für ihn subjektive Einschätzungen, für die wir uns zwar nach Möglichkeit an objektiven Gegebenheiten orientieren sollen, die aber letztlich doch einen subjektiven Charakter haben.

Der Bayesianismus ist dann vor allem eine Theorie darüber, wie sich die Wahrscheinlichkeiten *ändern*, wenn wir neue Informationen erhalten. Insbesondere denken wir dabei daran, dass wir Theorien bzw. Hypothesen H betrachten und bestimmte Daten E erhalten, durch die die Theorien mehr oder weniger gestützt werden. Die Grundidee des Updatens durch Konditionalisierung hatten wir bereits kennengelernt. Man sehe sich dazu noch einmal das Beispiel des Dschungelfiebers zu Beginn des Kapitels an. Im Weiteren werden wir mit P^+ immer die neuen Wahrscheinlichkeiten nach dem Updaten von P bezeichnen:

(BU) Bayesianisches Updaten: Die Konditionalisierungsregel

$$P^+(H) = P(H|E) = P(H)\ [P(E|H)/P(E)]$$

Wenn wir also die neue Beobachtung E machen, wird die Wahrscheinlichkeit der Hypothese H damit »upgedatet«, aber zugleich wird auch jede andere unserer Überzeugungen mit Hilfe der Regel (BU) upgedatet. Wir gehen insgesamt von einer Wahrscheinlichkeitsverteilung P (bzw. dem Überzeugungssystem P) zu der neuen Wahrscheinlichkeitsverteilung P^+ (bzw. dem neuen Überzeugungssystem P^+) über. Daraus ergibt sich auch die neue Wahrscheinlichkeit von E zu: $P^+(E) = P(E|E) = 1$. Hier verletzt der klassische Bayesianismus also bereits das Dogmatismusverbot, indem zumindest die gemachten Beobachtungen ausgenommen werden. Sie

erhalten nach der Beobachtung direkt den Status sicheren Wissens. Wir können das als eine Art von Idealisierung betrachten, und außerdem gibt es jedoch noch die Jeffrey-Konditionalisierung, die andere Wahrscheinlichkeiten für das Datum E erlaubt, die wir später kennenlernen werden.

Im klassischen bayesianischen Rahmen nehmen wir jedenfalls erst einmal an, dass durch eine Beobachtung die entsprechende Beobachtungsaussage zu unserem sicheren Wissen gezählt werden darf und die Wahrscheinlichkeit 1 erhält. Das ist ein exogener Einfluss auf das Netz unserer Überzeugungen. Wir greifen eine Überzeugung heraus und setzen ihren Wert auf 1. Alle anderen Überzeugungen müssen dann entsprechend angepasst werden. Zunächst ist klar, dass $P^+(\neg E) = 0$ sein muss. Für alle Hypothesen T, aus denen ¬E logisch folgt, gilt das Entsprechende, dass sie ebenso auf 0 gesetzt werden müssen. Die Regel (BU) gibt also für alle anderen Fälle an, wie wir die neuen Wahrscheinlichkeiten berechnen müssen. Somit liefert schließlich P^+ die neue Wahrscheinlichkeitsverteilung auf unserem Netz von Überzeugungen. Das kann so fortgesetzt werden und wir erhalten durch weiteres Updaten mit weiteren Daten eine Wahrscheinlichkeit P^{++} usf. Dazu haben wir uns schon überlegt, dass die Reihenfolge, mit der die Daten eintreffen, und in der wir dann mit ihnen »updaten«, dabei keine Rolle spielt.

Außerdem gibt es einige andere sichtbare Effekte, auf die ich gleich hinweisen möchte. Sollte E sogar deduktiv aus H folgen, so ist $P(E|H) = 1$. Da $P(E) < 1$ gemäß dem Dogmatismusverbot gelten sollte, gilt dann also $P^+(H) > P(H)$, denn der Update-Faktor $P(E|H)/P(E)$ in (BU) wird größer als 1. Das heißt zugleich, H wurde durch das Auftreten des Datums E plausibler oder wir werden auch sagen, dass E unsere Hypothese H bestätigt hat. Das entspricht der hypothetisch-deduktiven Theorienbestätigung, die wir schon kennengelernt haben, und wird von Bayesianern gern als Bestätigung ihrer Konzeption betrachtet; aber wir erinnern uns, dass das hypothetisch-deduktive Bestätigungsverfahren mit einer Reihe von Problemen wie dem der irrelevanten Bestätigung belastet war, die sich damit auf den Bayesianismus übertragen können. Da der Bayesianismus die logischen Beziehungen respektiert, ist zunächst kein Platz vorhanden, um z.B. mehr als logische Ableitbarkeit der Daten zu verlangen, wie etwa

die *Erklärung* der Daten durch die Theorien, denn die beste Beziehung zwischen Theorie und Daten ist für den Bayesianer, dass $P(E|H) = 1$ ist.

Auch die Falsifikation finden wir hier wieder. Sollte $\neg E$ logisch aus H folgen, so wird $P(E|H) = 0$ und es gilt: $P^+(H) = 0$, wenn E auftritt; unsere Hypothese H wurde also endgültig falsifiziert. Das zeigt zugleich, dass unserer Verletzung des Dogmatismusverbots in Bezug auf E nun eine weitere in Bezug auf H nach sich zieht.

Wenn wir auf den Update-Faktor $P(E|H)/P(E)$ schauen, werden außerdem zwei durchaus erwünschte Phänomene deutlich: 1. Je stärker H das Datum E vorhersagt, um so stärker ist ceteris paribus auch die Bestätigung durch E, wenn E tatsächlich auftritt. 2. Allerdings spielt für die Stärke der Bestätigung ebenso die Vorher-Wahrscheinlichkeit von E eine wichtige Rolle. Ist E eine Tautologie (also $P(E) = 1$), so findet keine Bestätigung mehr statt. Ist E schon beinahe selbstverständlich, fällt die Bestätigung relativ schwach aus, und erst wenn E vorher sehr unwahrscheinlich war und trotzdem von der Theorie vorhergesagt wurde, ergibt sich eine stärkere Bestätigung durch E.

Gerade dieser zweite Effekt scheint sich in der Wissenschaftsgeschichte wiederzufinden. Die Vorhersage der einsteinschen allgemeinen Relativitätstheorie, dass eine deutliche Lichtablenkung im Schwerefeld der Sonne zu beobachten sein würde, stellte, als diese während einer Sonnenfinsternis tatsächlich beobachtet werden konnte, einen der stärksten Gründe für viele Zeitgenossen dar, die Theorie zu akzeptieren. Doch Wissenschaftler sehen den Punkt auch kritisch. Es mag zwar psychologisch plausibel sein, einer Theorie zu vertrauen, die uns mit so überraschenden und dann zutreffenden Vorhersagen versorgt, aber sollte das tatsächlich über die Bestätigungsstärke durch ein Datum entscheiden? Ist das Ausmaß der Überraschung nicht nur subjektiver oder zufälliger Natur? Hätten wir das Datum zufälligerweise schon früher kennengelernt, wäre der Überraschungseffekt nicht mehr gegeben. Würde dann unsere Theorie durch die Daten weniger gut bestätigt werden? Das scheint uns nicht besonders überzeugend zu sein. Von subjektiven Einschätzungen und historischen Zufälligkeiten sollte die Bestätigungsstärke möglichst nicht abhängen. Was auf den ersten Blick ein besonders intuitiver Aspekt des bayesianischen Konzept war, entpuppt sich auf den zweiten Blick

schon als ein mögliches Problem. Dem werden wir in Form des *Problems der alten Evidenz* (»old evidence«) wiederbegegnen.

Eine andere Frage ist, ob die Reihenfolge, in der neue Daten auftreten, einen Unterschied ausmacht. Das ist solange nicht der Fall, wie unser übriges Hintergrundwissen stabil bleibt, d.h., solange die Veränderungen der Überzeugungswahrscheinlichkeiten P nur durch die Daten veranlasst werden und nicht durch einen anderen Sinneswandel. Ein solcher Sinneswandel könnte etwa die Wahrscheinlichkeiten P(E) verändern und damit sogleich die Stärke der Bestätigung verändern. In der Wissenschaft finden wir einen solchen Sinneswandel etwa an den Stellen, an denen jemand eine *neue Theorie* ins Spiel bringt. Wenn wir sie ins Netz unserer Überzeugungen aufnehmen, müssen wir zwangsläufig die Wahrscheinlichkeiten anpassen, obwohl keine neuen Daten vorliegen müssen. Der Bayesianer muss daher davon ausgehen, dass alle möglichen Theorien, auf die wir jemals kommen werden, von Anfang an in der Menge L, auf der unser P definiert ist, enthalten sind. Das ist eine weitere deutliche Idealisierung.

Betrachten wir das Problem der Reihenfolge und des mehrfachen »Updatens« kurz noch einmal in einer etwas anderen Berechnungsweise, wobei wir statt »e&f« kürzer »ef« schreiben werden. Bestimmte Abkürzungen dieser Art sind für Berechnungen oft hilfreich und üblich. Wir nehmen eine Hypothese H und zwei Daten e und f, die nacheinander oder gleichzeitig auftreten. Zunächst das Updaten bei gleichzeitigem Auftreten:

(1) $P^*(H) = P(H|ef) = P(H)\ [P(ef|H) / P(ef)] =$
$P(H)\ [P(e|Hf)P(f|H) / P(e|f)P(f)] =$
$P(H)\ [P(f|He)P(e|H) / P(f|e)P(e)]$

In der zweiten Reihe finden sich die beiden Möglichkeiten, den oberen Term weiter aufzulösen. Wie sieht es nun aus, wenn wir erst mit e und dann mit f updaten?

(2) $P^+(H) = P(H|e) = P(H)\ [P(e|H) / P(e)]$ und außerdem ist
$P^+(f) = P(f|e) = P(f)\ [P(e|f) / P(e)]$ und
$P^+(f|H) = P(f|He) = P(f|H)\ [P(e|Hf) / P(e|H)]$

Wenn wir nun noch mit f updaten ergibt sich durch Einsetzen:

(3) $P^{++}(H) = P^{+}(H|f) = P^{+}(H)\ [P^{+}(f|H)\ /\ P^{+}(f)] =$
$P(H)\ [P(e|H)\ /\ P(e)]\ (P(f|H)\ [P(e|Hf)/P(e|H)]\ /$
$P(f)\ [P(e|f)/P(e)]) = P(H)\ [P(f|H)P(e|Hf)\ /\ P(f)P(e|f)]$

Diese kleine Rechenübung zeigt zum einen, wie mehrfaches Updaten funktioniert und dass wir in (3) wieder dasselbe Ergebnis erzielen wie in (1), wo wir gleichzeitig mit e und f upgedatet haben. Wegen der Symmetrie in e und f zeigt das sogleich, dass die Reihenfolge des Updatens für das Endergebnis ohne Bedeutung bleibt, wenn wir jedenfalls daran festhalten, dass Veränderungen der Wahrscheinlichkeiten nur durch die Regel (BU) erfolgen und keine anderen Formen von veränderten Einschätzungen auftreten.

5.5.2 Hintergrundwissen im Bayesianismus

Wir hatten schon früher gesehen, dass wir aus Daten nicht einfach Schlüsse auf bestimmte Theorien ziehen können, ohne unser Hintergrundwissen einzubringen. Das induktive Schließen stützt sich immer auf weiteres Hintergrundwissen. Das sollte sich ebenfalls in den bedingten Wahrscheinlichkeiten wiederfinden. Wenn Fritz plötzlich und während einer Grippewelle hohes Fieber bekommt, dann spricht das dafür, dass ihn eine Grippe erwischt hat. Man könnte also vermuten, dass in diesem Fall gilt: P(Fritz hat Grippe|Fritz hat Fieber) sehr hoch ist. Schon für diese Einschätzung benötigten wir Hintergrundwissen. Aber vielleicht wissen wir außerdem noch, dass Fritz gegen Grippe geimpft wurde. Dann wird P(Fritz hat Grippe|Fritz hat Fieber) womöglich wieder klein. Diese bedingten Wahrscheinlichkeiten hängen also von unserem konkreten Hintergrundwissen K ab. Das hat einige Autoren (etwa Hajek 2003) dazu gebracht, dieses Hintergrundwissen immer explizit aufzuführen und davon zu sprechen, dass es sich bei P immer um eine bedingte Wahrscheinlichkeit handelt. Statt P(E) müssen wir dann eigentlich P(E|K) betrachten. Dabei ist K eine Menge von Aussagen bzw. ihre Konjunktion.

Doch das ist nicht unproblematisch. Zunächst einmal müsste K die Wahrscheinlichkeit 1 erhalten: P(K) = P(K|K) = 1. Um welche Aussagen sollte es sich dabei handeln? Das widerspricht zumindest dem Dogmatismusverbot. Wenn unsere Wahrscheinlichkeitsfunktion regulär wäre, d.h., wenn aus P(K) = 1 bereits folgt, dass K eine Tautologie ist,

dann könnten wir gleich nachweisen, dass wir darauf verzichten dürfen, K explizit aufzuführen; denn für den Fall ist ableitbar: P(H|E&K) = P(H|E). Doch das sollte in gleicher Weise im Normalfall gelten, denn unser Hintergrundwissen ist schließlich bereits in unserer Überzeugungsfunktion P kodiert. Gehört unser Wissen um die Grippeimpfung von Fritz zu unserem Hintergrundwissen, so sollte sich das darin äußern, dass die Wahrscheinlichkeit P(Fritz hat Grippe|Fritz hat Fieber) klein wird. Dann müssen wir das Hintergrundwissen nicht unbedingt noch einmal explizit anführen.

Wir könnten das Hintergrundwissen K auch als Index aufführen: P_K, wenn es denn überhaupt noch eine klare Bedeutung innerhalb des probabilistischen Ansatzes hätte. Doch das ist eigentlich nicht der Fall. Wir haben nur P und keine Auszeichnung bestimmter Überzeugungen als Hintergrundwissen mehr. Insbesondere gerät eine solche Auszeichnung mit dem Dogmatismusverbot in Konflikt, das wir zwar schon gelockert hatten, das aber trotzdem noch für viele Elemente unseres Hintergrundwissens sinnvoll zu sein scheint. Dass Fritz sich gegen Grippe hat impfen lassen und diese ihn gut vor der Grippe schützt, mag zu meinem Hintergrundwissen gehören, doch deshalb muss es noch nicht so irrtumssicher sein, dass ich ihm Wahrscheinlichkeit 1 geben sollte. Genau genommen ist daher die Rede von Hintergrundwissen im Rahmen des subjektiven Bayesianismus nicht besonders sinnvoll. An manchen Stellen, wo das hilfreich erscheint, werde ich trotzdem wieder auf K als Hintergrundwissen zurückgreifen und es manchmal explizit einbauen, schon um zu kennzeichnen, ob in verschiedenen Situationen dasselbe oder unterschiedliches Hintergrundwissen vorliegt.

Vor allem dort, wo jemand wie Hawthorne (2010) versucht, auf objektivere Wahrscheinlichkeitsfunktionen (bei ihm konditionale Popper-Funktionen) zurückzugreifen, die nicht mehr nur den subjektiven Gehalt ganz bestimmter Überzeugungssysteme wiedergeben sollen, kann es wieder sinnvoller erscheinen, ein bestimmtes Hintergrundwissen explizit in die Wahrscheinlichkeitsfunktionen aufzunehmen. Sie werden gewissermaßen nicht als das Hintergrundwissen einer bestimmten Person einfach in der Wahrscheinlichkeit »versteckt«, sondern wieder offen als Bedingungen bestimmter Hintergrundannahmen aufgeführt. Wir werden später sehen, dass Hawthorne eine Behauptung wie »P(A|B) = r«

dann so liest: »In den möglichen B-Welten ist der Anteil der A-Welten gerade r.« Es handelt sich also schon eher um eine logische Beziehung zwischen Aussagen, für die wir dann natürlich alle beteiligten Aussagen offenlegen sollten. Die objektiven Bayesianer versuchen hier vor allem, die Ausgangswahrscheinlichkeiten objektiver zu gestalten.

5.5.3 Die Rechtfertigung der Konditionalisierungsregel

Wie im Falle der Wahrscheinlichkeitsaxiome können wir auch hier wiederum fragen, warum wir der Konditionalisierungsregel folgen sollten. In unserer jetzigen Schreibweise fragen wir also, warum wir als neue Wahrscheinlichkeit P^+ gemäß der folgenden Regel wählen sollten:

Bayesianische Konditionalisierung: $P^+(H) = P(H|E)$

Diese Vorschrift besagt anschaulich, dass wir einer bestimmten Theorie H, die wir gestern etwa noch für unwahrscheinlich hielten, heute die Wahrscheinlichkeit P(H) = 0,9 geben sollten, wenn zwei Dinge gelten: 1. Wir haben gestern gesagt, dass sie die Wahrscheinlichkeit 0,9 hätte, sollte E eintreten, und 2. E tatsächlich inzwischen eingetreten ist. So wie die Wahrscheinlichkeitsaxiome synchronische Konsistenzbedingungen darstellen, findet sich in der Konditionalisierungsregel eine diachronische Konsistenzbedingung. Meine heutigen Glaubensgrade sollten zu meinen gestrigen passen. Das erscheint zunächst plausibel und als eine Art von diachronischer Konsistenzforderung. Bayesianer stützen sich dazu außerdem auf ein Dutch-Book-Argument, um es als einzig rationale Verhaltensweise zu begründen.

Dazu ist zu klären, wie *bedingte* Glaubensgrade P(H|E) zu einem bestimmten Wettverhalten passen. Die entsprechende Wette auf H gegeben E soll dabei nur dann zur Auszahlung kommen, wenn E eintritt, ansonsten passiert einfach nichts. Nehmen wir nun an, jemand hat die folgenden Glaubensgrade: P(H|E) = 0,3 und $P^+(H) = 0{,}7$ (was durch Updaten mit E entsteht). Dann sollte er die folgenden Wetten mit einer Auszahlung von 10 Euro für den Fall akzeptieren, dass E eintritt:

(1) Wette 1 (heute): 7 Euro auf ¬A (falls E eintritt)
(2) Wette 2 (morgen nach dem Eintreffen von E): 7 Euro auf A

Im Falle von ¬E passiert hier nichts, denn die erste Wette wird annulliert und die zweite wird erst gar nicht abgeschlossen. Aber im Falle des Eintreffens von E werden beide Wetten abgeschlossen und auch aktiviert. Dann zahlt der Akteur 14 Euro ein und erhält nur 10 Euro ausgezahlt. Er kann also nur verlieren bei diesem Wettsystem. Man kann die Wetten noch durch geeignete Wetten auf E aufpeppen, so dass der Akteur in jedem Fall draufzahlt. Aber auch ohne diese Ergänzungen dürfte die Problematik von diachronisch nicht zusammenpassenden Glaubensgraden offenbar geworden sein.

Bayesianer verweisen darauf, dass wir, selbst wenn wir die Wetten etwa nicht als Paket akzeptieren wollen, wir doch bemerken müssten, dass etwas in unserem Überzeugungssystem nicht stimmt. Der Vorzug passender Glaubensgrade sollte für uns intuitiv erkennbar sein. Doch wir müssen auch an das Beispiel des einfachen Reflexionsprinzips denken, das keineswegs so plausibel war, aber trotzdem durch ein diachronisches Dutch-Book-Argument gerechtfertigt werden konnte. Es gibt natürlich daneben wieder andere axiomatische Ansätze, um die Konditionalisierungsregel zu begründen (vgl. etwa Earman 1992, 46 ff.), doch die Frage bleibt immer, wie plausibel uns die grundlegenden Forderungen in diesen anderen Ansätzen erscheinen und ob sie tatsächlich plausibler sind, als es die Konditionalisierungsregel selbst es schon ist.

Wir hatten aber bereits in Kapitel 5.3.14 gesehen, dass gerade für diachronische Zusammenhänge die Dutch-Book-Argumente nicht wirklich überzeugen können. Daher sind neben der intuitiven Einschätzung des Verfahrens andere Argumente erforderlich, um die Konditionalisierungsregel zu begründen. In Kapitel 5.3.5 hatten wir ein solches Argument kennengelernt. Die Konditionalisierungsregel ist die einzige Update-Regel, die die zu *erwartende Wahrheitsnähe* bzw. den *epistemischen Erwartungsnutzen* maximiert.

Außerdem werden wir gleich (in Kap. 5.5.5) im Zusammenhang mit einer allgemeineren Konditionalisierungsregel erfahren, dass sie noch eine andere ausgezeichnete Eigenschaft besitzt, nämlich sie ist die Update-Regel, die nur *minimale Änderung* im Lichte der Daten vornimmt. Es wird nur insoweit von den bisherigen Glaubensgraden abgewichen, wie die Daten das erzwingen, aber nicht in willkürlicher und unbegründbarer Weise darüber hinausgegangen. Das ist ein großer

Vorzug der Konditionalisierungsregel und entspricht auch den Verfahrensweisen, die man im Rahmen der klassischen Erkenntnistheorie in den »belief change«-Ansätzen verfolgt.

Williamson (2011) diskutiert allerdings unterschiedliche Fälle, in denen die Konditionalisierungsregel nicht mehr sinnvoll anwendbar ist und plädiert jeweils dafür, stattdessen im Sinne eines objektiven Bayesianismus mit Hilfe der Maximum-Entropie-Regel upzudaten. In vielen Fällen stimmen die Ergebnisse überein (nämlich wenn die neue Information eine einfache Aussage ist und wir unsere Sprache nicht erweitern müssen), aber wenn das nicht der Fall ist, wird uns die Konditionalisierung nach Williamson keine plausiblen Ergebnisse mehr liefern. Um das entscheiden zu können, wird man weiter solche Beispiele wie die in Williamson (2011) analysieren müssen (vgl. dazu a. Kap. 5.5.5).

5.5.4 Sind *vernünftige* Glaubensgrade messbar?

Manchmal nehmen Bayesianer für sich in Anspruch, dass ihre Grundgrößen – die Glaubensgrade – messbar sind, und betrachten das als Vorteil ihres Ansatzes. Wir hatten bereits gesehen, dass die Verbindung der Glaubensgrade zu einem bestimmten Wettverhalten keineswegs so eng ist, wie das Probabilisten gerne darstellen, weshalb wir dafür eintraten, die Glaubensgrade eher als theoretische Größen einzuschätzen. Nun kommt noch ein weiteres Problem hinzu. Psychologen wie Kahneman und Tversky (z.B. in 1982) haben uns gelehrt, dass Menschen ihre Glaubensgrade normalerweise nicht von selbst nach den Wahrscheinlichkeitsaxiomen richten, sondern dazu eine spezielle Ausbildung benötigen. In Fällen wie dem Dschungelfieberbeispiel wird selbst von Ärzten und anderen Fachleuten, aber noch mehr von Laien, systematisch der Effekt des Tests auf die neue Wahrscheinlichkeit überbewertet gegenüber dem Effekt der geringen Prävalenz der Krankheit in der Grundgesamtheit. Man nennt das den »base-rate-fallacy«, weil man eben diese Basisrate vergisst. Es kommen viele andere systematische Fehler dieser Art hinzu. Daher werden schon die Ausgangsglaubensgrade nicht kohärent bestimmt sein, aber weitere Abweichungen von den Wahrscheinlichkeitsaxiomen entstehen danach beim Updaten. Selbst wenn wir also die *tatsächlichen* Glaubensgrade messen könnten, gilt das

nicht mehr für die *vernünftigen* Glaubensgrade, die korrigiert gegenüber den realen Glaubensgraden den Wahrscheinlichkeitsaxiomen gehorchen.

Überhaupt können wir die vernünftigen Glaubensgrade (die sich nach den Wahrscheinlichkeitsaxiomen und der Konditionalisierungsregel richten) eigentlich keiner realen Person mehr zuschreiben. Nehmen wir an, wir hätten eine reale Person s mit den Glaubensgraden P_s, die aber bestimmte Inkohärenzen beinhalten. Die wären vermutlich etwas versteckt, aber wir nehmen der Einfachheit halber an, dass schlicht und einfach $P_s(A) = 0{,}6$ und $P_s(\neg A) = 0{,}7$ sei. Wie sähen denn dann die korrigierten vernünftigen Glaubensgrade von s aus? Für dieses Paar von Überzeugungen könnten wir etwa die Paare [0,6;0,4] oder [0,3;0,7] oder [0,46;0,54] oder ganz andere Werte wählen und müssten jeweils weitere Anpassungen im restlichen Überzeugungssystem vornehmen. Tatsächlich führt kein vernünftiger Weg zu nur einer dieser speziellen Lösungen. Wir erhalten also keine eindeutige Zuweisung von *vernünftigen* Glaubensgraden zu s mehr in einem solchen Fall.

Wir könnten stattdessen natürlich gleich eine Menge von Glaubensgradfunktionen betrachten, die alle mit den unproblematischen Überzeugungen von s übereinstimmen und an der Stelle A alle kohärenten Möglichkeiten erlauben. Doch wir erhalten so schnell eine große Menge von Überzeugungssystemen und die Frage bleibt, ob uns das tatsächlich voranbringt.

Wir erfinden also eine fiktive Person t und schreiben ihr vernünftige Glaubensgrade also Wahrscheinlichkeiten P_t zu, die vielleicht noch gewisse Ähnlichkeiten zu den Glaubensgraden von s aufweisen. Eigentlich sind wir damit bereits einen kleinen Schritt hin zu einer induktiven Logik gegangen, nämlich weg von realen Glaubensgraden hin zu vernünftigen Glaubensgraden, die sich an bestimmte epistemisch rationale Spielregeln halten. Je mehr solcher Spielregeln wir jeweils einführen, um so mehr nähern wir uns einer modernen induktiven Logik (im Sinne von Hawthorne), die keineswegs verlangt, dass ihre Regeln so stark sein müssen, dass sie bereits alle Wahrscheinlichkeiten eindeutig festlegen. Man sollte dann wohl eher von *Graden der Bestätigung* als von Glaubensgraden sprechen.

Wir werden später noch sehen, dass jeder dieser Schritte (der Einführung objektiver Regeln) einen erkenntnistheoretischen Fortschritt

bedeutet. Ohne solche Regeln liefert der Bayesianismus weder eine Erkenntnistheorie noch eine Methodologie, sondern bestenfalls ein Buchhaltungsverfahren, das unsere anderweitig bestimmten Glaubensgrade notiert. Daher müssen wir diese Spielregeln so genau betrachten und sind darauf zwingend angewiesen. Zumindest gewinnen wir hier jedoch die Erkenntnis, dass die vernünftigen Glaubensgrade nicht messbar oder anderweitig operationalisierbar sind, weil kein direkter Weg von aktualen zu vernünftigen Glaubensgraden führt.

5.5.5 Jeffreys Konditionalisierungsregel und minimale Änderungen

Ein Problem des klassischen Bayesianismus ist die Verletzung des Dogmatismusverbots, wonach den Daten, die wir beobachtet haben, die Wahrscheinlichkeit 1 zugewiesen wird. Das scheint auch erkenntnistheoretisch keineswegs so unproblematisch zu sein, denn natürlich müssen wir zugestehen, dass wir uns bei unseren Beobachtungen zumindest manchmal irren können, und daher die neue Wahrscheinlichkeit für eine Beobachtungsaussage E zumindest etwas kleiner als 1 sein sollte. Außerdem haben wir es in der Wissenschaft zum einen mit Rohdaten in Form von einfachen Beobachtungen zu tun, doch die sind noch nicht mit unseren Theorien vergleichbar, weshalb wir diese Rohdaten interpretieren müssen, um sie dann mit unseren Theorien konfrontieren zu können, wodurch weitere Irrtumsrisiken ins Spiel kommen.

Wollen wir unsere Theorie der Elektrodynamik testen und müssen dazu eine Stromstärke mit Hilfe eines Amperemeters messen, so ist etwa ein Rohdatum, dass das Anzeigegerät auf der 3 steht. Doch das sagt zunächst noch nichts über die Elektrodynamik. Erst wenn wir dieses Datum interpretieren als: »Es fließt ein Strom von 3 Milliampere in unserem Leiter«, sagt es etwas über elektromagnetische Phänomene und lässt sich direkt mit den Vorhersagen unserer Theorie vergleichen. Doch spätestens der Schritt vom Rohdatum zum interpretierten Datum ist fehleranfällig. Er muss sich auf eine Theorie stützen (in diesem Fall pikanterweise sogar u.a. auf die Elektrodynamik selbst), und die könnte zumindest fehlerhaft sein. Außerdem könnte das Messgerät defekt sein und so zu Messfehlern führen. Also ist es nur vernünftig, wenn wir das

in unserer Modellierung des Vorgangs auch berücksichtigen, und so das interpretierte Datum eine Wahrscheinlichkeit kleiner 1 aufweist.

Natürlich kann man auch versuchen, den ganzen Prozess mit allen beteiligten Hilfsannahmen bayesianisch zu rekonstruieren, doch das würde den Einsatz des Bayesianismus erheblich erschweren und ließe uns nicht direkt die spannenden Zusammenhänge zwischen Theorien und Daten diskutieren (vgl. auch Bovens & Hartmann 2006). Wir werden darauf später noch zurückkommen. Also sollten wir die schon *interpretierten Daten* als Daten zum Updaten von Theorien zulassen, dann müssen wir umso mehr einräumen, dass es für sie eine Fehlermöglichkeit gibt, denn natürlich könnten unsere Messverfahren fehlerhaft sein.

Das scheint auch bestimmte Fälle von Wahrnehmungen e besser zu beschreiben. Wenn wir das Ergebnis der Wahrnehmung e nicht durch einen einfachen Satz darstellen können, sondern nur durch eine Veränderung unserer Wahrscheinlichkeiten für unsere Überzeugungen. Wir glauben z.B. eine bekannte Person an einem bestimmten Ort gesehen zu haben (E). Wir haben sie aber nur kurz und nur von hinten gesehen. Dann sollten wir nun nicht gleich mit E klassisch Updaten, sondern die Wahrscheinlichkeit von E nur leicht anheben, etwa von 0,2 auf 0,7.

Um diesen Fall von nicht-irrtumssicheren Daten korrekt beschreiben zu können, kann der Bayesianer die Konditionalisierungsregel von Jeffrey verwenden. Hier geht man nicht mehr davon aus, dass die Wahrscheinlichkeit einer Aussage auf 1 wächst, sondern nur, dass sie sich verändert und wir dann die anderen Wahrscheinlichkeiten anpassen müssen. Auch Jeffrey selbst motivierte seine Regel unter Hinweis auf Beobachtungen, bei denen wir uns nicht sicher sind – etwa weil die Lichtverhältnisse suboptimal sind. Trotzdem sollten wir die daraus resultierenden Informationen nicht verlieren, sondern weiterverarbeiten. Dazu benötigen wir eine Regel, die keine Sicherheit der Daten voraussetzt. Nehmen wir also an, dass wir durch Beobachtungen die Wahrscheinlichkeit einer Aussage E verändern zu $P^+(E)$. Dann ändern sich die Wahrscheinlichkeiten der anderen Aussagen gemäß der:

(1) Jeffrey Konditionalisierung:

$$P^+(H) = P(H|E)\cdot P^+(E) + P(H|\neg E)\cdot P^+(\neg E)$$

Das führt zur klassischen Konditionalisierungsregel für den Spezialfall $P^+(E) = 1$ und damit $P^+(\neg E) = 0$. Es können sogar Fälle auftreten, in denen wir unsere Wahrscheinlichkeit auf eine vollständige Zerlegung $\{E_1,\ldots,E_n\}$ aufteilen, die wir als unsere *Evidenz-* oder *Beobachtungsbasis* bezeichnen wollen. (Bisher haben wir nur die einfache Zerlegung $\{E,\neg E\}$ betrachtet.) Nehmen wir an, wir würden nicht genau erkennen, welche Farbe ein Objekt hat (aufgrund sehr schlechter Beleuchtung), und es würden nur die folgenden drei Farben in Frage kommen: $P^+(\text{blau}) = 0{,}2$ und $P^+(\text{grün}) = 0{,}7$ und $P^+(\text{rot}) = 0{,}1$. Dann könnten wir auch diese neue Verteilung auf unsere Datenaussagen E_i zum Updaten heranziehen und erhielten:

(2) Jeffrey Konditionalisierung: $P^+(H) = \sum_i P(H|E_i)\cdot P^+(E_i)$

Obwohl diese Konditionalisierungsregel so plausibel erscheint, wird sie von klassischen Bayesianern eher abgelehnt. Eine ganze Reihe von Kritikpunkten finden wir bei Weisberg (2009). Ein zentraler Einwand ist, dass es bei dieser Regel auf die Reihenfolge der Daten ankommen kann, in der mit ihnen upgedatet wird. Es kann also passieren, dass zwei Bayesianer mit denselben Startwahrscheinlichkeiten zwei gleiche Beobachtungen machen, die auch entsprechend in neue Wahrscheinlichkeiten umsetzen, aber trotzdem zu anderen Nachher-Wahrscheinlichkeiten gelangen, weil sie ihre Beobachtungen in einer anderen Reihenfolge gemacht haben. Die formalen Eigenschaften haben schon Diaconis & Zabell (1982) untersucht. Wenn wir etwa zweimal bezüglich derselben Zerlegung updaten, so bestimmt jeweils das zweite Updaten die Werte $P(E_i)$, während das erste damit für unsere Datenaussagen E_i wieder »verlorengeht«.

Somit scheint ein eigentlich kontingentes Faktum, nämlich die Reihenfolge, in der wir die Daten erhalten, einen zu großen Einfluss auf das Ergebnis zu haben. Es gibt allerdings auch Verteidiger dieses Aspekts der Jeffrey Konditionalisierung. Lange (2000) zeigt an speziellen Beispielen, dass die Reihenfolge durchaus einmal relevant sein kann. Die jeweils spätere Information kann z.B. die frühere dabei ein Stück weit diskreditieren. Doch das scheinen recht spezielle Fälle zu sein. Diaconis & Zabell (1982) geben zu der *fehlenden Kommutativität* weitere Beispiele an und diskutieren unterschiedliche Bedingungen, unter denen die Kommutativität doch wieder hergestellt wird. Auf die werden wir im nächsten Kapitel eingehen.

Allerdings benötigen wir jeweils eine Zerlegung $\{E_1,\ldots,E_n\}$, um die Regel anwenden zu können, was eine gewisse Willkür in der Anwendung bedeutet, und bei mehrfachem Updaten können sich die Zerlegungen wiederum unterscheiden, wodurch die Problematik sich verschärft.

Die Jeffrey-Konditionalisierung besitzt allerdings einige schöne Eigenschaften, die wir für eine echte *Wahrscheinlichkeitskinematik* erwarten. Wenn etwa für alle i die Wahrscheinlichkeit $P(E_i) > 0$ ist (was man meisten als sinnvolle Forderung einbringt), so ergibt ein erneutes Updaten mit den alten Wahrscheinlichkeiten für die E_i wieder die ursprüngliche Verteilung. Insbesondere ist sie dadurch bereits ausgezeichnet und festgelegt, dass für alle Aussagen A und alle i gilt: $P^+(A|E_i) = P(A|E_i)$. Das finden etwa Howson & Franklin (1992) besonders bedeutsam, da es ihrer Meinung nach geradezu zur Logik des Updatens gehört. Diese Form der *Rigiditätsforderung* (die wir bereits kurz beschrieben haben) sorgt dafür, dass das Updaten für andere Aussagen nur über die Veränderung der Wahrscheinlichkeiten für die E_i geschieht. Liegen die fest, werden die Wahrscheinlichkeiten für die anderen Aussagen vorher und nachher dadurch gegeneinander abgeschirmt. Es gibt auch eine dazu äquivalente Darstellung, die sich u.a. bei James Joyce (2009) findet, die er als Prinzip der minimalen Änderung (MC_2) bezeichnet:

> **Minimale Änderung**: Für alle Aussagen $A\&E_i$ und $B\&E_i$, die aus den E_i folgen (für alle i), bleibt das Verhältnis der Wahrscheinlichkeiten beim Updaten konstant: $P^+(A\&E_i)/P^+(B\&E_i) = P(A\&E_i)/P(B\&E_i)$

Das bedeutet, dass die Veränderungen der Wahrscheinlichkeiten beim Updaten jeweils anteilig auf die anderen Aussagen verteilt werden. In das Verhältnis dieser anderen Aussagen zueinander wird nicht weiter eingegriffen. Tatsächlich führt das Jeffrey-Updaten, das aus der Rigiditätsregel folgt, zu einer nur minimalen Veränderung der Wahrscheinlichkeiten beim Updaten im Hinblick auf unterschiedliche Abstandsmaße. Das überträgt sich dann natürlich auch auf den Grenzfall der klassischen Konditionalisierungsregel.

Zwar wurden auch für Jeffreys Regel wieder Dutch-Book-Argumente angeführt, aber spannender ist hier ein anderer Begründungsansatz, der das *Prinzip der minimalen Änderung* ausnutzt. Man kann argumentieren,

dass das Jeffrey-Updaten eine Lösung für die folgende Vorgehensweise liefert: Suche zu einer Wahrscheinlichkeitsverteilung P und einigen neuen Werten auf einer Zerlegung E_i die zu P *nächstgelegene* Wahrscheinlichkeitsfunktion P*, die auf E_i diese neuen Werte annimmt. Die Idee ist eine klassische Idee aus dem Repertoire der Verfahren zur Überzeugungsänderung. Man nehme nur so viele Abänderungen vor, wie tatsächlich durch die neuen Informationen erzwungen werden, aber nicht mehr. Im Übrigen bleiben wir konservativ und so nahe wie möglich an den bisherigen Überzeugungen. Das kann man wiederum als eine Anwendung eines Prinzips der Unvoreingenommenheit betrachten. Im Vergleich zweier Aussagen A und B verändern wir ihre Wahrscheinlichkeiten nur insoweit, wie es die neuen Daten verlangen, aber gehen nicht zugunsten der einen Aussage darüber hinaus. Würden wir dagegen A gegenüber B beim Updaten noch weitergehend bevorzugen, wäre das wiederum ein Fall von Voreingenommenheit.

Es scheint zumindest auch aus praktischen Erwägungen heraus eine sinnvolle Vorgehensweise zu sein. Wir müssen nicht eine komplette Neubewertung vornehmen und haben nicht fortlaufend weitere radikale Neubewertungen zu gegenwärtigen. Wie gut die erkenntnistheoretische Begründung ist, ist schon nicht ganz so einfach zu entscheiden.

Sie hat jedenfalls bestimmte erkenntnistheoretische Intuitionen auf ihrer Seite, was ich noch an einem konkreten Beispiel erläutern möchte. Nehmen wir an, Kommissar X hat für einen Mord genau drei Verdächtige Franz, Kevin und Joe. Alle drei sind bisher gleichverdächtig, so dass Kommissar X jedem die Wahrscheinlichkeit 1/3 gibt, der Täter zu sein. Nun erhält er die absolut zuverlässige Information, dass der Täter blond ist und nur die ersten beiden Verdächtigen sind blond. Damit scheidet Joe aus: P(Joe)=0. Wie verteilen wir die freigewordene Wahrscheinlichkeit auf die beiden anderen Verdächtigen? Wenn Kommissar X hier vom klassischen Updaten abweichen würde mit P(Franz)=0,6 und P(Kevin)=0,4, so würde er ohne nachvollziehbaren Grund nun Franz als Täter bevorzugen. Da die Information aber in gleicher Weise für die beiden ersten Verdächtigen spricht, wäre das ein Fall von epistemischer Voreingenommenheit, der uns irrational erschiene. Auf das Verhältnis der Verdächtigkeit von Franz und Kevin sollte die Information keinen Einfluss haben, so wie das Beispiel gestaltet ist.

Das sollte auch für unsichere Informationen gelten. Nehmen wir an, Kommissar X hat für einen anderen Mord wieder genau drei gleichermaßen Verdächtige Franz, Kevin und Ute. Ute hat sich aus Sicht von X irgendwie verdächtig verhalten und erhöht daher die Wahrscheinlichkeit für die Annahme (H_3), dass sie die Täterin war, auf 0,5, während die anderen beiden Hypothesen dann noch 0,25 erhalten. Weiterhin folge anhand unseres Hintergrundwissens, dass ihr dann ihr Sohn Uwe ziemlich sicher geholfen haben muss, wenn sie die Tat begangen hat. Dann sollte die Wahrscheinlichkeit für die Mittäterschaft von Uwe nun auch entsprechend anwachsen, aber doch nicht unverhältnismäßig stark, sondern gemäß der ursprünglichen Relation zu den Wahrscheinlichkeiten für die Folgen aus den anderen Annahmen H_1 und H_2. Wenn Kommissar X einfach schlösse, damit sei nun zu 90% sicher, dass Uwe ein Mittäter war, würde er offensichtlich etwas falsch machen und gegen eine Art von epistemischem Fairnessgebot verstoßen. Er sollte die Wahrscheinlichkeit für die Mittäterschaft von Uwe nur soweit erhöhen, wie es die Daten verlangen, also der Konzeption folgen, dass man wieder eine Wahrscheinlichkeitsfunktion erhält, die die neuen Werte aufgreift und ansonsten nur minimale Änderungen vornimmt.

Williamson (2010) und Landes & Williamson (2013) treten sogar ganz allgemein dafür ein, dass wir immer nach diesem Verfahren Updaten sollten. Demnach sollten wir zunächst aus der Menge aller Wahrscheinlichkeitsfunktionen die Teilmenge der Funktionen auswählen, die noch zu unseren Daten und unserem Hintergrundwissen passen. Das nennt er *Kalibrierung*. Dazu sollten wir dann den konvexen Abschluss bilden und daraus die Glaubensfunktion auswählen, die sich am nächsten an der Gleichverteilung $P_=$ befindet. Das bezeichnet er als *Äquivokation*. Es gibt eine ganze Reihe von Argumenten für diese Form eines objektiven Bayesianismus, aber natürlich auch einige Probleme.

Positiv ist zunächst einmal, dass das Verfahren nicht auf das Updaten mit einfachen Aussagen angewiesen ist, sondern auch für andere Informationen offen bleibt. Andererseits hängt es dann wesentlich davon ab, mit welcher Abstandsfunktion man arbeiten möchte. Williamson wählt den Kullback-Leibler Abstand (s.u.) bzw. tritt einfach für eine *Entropiemaximierung* ein, was in diesem Fall auf dasselbe hinausläuft. Er geht dabei davon aus, dass wir im Falle, dass überhaupt keine

Informationen vorliegen, mit der Gleichverteilung arbeiten. Dem wollen allerdings viele Bayesianer nicht folgen. Obwohl sie die Forderung nach minimalen Änderungen beim Updaten für gut begründet erachten, sind sie nicht bereit, auch diesen Grenzfall keiner vorliegenden Daten entsprechend mit einem Indifferenzprinzip zu lösen. Diese Unterscheidung zu treffen, zwischen den Fällen mit bestimmten Informationen und dann auf möglichst minimale Änderungen zu setzen, und denen ohne Informationen und für diesen Fall keine Regel zu akzeptieren, passt m.E. jedoch nicht wirklich zusammen.

Für das Jeffrey-Updaten gilt jedenfalls noch, dass es zu minimalen Änderungen für eine größere Bandbreite von Abstandsmaßen führt. Diaconis und Zabell (1982) zeigen, dass Jeffreys Regel insbesondere den Kullback-Leibler-Abstand I(P*,P) (auch genannt: relative Entropie von P* bzgl. P) minimiert. Dieses Abstandsmaß ist zwar nicht symmetrisch, hat aber eine ganze Reihe wünschenswerter Eigenschaften für ein Abstandsmaß. So ist es z.B. positiv definit (vgl. Kullback, 1968). Es wird daher in vielen Kontexten eingesetzt. Bei Weisberg (2009) wird es wie bei Williamson (2010) innerhalb der Regel *Infomin* verwendet, die eine sehr allgemeine Regel zum Updaten von Wahrscheinlichkeiten darstellt und für die Jeffreys Regel eindeutige Ergebnisse dort liefert, wo wir die Regel von Jeffrey einsetzen können.

> **Infomin:** Wenn unsere Ausgangswahrscheinlichkeit durch P gegeben ist und wir nun die Zusatzinformation erhalten, dass nur noch Wahrscheinlichkeitsverteilungen P* aus einer bestimmten Menge S von Verteilungen zugelassen sind, dann wähle die Funktion $P^* \in S$, für die die folgende Größe minimal wird auf der Menge A_i aller Vollkonjunktionen aus L:
>
> $I(P^*,P) = \sum_i P(A_i) \log[P(A_i)/P^*(A_i)]$

Man sagt auch: Die Funktion I liefert den Erwartungswert der relativen Information von P zu P*, da der Informationsgehalt eines Ergebnisses A_i gerade $\log(1/P(A_i))$ ist. Eine Minimierung von I(P*,P) durch eine Wahl von P* bedeutet demnach, dass wir die Verteilung P* in unserer Menge S suchen, die *aus Sicht von P* am wenigsten Informationen aufzuweisen hat.

Das ist eine ganz ähnliche Idee wie im schon erwähnten Maximum-Entropie-Prinzip. Man könnte auch hier sagen, dass die Idee des Prinzips vom zureichenden Grunde dahinter steht. Wähle nur die Abweichungen vom bisherigen Überzeugungssystem P, die tatsächlich durch unsere neuen Informationen erzwungen werden und keine darüber hinaus gehenden Abweichungen. Dann hängt natürlich Vieles davon ab, dass unser Abstandsmaß intuitiv das leistet, was wir uns von ihm versprechen.

Doch für die Jeffrey-Regel sind wir nicht allein auf Infomin angewiesen, denn auch für andere Abstandsfunktionen liefert die Jeffrey-Regel ein minimales Ergebnis; allerdings muss dieses nicht unbedingt eindeutig bestimmt sein. Diaconis & Zabell (1982) und Joyce (2009a) geben noch andere Abstandsmaße an, für die das ebenfalls gilt. Dazu betrachten wir alle Vollkonjunktionen ω unserer Sprache und dann die folgenden Abstandsmaße:

– Supremumsnorm:	$S(P,P^*) = \sup_\omega \lvert P(\omega) - P^*(\omega)\rvert$
– Brier-Score:	$B(P,P^*) = \sum_\omega (P(\omega) - P^*(\omega))^2$
– Hellinger Abstand:	$H(P,P^*) = \sum_\omega P(\omega)+P^*(\omega) - (P(\omega)\cdot P^*(\omega))^{½}$
– Kullback-Leibler Abstand:	$I(P,P^*) = \sum_\omega P^*(\omega) \log(P^*(\omega)/P(\omega))$

Für alle diese Maße ergibt die Jeffrey-Regel die nächstgelegene Wahrscheinlichkeitsverteilung P*, die auf der Beobachtungsbasis gerade die upgedateten Werte liefert.

Diese Form der minimalen Änderung wird eigentlich von allen Bayesianern als eindeutiger Vorteil der Jeffrey-Regel angesehen. Man sollte beim Updaten nur die Daten auswerten, aber nicht darüber hinausgehen. Das wird als sinnvolles epistemisches Prinzip betrachtet nur erstaunlicherweise nicht mehr für den Grenzfall überhaupt keiner Daten, für den der objektive Bayesianismus die entsprechende Regel ebenso anwenden möchte. Wir sollten demnach auch dort nicht über die Daten hinausgehen, und solange keine Daten vorliegen sollten wir nach dem Prinzip MaxEnt (oder einem ähnlichen Prinzip) verfahren und alle möglichen Zustände als gleichberechtigt ansehen und ihnen dieselbe Wahrscheinlichkeit zuweisen. Doch darauf kommen wir in Kapitel 5.7 noch zu sprechen.

Die Jeffrey-Regel bietet also eine gut begründete Ergänzung des bayesianischen Ansatzes, die bestimmte Idealisierungen des Ansatzes aufhebt und damit neue Einsichten ermöglichen kann. Sie ist noch dynamischer als die einfache Update-Regel, weil sie es gestattet, auch die Datenaussagen wieder zu revidieren.

Außerdem haben wir zugleich eine weitere Rechtfertigung für die klassische Konditionalisierungsregel erhalten. In den einfachen Fällen, in denen sie anwendbar ist (in denen also eine Beobachtungsaussage auf Wahrscheinlichkeit eins gesetzt wird), liefert sie eine neue Glaubensgradfunktion, die nur durch minimale Änderungen aus der alten hervorgegangen ist. Diese konservative Strategie erscheint als der sinnvollste Weg, neue Daten in unser Überzeugungssystem einzubringen und entspricht auch dem Vorgehen der *Belief-change-Ansätze* im Falle der klassischen Erkenntnistheorie. Diese Vorgehensweise stellt eine Erweiterung des epistemischen Gleichbehandlungsprinzips dar, bei dem keine Aussage gegenüber anderen bevorzugt werden soll, solange keine Daten vorliegen, die das erzwingen.

Das Judy-Benjamin-Beispiel. Eine spezielle Frage ist aber, ob die Jeffrey-Regel schon alle Fälle von Informationsgewinn abdeckt. Van Fraassen (1981) hat dazu als Problemfall das Judy-Benjamin-Beispiel entwickelt, und ähnliche Beispiele wurden von etwa von Richard Bradley (2005) vorgeschlagen (s. Joyce 2009a, Weisberg 2009). Judy Benjamin sprang mit dem Fallschirm aus einem Flugzeug in einen bestimmten Bereich, der in vier gleiche Sektoren (Quadranten) nach Himmelsrichtungen aufgeteilt wurde: NW, NO, SW und SO. Judy glaubt nun zunächst, dass jeder der Sektoren dieselbe Chance hat, getroffen zu werden: Für jeden Quadranten haben wir die Wahrscheinlichkeit 1/4. Sie schildert nach der Landung über Funk ihre Umgebung und erhält dann die Information, dass man zwar nicht wisse, ob sie im Norden oder Süden gelandet sei, aber wenn sie im Norden gelandet sei, dann wären ihre Chancen 3:1, dass sie im Osten sei. Wie soll Judy diese Informationen nun verarbeiten?

Das Problem ist hier, dass die Information, die sie erhalten hat, nicht wie für die Konditionalisierungsregel vorgesehen eine unbedingte Wahrscheinlichkeit liefert, sondern nur eine *bedingte*, nämlich dass $P(NO|N) = 3/4$ ist. Um solche Fälle einer Veränderung bei den bedingten

Wahrscheinlichkeiten ebenfalls behandeln zu können, können wir nach noch allgemeineren Regeln des Updatens Ausschau halten oder wir können versuchen, es auf die bisherigen Regeln zurückzuführen, indem wir die Aussage über Funk selbst mit einer bestimmten Wahrscheinlichkeit wahr zu sein versehen und dann damit updaten (vgl. Grove & Halpern 1997). Den Vorschlag werden wir hier aber nicht weiter verfolgen.

Auch die Regel, die Wahrscheinlichkeitsfunktion zu wählen, die die neuen Informationen berücksichtigt und ansonsten so nahe wie möglich an der ursprünglichen Glaubensfunktion liegt, führt hier zumindest nicht in einfacher Weise zum Ziel. Joyce (2009a) diskutiert die dabei auftretenden unterschiedlichen Lösungen, je nachdem, welche der Abstandsmaße wir einsetzen. Joyce spricht sich selbst für den sogenannten Brier-Score aus, der aus der Beurteilung von Wettervorhersagen stammt, aber heutzutage ebenso in der theoretischen Statistik zum Einsatz kommt. Joyce setzt ihn u.a. dazu ein, die Wahrscheinlichkeitsaxiome für Glaubensgrade zu begründen. Williamson (2010) und Landes & Williamson (2013) setzen dagegen auf den Kullback-Leibler-Abstand, der in der Informationstheorie sehr beliebt ist und ebenfalls in der theoretischen Statistik zum Einsatz kommt. Sie begründen damit auch die Gültigkeit der Wahrscheinlichkeitsaxiome als Rationalitätsforderungen für unsere Glaubensgrade. Für das Judy-Benjamin-Beispiel führen die verschiedenen Maße allerdings zu unterschiedlichen Lösungen (vgl. Joyce 2009a, 454 ff.), und wir können bisher nicht sagen, dass eine davon die eindeutig beste wäre.

Howson & Franklin (1992) schlagen für Fälle unsicherer Daten einen ganz anderen Weg im Rahmen des klassischen Bayesianismus vor, nämlich eine weitere Aussage einzuführen, die etwas über die Zuverlässigkeit eines Zeugen oder Messgerätes oder einer anderen Informationsquelle behauptet und sogar selbst wieder mit einer gewissen Wahrscheinlichkeit versehen wird, die unsere Unsicherheit wiedergibt; so wie wir das in Kapitel 5.6.8 kennenlernen werden.

5.5.6 Wahrscheinlichkeitskinematik

Die Update-Regel von Jeffrey führt in vielen Fällen dazu, dass die Reihenfolge, in der Daten bzw. neue Informationen aufgenommen werden,

über das Endresultat entscheidet. Das liegt daran, dass die Update-Informationen auf eine bestimmte Weise eingebracht werden, nämlich durch das direkte Festsetzen der $P^+(E_i)$ für eine Beobachtungsbasis E_i (die eine disjunkte Zerlegung bilden): $P^+(E_i) = r_i$. Es gibt dazu alternative Verfahren, die Informationen z.B. aus unserer Wahrnehmung auszuwerten. Die haben u.a. Field (1978), Diaconis und Zabel (1982), Hawthorne (2004) und Joyce (2009a) weiter untersucht. Folgen möchte ich vor allem den Überlegungen von Hawthorne (2004), da wir hier die m.E. ausführlichste und plausibelste Erörterung zu diesem Problem finden. Beginnen wir zunächst mit dem einfachen Updaten P_e mit einer Information e und der Beobachtungsbasis E_i. Die Rigiditätsbedingung sorgt dafür, dass nun die Glaubensgrade für alle anderen Aussagen S mit Hilfe der neuen Werte für die E_i bestimmt werden:

$P_e(S|E_i) = P(S|E_i)$ (*einfache Rigidität*)

Damit erhalten wir für beliebige Aussagen S in einem ersten Schritt:

$$P_e(S) = \sum_i P(S|E_i)\cdot P_e(E_i) = P(S\&E_i)\cdot(P_e(E_i)/P(E_i)),$$

wobei in der letzten Summe nur über die i summiert werden darf, für die $P(E_i)$ echt größer als 0 sind. Um die Formeln übersichtlicher zu gestalten und im Einklang mit dem Dogmatismusverbot, werde ich das auch in den nächsten Gleichungen immer für alle entsprechenden Wahrscheinlichkeiten für die Basen annehmen. Sonst müssen wir die Summation einfach auf die anderen Summanden beschränken.

Wenn wir nun eine Folge weiterer Informationen a, b, c, d, ... betrachten mit den jeweiligen Beobachtungsbasen A_i, B_i, C_i, D_i, ... und dann Updaten zu P_{abcd} so erhalten wir im Sinne der letzten Gleichung die folgende Update-Formel:

$$P_{abcd}(S) = \sum_i \ldots \sum_k P(S\&A_k\&\ldots\&D_i)\cdot[P_a(A_k)/P(A_k)]\cdot\ldots\cdot[P_{abcd}(D_i)/P_{abc}(D_i)]$$

Dabei nennt Hawthorne die Terme in eckigen Klammern *normierte Likelihood Update Faktoren*. (Wären die Informationen a, b, c, d durch Aussagen darstellbar, so hätten wir es mit echten Likelihoods zu tun, die dann noch durch den jeweiligen Nenner normiert werden.) Man

sieht schnell, wie die Formel verallgemeinert werden kann und welche besondere Bedeutung diesen Update Faktoren dabei zukommt.

Beim bisherigen Jeffrey-Updaten setzen wir direkt die jeweils neuen P_{ab} auf den jeweiligen Basen fest. Haben wir es etwa immer mit denselben Basen E_i zu tun, ergibt sich demnach $P_{abc}(E_i) = P_{ac}(E_i)$. Hawthorne spricht daher von *amnesischen Updaten*, denn es gehen immer wieder Informationen wie hier b durch späteres Updaten mit d verloren. Es ist daher kein Wunder, dass das nicht kommutativ ist. Frühere Informationen werden teilweise ausgelöscht und der letzten Information kommt dabei ein besonderes Gewicht zu. Das mag im Einzelfall durchaus sinnvoll sein. Unsere Informationen können implizieren, dass wir unseren früheren Wahrnehmungen lieber nicht mehr trauen sollten oder die verwendeten Messgeräte grob fehlerhaft waren. Dann ist das amnesische Updaten sicher sinnvoll und wir sollten auch keine Kommutativität erwarten. Aber in vielen Fällen werden die neuen Informationen einfach kumulativ auszuwerten sein und die früheren nicht auslöschen. Dafür sollten wir nun ein Modell einer *echten Wahrscheinlichkeitskinematik* suchen.

Man könnte sich das Updaten auch so vorstellen, dass es in geeigneter Weise etwas über die beteiligten Update Faktoren sagt und nicht über die neuen Wahrscheinlichkeiten für die Basen. Wenn man sich hier auf bestimmte Fälle beschränkt und eine erweiterte Rigiditätsanforderung aufstellt, kann man durchaus kommutative Formen des Updatens erhalten. Hawthorne (2004, Abschnitt 6) stellt eine entsprechendes Update-Modell vor.

Das möchte ich hier nicht verfolgen, denn intuitiver ist das dritte Modell von Hawthorne, das auf Field (1978) zurückgeht und das wir daher als *Field-Updaten* bezeichnen werden. (Das allerdings von Hawthorne etwas angepasst wird, um bestimmten Einwänden von Garber (1980) zu entgehen.) Danach legen wir nicht die normierten Likelihood Update Faktoren beim Updaten fest, sondern die sogenannten Likelihood-Quotienten, die im Bayesianismus auch manchmal als Bayes-Faktoren bezeichnet werden und eine intuitive Bedeutung besitzen.

Dabei stehe $LR[P_{ab}, c, C_i, C_j]$ für den Likelihoodquotienten, der sich als Update-Faktor für das Updaten von P_{ab} mit c auf unseren Basen C_i ergibt:

$P_{abc}(C_j)/P_{abc}(C_i) = LR[P_{ab}, c, C_i, C_j] \cdot [P_{ab}(C_j)/P_{ab}(C_i)]$

Der Likelihoodquotient $LR[P_{ab}, c, C_i, C_j]$ gibt also an, wie sich das Verhältnis der Wahrscheinlichkeiten zwischen unseren Basisaussagen ändert, wenn wir mit c updaten. In der bayesianischen Statistik spricht man hier auch vom *Bayes-Faktor*, der uns angibt, in welchem Maße nun C_j gegen über C_i durch c bestätigt wird.

Es wird dort sogar eine Einteilung angegeben, nach der Bestätigungen unterhalb von 3 vernachlässigbar klein sind, die von 3-10 immer noch schwach sind, und wir es erst über 10 mit einer starken Bestätigung zu tun haben. Über 100 handelt es sich dann schon um eine sehr starke Bestätigung. Das ergibt sich daraus, wie sich die Wahrscheinlichkeiten für zwei einander ausschließende Aussagen A und B beim entsprechenden Updaten verändern, wenn sie beide bei 0,5 starten und dann mit entsprechendem LR upgedatet wird. Ist der LR=10, so steigt die erste Aussage auf eine Wahrscheinlichkeit von 90%. Darauf kommen wir noch in der Debatte um Signifikanztests zu sprechen.

Es bietet sich jedenfalls an, beim Updaten den Likelihoodquotienten zu bestimmen, der festlegt, wie stark sich die Wahrscheinlichkeiten zwischen den Basisaussagen verschieben, weil unsere Informationen bestimmte Basisaussagen relativ zu anderen bestätigen. Damit sich dann aber auch die gewünschte Kommutativität einstellt, müssen wir noch eine erweiterte Rigiditätsforderung aufstellen:

> **Erweiterte Rigidität für Likelihoodquotienten:** Für eine Folge von Informationen e mit gemeinsamer Basis E_i und einer Information d gibt es ein Ek, so dass für jede Glaubensfunktion P_a für alle E_j die folgende Identität für die Likelihoodquotienten gilt:
> $LR[P_{ad}, e, E_j, E_k] = LR[P_a, e, E_j, E_k]$

Das heißt die Likelihoodquotienten, die zu einer Information e gehören, sind stabil und nicht von der speziellen Glaubensfunktion abhängig. Dabei habe ich noch eine technische Komplikation weggelassen, die Hawthorne von Field unterscheidet und eine Antwort auf den Einwand von Garber (1980) bietet.

Wir können diese Faktoren nun vereinfachen und erhalten dann als allgemeine *Field-Update-Formel* für alle Sätze S:

$$P_{abcd}(S) = \frac{\Sigma_i \ldots \Sigma_k P(S\&A_k\&B_i) \cdot LR[P, a, A_k, A] \cdot \ldots \cdot LR[P, b, B_i, B]}{\Sigma_i \ldots \Sigma_k P(A_k\&B_i) \cdot LR[P, a, A_k, A] \cdot \ldots \cdot LR[P, b, B_i, B]}$$

wobei A und B einfach spezielle Aussagen aus der entsprechenden Basis sind, für die unserer erweiterte Rigiditätsforderung erfüllt ist. Es lässt sich dann zeigen (s. Hawthorne 2004), dass das Field-Updaten, bei dem wir die Likelihoodquotienten als neue Daten einbringen, wie gewünscht *kommutativ* ist. Joyce (2009a, 453) kommt somit zu dem Fazit, dass das Jeffreys Updaten genau dann kommutativ wird, wenn es das auch sein sollte und sonst nicht. Handelt es sich um ein amnesisches Updaten, so werden frühere Informationen damit zerstört oder unterminiert und sollen dann auch nicht mehr berücksichtigt werden. Ist das Updaten aber nicht amnesisch, so sollte es in Form des Field-Updatens so geschehen, dass es die erweiterte Rigiditätsforderung auch noch erfüllt und damit kommutativ wird. Wir haben so eine echte Wahrscheinlichkeitskinematik erhalten.

Leider ist die Formel für das Field-Updaten so komplex, dass wir uns im Normalfall wohl mit der einfacheren klassischen Herangehensweise begnügen werden, zumal auch das Einbringen der Daten komplexer geworden ist, obwohl es zumindest eine intuitive Deutung der Likelihoodquotienten gibt. Aber in der Praxis wird man vermutlich eher zu klassischen bayesianischen Verfahren greifen und sich etwa im Rahmen der bayesianischen Netze mit bestimmten Extraknoten für die Zuverlässigkeit bestimmter Informationen begnügen. Zumindest ist aber das Jeffrey-Updaten theoretisch rehabilitiert und wir wissen nun genauer, unter welchen Bedingungen und aus welchen Gründen das Jeffrey-Updaten kommutativ oder nicht-kommutativ ist.

5.6 Bayesianische Bestätigungstheorien

Bisher haben wir vor allem das bayesianische Updaten unseres Überzeugungssystems dargestellt und dabei darauf geachtet, bestimmte Regeln zu formulieren, die vernünftige Glaubensgrade kennzeichnen. Damit haben wir mehr implizit als explizit gewisse Zusammenhänge zur Erkenntnistheorie hergestellt. Die wollen wir nun explizit machen und präzisieren. Wann *bestätigen* danach bestimmte Daten D eine Hypothese

H? Dazu müssen wir als elementaren Zusammenhang, der wohl von allen Bayesianern geteilt wird, die folgenden Regeln ansehen. Ein Bayesianer sollte diesen Regeln jedenfalls zustimmen, weil seine Glaubensgrade sonst für die Erkenntnistheorie nutzlos wären:

Bestätigung und Glaubensgrade
(1) $P(H|E) > P(H)$ gdw. E bestätigt H
(2) $P(H|E) = P(H)$ gdw. E ist neutral bzgl. H
(3) $P(H|E) < P(H)$ gdw. E schwächt H (bzw. E bestätigt ¬H)

Damit hätten wir einen ersten noch recht allgemeinen Zusammenhang zwischen Glaubensgraden und einer epistemischen Bestätigung. Doch, was besagt der, wenn wir einmal nur davon ausgehen, dass es sich um beliebige Glaubensgrade einer Person s mit der Glaubensgradfunktion P_s handelt? Da P_s einfach der Ersatz für unseren kategorialen Glauben darstellt, sagt »$P_s(H|E) > P_s(H)$« eigentlich nur, dass der bedingte Glaubensgrad von H gegeben E der Person s höher ist als ihr einfacher Glaubensgrad an H. Das müssten wir dann eigentlich so übersetzen: »*Subjekt s hält E für eine Bestätigung von H.*« Das sagt uns jedoch noch nichts über eine objektive Bestätigungsbeziehung. Es ist nur eine Beschreibung der Überzeugungen von s und nicht mehr. Solange noch nicht sichergestellt ist, dass P_s weitere objektive Anforderungen erfüllt, liefert die Ungleichung keine weiteren Informationen über den tatsächlichen Zusammenhang zwischen E und H.

Das heißt andererseits, dass jede zusätzliche Anforderung an P_s uns einen Schritt näher an eine genuine erkenntnistheoretische Position heranführt. Deshalb sind die Zusatzregeln so besonders wichtig für den Bayesianismus. Der klassische Bayesianer kennt praktisch nur die Wahrscheinlichkeitsaxiome und die Konditionalisierungsregel, doch das sind nur einfache Konsistenzforderungen. Man kann als Bayesianer also weiter ziemlich verrückte Werte für $P(H|E)$ und $P(H)$ aufweisen. Erst wenn es uns gelingt, diese sinnvoll zu beschränken, haben wir eine erkenntnistheoretische These aufgestellt. Man könnte das so formulieren: Jeder Schritt in Richtung einer induktiven Logik verhilft dem Bayesianismus erst zu einem erkenntnistheoretischen Gehalt. Ob

es sich dabei schließlich um einen fortschrittlichen Ansatz in der Erkenntnistheorie handelt, müssen wir aber trotzdem noch anhand seiner Anwendungen ermitteln.

Einen solchen Schritt fanden wir im statistischen Syllogismus, aber leider hilft er uns in der bayesianischen Beurteilung wissenschaftlicher Theorien nicht weiter, wie ich bereits angemerkt habe. Das wichtigste Instrument zur Objektivierung bayesianischer Überlegungen ist daher die Likelihoodanbindung. Objektive Likelihoods sind die Wahrscheinlichkeiten, die Theorien dem Auftreten bestimmter Daten geben. Sie sind Teil des Gehalts der Theorien. Hier kommen objektive Bayesianer wieder mit der klassischen Statistik zusammen, die ebenfalls auf die Likelihoods setzt. Der Bayesianer kann so auf den Vorwurf der rein subjektiven Auskunft reagieren: Die Ausgangswahrscheinlichkeiten bleiben subjektive Einschätzungen (von möglicherweise objektiven Wahrscheinlichkeiten), aber der Prozess des Updatens wird vor allem durch die objektiven Likelihoods dominiert und erhält so seinen objektiven Charakter – jedenfalls, wenn wir die Likelihoodanbindung unterschreiben. Damit mögen die Startwahrscheinlichkeiten unterschiedlicher Bayesianer zwar stark variieren, bei den Likelihoods sollte dann aber wieder Einigkeit herrschen. Das reicht aus, um letztlich Einigkeit bei den Nachher-Wahrscheinlichkeiten zu erzielen, wenn genügend Daten eingehen. Das soll jedenfalls die Aussage der bayesianischen Konvergenztheoreme sein. Leider weist uns Hawthorne (2005) darauf hin, dass diese schöne Idee nicht so einfach funktioniert. Doch dazu später mehr (vgl. Kap. 5.8).

Überhaupt bleibt immer die Frage zu beantworten, ob (1) geeignet ist, um damit möglichst viele unserer Redeweisen von Bestätigung abdecken oder sogar erklären zu können. In (1) sagen wir, dass es eine positive statistische Korrelation zwischen E und H gibt. Das ist u.a. eine symmetrische Relation, weshalb dann auch E durch H gestützt wird. Das muss nicht als problematisch betrachtet werden, erscheint aber nicht jedem als plausibel. Christensen (1999) sieht das schon als Abweichung von unserer üblichen Sicht der Bestätigungsbeziehung. Schwerwiegender sind wohl die Einwände von Peter Achinstein, die wir später (in Kap. 5.8.8) diskutieren werden. Außerdem bleibt natürlich die Frage zu beantworten, wie wir zu Startwahrscheinlichkeiten für wissenschaftliche Theorien kommen sollen. Vorteilhaft an der bayesia-

nischen Konzeption scheint andererseits zu sein, dass wir es nun in der Hand haben, das Ausmaß der Bestätigung von Theorien durch Daten genauer angeben zu können. Beim hypothetisch-deduktiven Verfahren war das Unterbestimmtheitsproblem offen geblieben. So kann ein Datum aus verschiedenen Kombinationen von Theorien und Daten abgeleitet werden (T,H ⇒ E und T',H' ⇒ E). Welche der Theorien wird dann durch die Daten bevorzugt? Ein Maß für die Bestätigungsstärke könnte hier eine erste Antwort bieten.

5.6.1 Maße der Bestätigung

Die Bestätigung einer Theorie durch bestimmte Daten kann offensichtlich recht unterschiedlich ausfallen. Rein qualitative Bestätigungstheorien können das nicht richtig nachzeichnen. Die Glaubensgrade sollten für uns das geeignete Mittel darstellen, nun ein solches Maß anzugeben. Ein sehr naheliegendes Maß für die Bestätigungsstärke ist das einfache Differenzmaß:

(1) **Differenzmaß**: $d(H,E) = P(H|E) - P(H)$

Doch Fitelson (1999) konnte zeigen, dass sich für dieses Maß bestimmte Paradoxien ergeben. So ist die Paradoxie von Popper und Miller (1983) auf eine bestimmte Additivitätseigenschaft angewiesen, die unser Maß d aufweist, andere Maße wie das Ratio-Maß aber nicht zeigen. Daher wurden inzwischen zahlreiche andere Maße vorgeschlagen und es gibt eine lebhafte Debatte um das richtige Maß. Typische Beispiele sind:

(2) ***Ratio-Maß***: $r(H,E) = \log[P(H|E)/P(H)]$
(3) ***Likelihood-Ratio-Maß***: $l(H,E) = \log[P(E|H)/P(E|\neg H)]$
(4) ***Zweites-Differenz-Maß***: $s(H,E) = P(H|E) - P(H|\neg E)$

Rosenkrantz (1994) stützt sich dagegen speziell auf das Maß d, um eine Antwort auf das Problem der irrelevanten Konjunktionen zu finden. Fitelson (1999 und 2001) stellt die jeweils relevanten Eigenschaften sogar in Form einer Tabelle zusammen, wobei er sich in (1999) aber als viertes Maß auf ein von Carnap vorgeschlagenes etwas komplizierteres Maß bezieht. Man sieht schnell die unterschiedlichen Einschätzungen

der Bestätigungsstärke: Steigt die Wahrscheinlichkeit von H von 0,0001 auf 0,01, so ergibt das im Differenzmaß nur eine Steigerung von ca. 0,01, was uns eher als kleinerer Betrag erscheint, während das Ratio-Maß schon den Wert 100 liefert, was uns recht groß erscheint. Doch genau genommen sind die jeweiligen Werte nur schwer miteinander vergleichbar und letztlich wird es darum gehen, welches der Maße unsere intuitiven Einschätzungen in konkreten Anwendungsfällen und natürlich in den paradoxen Beispielen besser rekonstruieren kann.

Insbesondere sind die Maße untereinander nicht äquivalent, in dem Sinne, dass für alle Paare (H,E) und (H',E') die Maße M und M* zumindest in der Reihenfolge übereinstimmen: M(H,E) ≥ M(H',E') gdw. M*(H,E) ≥ M*(H',E'). Selbst diese ordinale Äquivalenz ist für unsere Maße nicht gegeben. Wir können sie also keineswegs als im Grunde doch gleichwertige Maße betrachten.

Ein Argument in der Debatte um das richtige Maß ist z.B. die mangelnde Sensibilität von r gegenüber deduktiver Irrelevanz, die etwa von Rosenkrantz (1994) und Gillies (1986) bemängelt wurde:

Wenn H ⇒ E, dann gilt r(H,E) = r(H&X,E), für alle Aussagen X.

Es wirkt eher unplausibel, dass die stärkere Hypothese H&X genauso gut durch die Daten gestützt wird (wohlgemerkt für ein beliebiges X) wie die schwächere Teilhypothese H selbst. Das spricht intuitiv deutlich gegen das Ratio-Maß.

Fitelson (2001) diskutiert und expliziert eine ganze Reihe von Anforderungen an solche Bestätigungsmaße, wobei das Likelihood-Ratio-Maß l seiner Meinung nach am besten abschneidet. Die bisherige Debatte zeigt aber wohl, dass es dabei keinen eindeutigen klaren Sieger gibt. Fitelsons Favorit ist wie gesagt das Likelihood-Ratio-Maß, da es auch in der Paradoxienbekämpfung am besten abschneidet. Christensen (1999) ist eher pessimistisch, dass es gelingt, ein Maß für alle Anwendungsfälle zu finden. Alle Maße werden auf jeden Fall von dem »old-evidence«-Problem geplagt, auf das wir noch gesondert eingehen werden (vgl. Kap. 5.8).

Tentori et al. (2007) untersuchten die unterschiedlichen Maße empirisch, indem sie ermittelten, in welchen Fällen von positiven Maßen auch

die Versuchspersonen von einer bestimmten Bestätigung ausgingen. Hier schnitt zumindest das Likelihood-Ratio-Maß l neben anderen Maßen, die wir hier nicht betrachten, wieder recht gut ab.

Das überrascht mich einerseits nicht, denn dieses Maß beschreibt am besten, was die beteiligten Theorien über die Daten sagen, andererseits ist es kein rein bayesianisches Maß mehr, wenn wir die Likelihoodanbindung akzeptieren. Jedenfalls kann für die objektiven Fälle von Likelihoods auch der klassische Statistiker diese Größen verwenden. Und in Kapitel 6.4.2 werden wir den Bayes-Faktor im Rahmen der bayesianischen Statistik für die Hypothesenwahl kennenlernen, der die relative Stärke der Bestätigung zwischen bestimmten Hypothesen angibt. Er ist eng verwandt mit dem Likelihood-Ratio-Maß l und spricht wiederum dafür, das als zentrales Maß im Rahmen der Theorienbestätigung anzusehen. Das passt auch zu unserer Vorgehensweise für das abduktive Schließen, bei der wir den Likelihoodquotienten ebenfalls als einen wesentlichen Aspekt bei Vergleichen der Erklärungsstärke eingestuft haben. Allerdings hatten wir bereits gesehen, dass es im Falle disjunktiver Hypothesen keine eindeutigen objektiven Likelihoods gibt, sondern dass wir hier auf bestimmte subjektive Wahrscheinlichkeiten angewiesen sind, um die Likelihoods zu berechnen. Überraschend ist aber doch, dass die eigentlichen epistemischen Wahrscheinlichkeiten der Hypothesen selbst nicht im Zentrum des ausgewählten Maßes stehen.

Eine neue Klasse von Maßen auf der Grundlage des Z-Maßes haben Crupi et al. (2007) vorgeschlagen. Das Z-Maß hat noch mehr wünschenswerte logische Symmetrieeigenschaften und kommt unseren intuitiven Urteilen in Tentori et al. (2007) noch weiter entgegen. Das zeigt, dass hier noch Raum für weitere Forschungen besteht und ich möchte das Maß Z zumindest noch erwähnen:

$$Z(h,e) = \begin{cases} [P(h|e) - P(h)]/[1 - P(h)], & \text{falls } P(h|e) > P(h) \\ [P(h|e) - P(h)]/P(h), & \text{sonst} \end{cases}$$

Wir können diese Maße nun einfach einmal auf unser erstes Beispielsystem aus Kapitel 5.3 anwenden. Einmal für das erste Updaten mit E und dann für das folgende zweite Updaten mit D (also für die Wahrscheinlichkeitsfunktion P^+). Dann erhalten wir (mit dem Logarithmus zur Basis 10):

d(H,E) = 0,7 – 0,5 = 0,2
d(H,D) = 0,833 – 0,7 = 0,133

r(H,E) = log(0,7/0,5) = log(1,4) = 0,146
r(H,D) = log(0,833/0,7) = log(1,19) = 0,076

l(H,E) = log(0,7/0,3) = log(2,333) = 0,368
l(H,D) = log(0,714/0,333) = log(2,144) = 0,331

s(H,E) = 0,7 – 0,3 = 0,4
s(H,D) = 0,833 – 0,5 = 0,333

Z(H,E) = 0,4
Z(H,D) = 0,443

Hieran sieht man zunächst, dass alle Maße bis auf Z darin übereinstimmen, dass in unserem Beispiel E eine stärkere Bestätigung für H bietet als D, wobei aber die Zahlenverhältnisse für die Bestätigungswerte durchaus recht unterschiedlich ausfallen. Für Z sieht es aber überraschenderweise sogar umgekehrt aus, dass D die stärkere Bestätigung darstellt. Daran erkennt man schon, dass die Maße nicht stabil sind, was die Reihenfolge der Bestätigungsstärke angeht (nicht ordinal äquivalent sind). Betrachten wir die entsprechenden Werte noch für das entsprechende Beispiel aus Kapitel 5.3.10 mit der dort verlangten stabilen Likelihoodanbindung:

d(H,E) = 0,7 – 0,5 = 0,2
d(H,D) = 0,845 – 0,7 = 0,145

r(H,E) = log(0,7/0,5) = log(1,4) = 0,146
r(H,D) = log(0,845/0,7) = log(1,207) = 0,081

l(H,E) = log(0,7/0,3) = log(2,333) = 0,368
l(H,D) = log(0,7/0,3) = log(2,333) = 0,368

s(H,E) = 0,7 – 0,3 = 0,4
s(H,D) = 0,845 – 0,5 = 0,345

Z(H,E) = 0,4
Z(H,D) = 0,483

Hier fällt vor allem auf, dass nun das Likelihoodmaß l für beide Daten dieselbe Bestätigungswirkung liefert, was auch nicht anders zu erwarten

war. Aber für die anderen Maße gilt weiterhin, dass E unsere Hypothese H stärker bestätigt als das Datum D. Daran ändert auch die Likelihoodanbindung nichts. Wenn Bayesianer das gerade wünschen, weil sie etwa erwarten, dass das wiederholte Datum Kopf unsere Hypothese H beim zweiten Mal nicht so stark bestätigt wie beim ersten Mal, dann wird das für die meisten Maße weiterhin gewährleistet. Man sieht allerdings sehr gut, wie stark solche Aussagen von der jeweiligen Wahl eines Maßes abhängen. Beim Maß Z ist die Bestätigung durch D nun sogar noch einmal größer geworden.

5.6.2 Bayesianische Konvergenz

Klassische Statistiker werden gegen den Bayesianismus vor allem vorbringen, dass er zu subjektiv gefärbt sei, denn die verwendeten Glaubensgrade sind jeweils spezifisch für bestimmte epistemische Subjekte – und können also von Wissenschaftler zu Wissenschaftler stark variieren. Was haben solche Größen in der wissenschaftlichen Beurteilung einer Theorie zu suchen? Seine persönlichen Einschätzungen mag jeder Wissenschaftler haben, aber warum sollten die irgendeine Rolle in der wissenschaftlichen Bewertung unserer Theorien spielen? Vielmehr sollten dort lieber allein das (allgemein anerkannte) Hintergrundwissen und vor allem unsere empirischen Daten im Hinblick auf die Theorie ausgewertet werden.

Argumente für oder gegen eine Theorie sollten dann z.B. die Form haben: Unser Hintergrundwissen über die vier wirksamen Grundkräfte sagt uns, dass es Einflussmechanismen wie die in der Astrologie postulierten nicht geben kann, und die Daten sprechen damit gegen die Vorhersagekraft astrologischer Berechnungen: Deshalb ist die Astrologie als unplausibel abzulehnen. Ein Bayesianer müsste dagegen ansetzen: Mein persönlicher Glaubensgrad an die Astrologie ist von Beginn an schon sehr klein und durch die Updates mit den Daten noch kleiner geworden. Der Einwand ist natürlich, dass der erste Teilsatz nur eine nette autobiographische Auskunft ist, die aber für die weitere Beurteilung der Astrologie nicht weiter ins Gewicht fallen sollte. Einstein hätte das Entsprechende vielleicht sogar von der Quantenmechanik gesagt, aber selbst in dem Fall hätten wir doch darauf bestanden, dass letztlich nur

die Fakten – also die Erklärungserfolge und Vorhersageleistungen der Quantenmechanik – zählen und nicht die persönlichen Einschätzungen eines Wissenschaftler – ganz gleich wie berühmt und genial er sein mag.

Es gibt eine Standardantwort, die Bayesianer immer gegen den Vorwurf der Subjektivität ihrer Glaubensgrade vorbringen, und das ist die der *Konvergenz*. Die Glaubensgrade mögen zu Beginn divergieren, aber wenn wir nur genügend Daten sammeln, werden sie sich beim Updaten mit diesen Daten schnell einander annähern. Das kann man zunächst in Beispielen gut beobachten, wenn wir mit normalen Startwahrscheinlichkeiten beginnen und einige plausible Annahmen (wie die Likelihoodanbindung) unterschreiben. Das illustrierten schon unsere früheren konkreten Beispiele (vgl. Kap. 5.3.12). Alexander Bird (1998, 208 f.) bringt ein weiteres Zahlenbeispiel.

Leider ist es beim bayesianischen Updaten besonders mühselig, die jeweiligen Nenner P(E) zu bestimmen. Das geschieht normalerweise mit Hilfe des Theorems der totalen Wahrscheinlichkeit, wenn wir schon über eine erschöpfende Menge an einander ausschließenden Hypothesen $\{H_1,\ldots,H_n\}$ verfügen. Wir können hier das bayessche Theorem dann so beschreiben, dass wir ansetzen:

$$P^+(H_i) = P(H_i|E) \propto P(H_i)\cdot P(E|H_i)$$

Dabei steht »$\propto$« für »ist proportional zu«. Die neue Wahrscheinlichkeit für eine Hypothese H_i ist daher proportional dem Produkt aus der alten Wahrscheinlichkeit von H_i und der Likelihood von E gegeben H_i. Der *Proportionalitätsfaktor* x zwischen $P^+(H_i)$ und $P(H_i)\cdot P(E|H_i)$ bestimmt sich dann daraus, dass die neue Wahrscheinlichkeitsverteilung für die Hypothesen wieder zusammen 1 ergeben muss:

$$1 = \sum_i P^+(H_i) = \sum_i x\cdot P(H_i)\cdot P(E|H_i) = x\cdot\sum_i P(H_i)\cdot P(E|H_i),$$
$$\text{also ist } x = 1/[\sum_i P(H_i)\cdot P(E|H_i)] = 1/P(E)$$

Diese Form des bayesschen Theorems lässt sich im Rahmen der bayesschen Statistik oft gut einsetzen, um die Nachher-Wahrscheinlichkeiten zu finden oder abzuschätzen.

Einfacher werden Berechnungen der upgedateten Wahrscheinlichkeitsfunktionen gleichfalls, wenn wir die Quotientenform des bayesianischen Theorems verwenden (dabei fällt der Proportionalitätsfaktor

weg), womit wir uns zugleich ein wenig den Likelihoodisten annähern. James Hawthorne hat dieses Verfahren für viele Zwecke immer wieder erfolgreich eingesetzt. Nehmen wir zunächst an, wir hätten nur zwei einander ausschließende Hypothesen H_1 und H_2 und eine davon wäre wahr. Aus dem bayesschen Theorem wird dann:

Bayessches Theorem in Quotientenform (BTQ)

$$\frac{P^+(H_1)}{P^+(H_2)} = \frac{P(H_1) \cdot P(E|H_1)}{P(H_2) \cdot P(E|H_2)}$$

Hier besteht der rechte Term aus den Vorher-Wahrscheinlichkeiten der Hypothesen und den Likelihoods der jeweiligen Daten E. Wir können nun recht leicht den hier *Quotienten-Update-Faktor* (QUF) genannten Faktor $Q(E) = P(E|H_1)/P(E|H_2)$ für verschiedene Daten E betrachten (der häufig als *Bayes-Faktor* BF_{12} bezeichnet wird vgl. Kap. 6). Geht dieser Faktor gegen 0, so geht auch $P^+(H_2)$ gegen 0 und damit muss $P^+(H_1)$ gegen 1 gehen. Das können wir dann auf den Fall von n Hypothesen ausdehnen. Bleiben wir aber zunächst bei zwei Hypothesen. Nehmen wir z.B. eine Münze, die entweder eine Wahrscheinlichkeit von 0,6 für Kopf hat (H_1 besagt $P(\text{Kopf}|H_1) = 0{,}6$) oder nur 0,3 ($H_2$ besagt $P(\text{Kopf}|H_2) = 0{,}3$). Dann können wir den Faktor Q(E) für verschiedene Ergebnisse E ausrechnen (unter Einsatz der Likelihoodanbindung), die hier immer aus konkreten Folgen von Kopf und Zahl besteht. Nehmen wir außerdem an, dass der wahre Wert gerade 0,3 ist und deshalb die Anzahl der Köpfe auf lange Sicht mit einer relativen Häufigkeit von ungefähr 0,3 auftritt. So könnten etwa die für eine einzelne Folge von Münzwürfen die Ergebnisse auftreten:

Ein konkretes Zahlenbeispiel:

E_1 = 4 Köpfe von 10; E_2= 33 Köpfe von 100;
E_3 = 290 Köpfe von 1000, dann ist:

$Q(E_1) = 0{,}6^4 \cdot 0{,}4^6 \;/\; 0{,}3^4 \cdot 0{,}7^6 = 5{,}57 \cdot 10^{-1} = 0{,}557$
$Q(E_2) = 0{,}6^{33} \cdot 0{,}4^{67} \;/\; 0{,}3^{33} \cdot 0{,}7^{67} = 4{,}47 \cdot 10^{-7} = 0{,}000000447$
$Q(E_3) = 0{,}6^{290} \cdot 0{,}4^{710} \;/\; 0{,}3^{290} \cdot 0{,}7^{710} = 5{,}52 \cdot 10^{-86}$

Jetzt können wir für verschiedene Startwahrscheinlichkeiten von H_1 und H_2 die neuen Endwahrscheinlichkeiten ausrechnen, denn $P^+(H_1) = 1 - P^+(H_2)$.

Formel für die Endwahrscheinlichkeiten:
Setze: h: = $P(H_1)/P(H_2)$ und Q(E): = $P(E|H_1)/P(E|H_2)$,
dann erhalten wir:
$P^+(H_1) = 1/[1+h\cdot Q(E)]$ und
$P^+(H_2) = h\cdot Q(E)/[1+h\cdot Q(E)]$

Man erkennt hieran sehr schnell, was passiert, wenn der Quotienten-Update-Faktor Q(E) (bzw. der Bayes-Faktor) gegen Null oder gegen Unendlich geht. Im ersten Fall geht $P^+(H_1)$ für festes h gegen Eins und $P^+(H_2)$ gegen Null. Im zweiten Fall sieht es offensichtlich andersherum aus. Wenn wir etwa davon ausgehen, dass H_1 und H_2 zu Beginn ungefähr dieselbe Wahrscheinlichkeit von 0,5 haben, so wird h = 1 und damit sieht man für sehr kleine Werte von Q(E), dass $P^+(H_1)$ nahe 1 liegt und $P^+(H_2)$ ungefähr bei Q(E).

In unserem konkreten Beispiel sieht man, wie schnell die Q(E) mit größeren Zahlen an Würfen gegen Null gehen. Damit haben wir schon die Grundidee einer Variante der bayesianischen Konvergenztheoreme kennengelernt. Weil die allermeisten Daten, die bei einer so gestalteten Münze (mit P(Kopf) = 0,3) auftreten werden, in Bezug auf die relative Häufigkeit von Kopf nahe bei unseren Daten liegen, erwarten wir auch in den allermeisten Fällen ähnliche Werte für Q(E). Damit geht in den allermeisten Fällen $P^+(H_1)$ gegen Null und $P^+(H_2)$ gegen 1. Nun müssen wir nur noch »allermeisten« durch »mit hoher Wahrscheinlichkeit« ersetzen und finden für unser Beispiel und ähnliche Fälle die gesuchte Konvergenz (s.u.).

Für den Fall von n Hypothesen ist die Berechnung nur etwas komplexer. Aber wenn genügend andere Hypothesen praktisch falsifiziert wurden, indem ihre Wahrscheinlichkeiten sich der Null angenähert haben, so verteilt sich die Wahrscheinlichkeit auf die verbliebenen Hypothesen, also im günstigen Fall auf die eine zurückbleibende Hypothese (vgl. Hawthorne 2008, 2005, 2010).

Zu den Voraussetzungen dieser Konvergenz gehört wieder (wie im Falle des abduktiven Schließens), dass wir eine Liste von Hypothesen haben, die die wahre Hypothese enthält. Eine weitere Voraussetzung, die wir in den weiteren Konvergenztheoremen explizit wiederfinden werden, ist,

dass die Konkurrenten zur wahren Hypothese einen gewissen Abstand aufweisen müssen.

Tatsächlich finden sich eine ganze Reihe leicht unterschiedlicher Konvergenztheoreme, von denen ich einige kurz skizzieren möchte. Zunächst ein Theorem in einer recht allgemeinen Form mit möglicherweise unendlich vielen Konkurrenzhypothesen und einem größeren Wertebereich für die Daten. So finden wir bei Leonhard Held (2008, 172) in etwa (leicht angepasst an unsere Notation und Redeweisen) das folgende Theorem:

> **Theorem**: Es sei $H = \{h_1, h_2, \ldots\}$ eine abzählbare Menge von Hypothesen mit $P(h_i) > 0$ für alle $i \neq t$ und es sei die wahre Hypothese h_t in H und x^n sei der Vektor der Daten als Ergebnisse einer Zufallstichprobe von n Zufallsvariablen $X_1, \ldots, X_n$. Mit $f(x^n|h_i)$ sei die Wahrscheinlichkeitsdichte der Hypothese h_i gemeint. Außerdem sei der Kullback-Leibler-Abstand von h_t und h_i größer 0:
>
> $I(h_t,h_i) = \int f(x^n|h_t) \log[f(x^n|h_t)/f(x^n|h_i)]\, dx^n > 0$ für alle $i \neq t$, dann gilt:
>
> (*) $\lim_{n\to\infty} P(x^n|h_t) = 1$ und $\lim_{n\to\infty} P(x^n|h_i) = 0$ für alle $i \neq t$.

Als Voraussetzung haben wir somit die Bedingung, dass die wahre Hypothese in unserer Menge zu finden ist und dass alle anderen Hypothesen andere Behauptungen über die Beobachtungsdaten aufstellen. Beobachtungsäquivalente Hypothesen sind natürlich durch das bloße Updaten mit neuen Daten nicht voneinander zu unterscheiden. Um das sicherzustellen, wird hier wieder der Kullback-Leibler-Abstand I eingesetzt, den wir oben schon kennengelernt haben, nur dass er hier in einer Version für Dichten auftritt und die Beobachtungsdaten x^n nun einen kontinuierlichen Bereich von Werten annehmen können, über den wir im Integral in I zu integrieren haben. Außerdem wird verlangt, dass alle Hypothesen noch nicht endgültig falsifiziert sind, sondern eine Wahrscheinlichkeit größer Null erhalten. Wir hatten bereits gesehen, dass dieses Dogmatismusverbot nicht ganz so harmlos ist, wie man auf den ersten Blick denken könnte. Etwas besser angepasst an einen probabilistischen Ansatz, der sich auf Aussagen bezieht, ist das Theorem von Gaifman und Snir (1982), das etwa bei Earman (1992, Kap. &.5) diskutiert wird.

Besonders spannend erscheint mir aber das Likelihood-Quotienten-Konvergenztheorem, das vor allem von Hawthorne (2008, 2005, 2010) in verschiedenen Varianten vorgestellt und propagiert wird. Es stellt praktisch die entsprechende Variante des Konvergenztheorems zu unserem obigen Zahlenbeispiel dar. Es kommt mit recht wenigen Annahmen aus. So wird noch nicht einmal verlangt, dass die Wahrscheinlichkeiten σ-additiv sind, wie das für die obigen Theoreme verlangt werden muss, und es kommen auch keine Wahrscheinlichkeiten 2. Ordnung (also Wahrscheinlichkeiten über Wahrscheinlichkeiten) zum Einsatz. Darüber hinaus kann auch der klassische Statistiker oder zumindest ein Likelihoodist etwas damit anfangen, jedenfalls dann, wenn die vorkommenden Likelihoods alle objektiver Natur sind. Wir hatten jedoch oben schon gesehen, dass das zumindest für disjunktive-Hypothesen nicht immer gegeben ist.

5.6.3 Das Likelihood-Quotienten-Konvergenztheorem von Hawthorne und die moderne induktive Logik

Hawthorne (2011c, 2005, 2010) bezieht sich auf das bayessche Theorem in Quotientenform, führt aber noch einige der zusätzlichen Bestandteile des bayesianischen Updatens explizit mit auf, was ich nun ebenfalls tun möchte. Die Ausdrücke werden dadurch zwar etwas komplizierter, aber wir sind im Prinzip in der Lage, die Einflüsse von Hilfsannahmen genauer anzugeben und mit dem bayesianischen Apparat nachzuzeichnen. Diese Möglichkeiten sollen nicht unterdrückt werden, da sonst der bayesianische Ansatz hier schon gegenüber dem hypothetisch-deduktiven als weniger detailliert erscheinen könnte. Solche Komplikationen habe ich bisher nur der besseren Übersichtlichkeit halber weggelassen bzw. die Hilfsannahmen schlicht als Bestandteile des Hintergrundwissens betrachtet. Hawthorne hat dazu eine recht übersichtliche Notation eingesetzt, die ich mit kleinen Abänderungen kurz erläutern werde, um zumindest eine relativ einfache Version seines Konvergenztheorems vorzustellen (vgl. Hawthorne 2005). Es handelt sich dabei um eine Anwendung des schwachen Gesetzes der großen Zahlen.

Hawthorne betrachtet die Likelihoods P(e|h&b&c), wobei e wieder unser Datum darstellt, h unsere in Frage stehende Hypothese, b unser

Hintergrundwissen (für »background knowledge«) und c unsere Annahmen über die spezielle Situation, in der das Datum e erhoben wurde. Das könnte bei einer Beobachtung etwa die speziellen Lichtverhältnisse betreffen, die dabei geherrscht haben. Dazu kommen einfache Abkürzungen (wir haben schon ähnliche verwendet), um mehrere Daten darzustellen. Mit $P(e^n|h\&b\&c^n)$ ist der Fall gemeint, dass wir n Daten konjunktiv zusammenfügen und die dazu jeweils passenden n (möglicherweise unterschiedlichen) Randbedingungen c^n annehmen. Weiterhin haben wir eine abzählbare Liste von Hypothesen $\{h_1, h_2, \ldots\}$ über ein Gebiet, die einander ausschließen und zusammen erschöpfend sind, und zur Not eben eine »catch-all«-Hypothese enthalten, um im endlichen Falle eine vollständige Liste zu ergeben. Darin sei h_i wahr. Dann wird aus dem bayesschen Theorem in Quotientenform:

Bayessches Theorem in Quotientenform (BTQ)

$$\frac{P(h_i|e^n\&b\&c^n)}{P(h_j|e^n\&b\&c^n)} = \frac{P(h_i|b)\cdot P(e^n|h_i\&b\&c^n)}{P(h_j|b)\cdot P(e^n|h_j\&b\&c^n)}$$

Das gilt zumindest, wenn wir annehmen, dass $P(c^n|h_j\&b) = P(c^n|h_i\&b)$ gilt, dass also unsere Beobachtungsbedingungen nicht von unseren jeweiligen Hypothesen abhängen, was zumindest auf den ersten Blick recht plausibel erscheint. Dann erhalten wir ein übersichtliches Theorem, dessen Termen wir wieder Namen geben können. Es sagt uns, wie die Wahrscheinlichkeit der wahren Hypothese h_i gegen 1 konvergiert im Vergleich mit einer beliebigen Hypothese h_j. Wir haben zunächst erneut unseren wichtigen Quotienten-Update-Faktor (QUF) $Q_{i,j}(e^n)$, der einfach wie gehabt der Quotient der entsprechenden Likelihoods ist:

$$Q_{i,j}(e^n) := P(e^n|h_i\&b\&c^n) \;/\; P(e^n|h_j\&b\&c^n)$$

Diese Faktoren können wachsen oder gegen Null gehen im Vergleich der Hypothesen. Ebenso haben wir wiederum die Quotienten von Vorher- und Nachher-Wahrscheinlichkeiten wie schon im oberen Fall. Bedeutsam ist natürlich vor allem, was mit dem QUF geschieht. Dazu betrachtet Hawthorne die Menge all der Daten e^n, die zu einem bestimmten Hypothesenpaar zu einem kleinen QUF führt.

$Q_{i,j}(\varepsilon) := \{e^n; Q_{i,j}(e^n) < \varepsilon\}$ mit $\varepsilon>0$

Daraus wird nun ein Satz $A_{i,j}(\varepsilon)$ gebildet als eine lange Disjunktion, die einfach nur besagt, dass ein Satz von Daten e^n aus dieser Menge auftritt. Damit besagt $A_{i,j}(\varepsilon)$, dass ein Satz von Daten auftritt, für den der Quotienten-Update-Faktor kleiner wird als das gewählte ε, bei Vorliegen der Bedingung $h_i\&b\&c^n$. So erhalten wir das:

Likelihood-Quotienten-Konvergenz-Theorem (LQKT)
$\forall\varepsilon>0\ \exists q_{i,j} > 0$ mit: $P(A_{i,j}(\varepsilon)|h_i\&b\&c^n) \geq 1 - (q_{i,j}/n)$

Dabei hängt der genaue Wert von $q_{i,j}$ ab vom empirischen Abstand der beiden Hypothesen h_i und h_j. Dabei werden ähnlich wie beim Kullback-Leibler Abstand alle möglichen empirischen Ergebnisse in Betracht gezogen und etwa anhand bestimmter Unabhängigkeitsannahmen weiter vereinfacht. So lässt sich dann $q_{i,j}$ berechnen (vgl. Hawthorne 2008, Kap. 5 und 2011a/c). Aber das Theorem lässt sich bereits in der Form (LQKT) verstehen und interpretieren. Zunächst ist klar, dass mit wachsender Anzahl n an Daten die in (LQKT) betrachtete Wahrscheinlichkeit gegen 1 wächst. Damit besagt das Theorem: Zu beliebig klein gewähltem ε gilt: Wenn h_i die wahre Hypothese ist und unser Hintergrundwissen b wahr ist sowie auch unsere Annahmen c^n über die Situation der Datenerhebung, dann werden für jede andere Hypothese h_j mit hoher Wahrscheinlichkeit (gegen 1 wachsend bei weiterer Datensammlung) unsere Daten e^n in einem Bereich liegen, für den der Quotienten-Update-Faktor kleiner wird als ε. Wie in unserem Beispiel oben bedeutet ein kleiner QUF, dass die Daten gegen h_j sprechen. Nur dass die Formel jetzt etwas komplexer geworden ist, weil wir mehr Hypothesen berücksichtigen müssen, als in unserer Beispielrechnung. Damit ist (LQKT) so zu verstehen, dass wir mit zunehmender Anzahl an Daten einen QUF kleiner ε mit einer hohen Wahrscheinlichkeit erwarten dürfen, wobei die Wahrscheinlichkeit immer größer wird bei wachsendem n. Die anderen Hypothesen werden hingegen dabei letztlich probabilistisch falsifiziert.

Die Abhängigkeit der Endwahrscheinlichkeiten von den Quotienten-Update-Faktoren lässt sich wieder erkennen, indem wir die Gleichungen etwas umschreiben und zu den *Quoten* unserer Hypothesen übergehen.

Dabei ist die Quote von A gegeben B gerade: $\Omega(\neg A|B) = P(\neg A|B) \,/\, P(A|B)$. Dann erhalten wir:

$$\Omega(\neg h_i|e^n\&b\&c^n) = \sum_{j\neq i} Q_{j,i}(e^n) \cdot [P(h_j|b) \,/\, P(h_i|b)]$$

Das verlangt natürlich, dass die $P(h_i)>0$ sind, damit die hinteren Brüche existieren. Damit können wir dann auch die Endwahrscheinlichkeit von h_i bestimmen:

Bayessches Theorem anhand von Quoten
$P(h_i|e^n\&b\&c^n) = 1/[1+\Omega(\neg h_i|e^n\&b\&c^n)]$

Wir können obendrein erkennen, was mit den verschiedenen Hypothesen beim Updaten geschieht. Gehen die $Q_{j,i}(e^n)$ gegen Null für alle $j\neq i$, dann geht auch deren Summe gegen Null (zumindest für endliche Mengen von Hypothesen) und damit wächst die Endwahrscheinlichkeit der wahren Hypothese h_i gegen 1 und damit auch die Wahrscheinlichkeit unserer Hypothese h_i für eine wachsende Datenmenge.

Dieses Grenzwertverhalten des QUF ist natürlich ebenfalls für andere Ansätze von Bedeutung, die die Likelihoods als wesentliche Faktoren für die Theorienbewertung erachten, wie etwa das abduktive Schließen und entsprechende Kohärenzansätze. In diesen Ansätzen standen Vergleiche der Hypothesen im Hinblick auf ihre Erklärungsleistung für die Daten ganz im Vordergrund und die Erklärungsleistung war wiederum an die auftretenden Likelihoods gekoppelt. Das Likelihood-Quotienten-Konvergenz-Theorem besagt dafür gerade, dass die Likelihood-Quotienten im Durchschnitt für falsche Hypothesen im Vergleich zu wahren gegen Null gehen, was also zumindest im Durchschnitt sicherstellt, dass wir mit diesem Verfahren die wahren Hypothesen ermitteln werden. Voraussetzung ist allerdings, dass wir die wahren Hypothesen überhaupt mit im Blick haben und die Daten konstant durch dieselben Mechanismen (wie sie in Hypothese h_i beschrieben werden) erzeugt werden.

Hawthorne (2005, 2008, 2010) verweist außerdem darauf, dass wir etwa als induktive Logiker mit einer ganzen Menge P_α, P_β,… von Ausgangswahrscheinlichkeiten starten können, die unsere anfängliche Vagheit in den Glaubensgraden ausdrücken; dann sorgt die Konvergenz

dafür, dass diese Menge immer kleiner wird und gegen bestimmte Endwahrscheinlichkeiten konvergiert (s.a. Kap. 5.5.9). Selbst wenn also das Indifferenzprinzip oder das MaxEnt-Prinzip keine eindeutigen Ergebnisse liefert, mag es immer noch hilfreich sein, mit einer Menge von Ausgangswahrscheinlichkeiten unser Updaten zu beginnen, die alle zumindest eine gewisse Begründung haben und nicht völlig willkürlich gewählt wurden.

Der nächste Absatz wird das noch einmal unterstreichen. Er zeigt aber zugleich die Gefahren einer völlig subjektiven Wahl der Ausgangswahrscheinlichkeiten auf. Jedenfalls können diese Ausgangswahrscheinlichkeiten bei unterschiedlichen epistemischen Subjekten anderenfalls stark variieren, aber wenigstens bei den Likelihoods sollte es einen Konsens unter den Bayesianern geben, dass wir die mit Hilfe der Likelihoodanbindung entsprechend den Aussagen der Theorie selbst wählen. Dann wären sie zumindest für alle Bayesianer dieselben und ihre Konvergenz wird durch das Likelihood-Konvergenz-Theorem sichergestellt.

Wie das Konvergenzverhalten dabei vom empirischen Abstand der Hypothesen (davon wie unterschiedlich ihre Vorhersagen für bestimmte Daten sind) abhängt, sehen wir, wenn wir unser obiges Zahlenbeispiel mit anderen Hypothesen durchrechnen. Es dürfte aber auch intuitiv klar sein, dass benachbarte Hypothesen schwieriger durch Daten zu unterscheiden sind (vgl. a. Kap. 5.3.12). Es sei weiterhin die Hypothese H_2 wahr und H_1 ihr falscher Konkurrent. Im Falle (a) mit $P(\text{Kopf}|H_1) = 0{,}4$ und $P(\text{Kopf}|H_2) = 0{,}3$ und (b) $P(\text{Kopf}|H_1) = 0{,}31$ und $P(\text{Kopf}|H_2) = 0{,}3$ erhalten wir für den QUF die folgenden Werte:

Ein konkretes Zahlenbeispiel (Fortsetzung):
E_1 = 4 Köpfe von 10; E_2 = 33 Köpfe von 100;
E_3 = 290 Köpfe von 1000

Resultat	QUF	
	(a)	(b)
E_1	1,05	1,25
E_2	0,43	1,13
E_3	$5{\cdot}10^{-12}$	0,49

Tabelle 5.8: Probleme durch benachbarte Hypothesen

Im Falle (a) sehen wir also zumindest bei 1000 Würfen noch ein deutliches Resultat (eine probabilistische Falsifikation von H_2, wenn H_1 und H_2 nicht zu extreme Ausgangswerte haben), aber im Falle (b) findet auch bei 1000 Würfen mit nur 290-mal Kopf kaum noch eine Verschiebung statt. Erst wenn wir 10000-mal werfen und genau 3000 Köpfe erhalten, sinkt der QUF auf ca. 0,1. Damit ließen sich jetzt die jeweiligen Endwahrscheinlichkeiten berechnen, wenn wir für H_1 und H_2 bestimmte Startwahrscheinlichkeiten wählen würden. Es sollte in jedem Fall deutlich geworden sein, wie stark der QUF vom Abstand der Hypothesen abhängt. Trotzdem bleibt natürlich das Konvergenzresultat bestehen, dass sich die wahre Hypothese letztlich mit hoher Wahrscheinlichkeit durchsetzt, solange sie noch empirische Unterschiede zu ihren Konkurrenten aufweist. Die Frage ist dann nur, was das Konvergenzresultat in der Praxis, in der wir es immer nur mit einer begrenzten Anzahl von Daten zu tun haben, tatsächlich aussagt. Das soll der nächste Abschnitt ein wenig beleuchten.

5.6.4 Zur Bedeutung der Konvergenztheoreme: Eine Glosse

Die Konvergenzresultate sind zunächst nur theoretische Ergebnisse, die auf lange Sicht gelten, die aber womöglich sehr lang sein kann. Was besagen sie für die Praxis? Dafür ist das Likelihood-Konvergenz-Theorem hilfreich, weil sich hier die Konvergenz besser abschätzen lässt. Zunächst ist klar, dass wir bei hinreichend getrennten Hypothesen und doch einigermaßen vielen Experimenten zwischen H_1 und H_2 etwa in unserem Basisfall recht gut unterscheiden können, sollten H_1 und H_2 nicht allzu extreme Ausgangswahrscheinlichkeiten aufweisen. Das ruft geradezu die induktiven Logiker auf den Plan, die sich von den subjektiven Bayesianern vor allem dadurch unterscheiden, dass sie für die Ausgangswahrscheinlichkeiten mit dem Indifferenzprinzip und seinen Abkömmlingen ein rationales Verfahren vorschlagen, wie man zu sinnvollen Startwahrscheinlichkeiten gelangen sollte. Das dient dazu, ausgefallene Startwahrscheinlichkeiten als irrational auszuschließen. Sollte ein subjektiver Bayesianer da nicht mitziehen, helfen ihm schließlich auch die Konvergenzresultate nicht mehr. Das möchte ich mit einer kurzen Glosse verdeutlichen:

Die Mobilfunkfirma Handyflat beauftragt den Wissenschaftler Professor Wichtig, einen bekannten (subjektiven) Bayesianisten, eine Studie zu möglichen Gefahren des uneingeschränkten Handy-Konsums durchzuführen. Wissenschaftler Wichtig hält überhaupt nichts von der Theorie T (T besagt: *10 Jahre Handykonsum von mindestens einer Stunde verursacht zwingend Hirnerweichung*.), die die Umweltaktivisten in den Raum stellen. Als Bayesianist muss er für T zunächst eine Vorher-Wahrscheinlichkeit festlegen. Er überlegt sich, dass die nicht so hoch sein sollte, da er T für völlig falsch hält. Da kommen verschiedene Zahlen in Frage: 1/1000, 1/1000000, 10^{-10}, 10^{-1000}, $10^{-1000000}$, $10^{-(10\hat{}100)}$, $10^{-(10\hat{}100000)}$, und viele weitere. Dabei soll $10^{-(10\hat{}100)}$ bedeuten, dass hier die Eins geteilt wird durch die recht große Zahl bestehend aus einer Eins mit 10^{100}=10^100 Nullen. (Diese großen Zahlen haben sogar Namen, denn 10^{100} wird *Googol* mit englischer Aussprache genannt und 10^{Googol}=$10^{(10\hat{}100)}$ wird als *Googolplex* bezeichnet. Manchmal erhalten selbst weitere Zahlen wie $10^{\text{Googolplex}}$ noch weitere Namen etc.) Allerdings lassen sich auf diesem Wege schnell noch größere Anzahlen und damit kleinere Wahrscheinlichkeiten entwerfen.

Da Professor Wichtig ein eher bescheidener Mensch mit bescheidenen Mathematikkenntnissen ist, entscheidet er sich für die moderate Einschätzung P(T)= $10^{-(10\hat{}1000)}$, die T noch eine Chance lassen soll, obwohl er sie eigentlich für völlig falsch hält. Diese Grundeinschätzung mag irgendwie auch mit seinem Auftraggeber »Handyflat« zusammenhängen, der seinen Kunden den exzessiven Mobilfunkkonsum anrät. Doch als guter (subjektiver) Bayesianist weiß er, dass er völlig frei in seiner Wahl einer Ausgangswahrscheinlichkeit ist. Allerdings hält er sich strikt an das Dogmatismusverbot, denn wir wollen jeder Hypothese − und sei sie noch so absurd wie die Theorie T − eine Chance geben, sich doch noch zu beweisen. Schließlich besagen die Konvergenztheoreme, dass sich die wahre Theorie dann schon durchsetzen wird.

Er kann seine Meinung später natürlich noch ändern und mit einer anderen Vorher-Wahrscheinlichkeit starten, die noch sehr viel ungünstiger für T ausfällt, sollte sich das als hilfreich erweisen. Ein Mathematiker könnte ihn beraten, wie man besonders große Zahlen für den Nenner finden kann. Da müsste sich doch noch etwa machen lassen. Sein Auftraggeber bietet sogar an, ihm solche großen Zahlen zu besorgen,

doch Professor Wichtig als unabhängiger und objektiver Wissenschaftler lehnt das Angebot natürlich dankend ab.

Doch nun frisch ans Werk: Professor Wichtig stellt erste Experimente an. Nehmen wir zunächst an, T sei deterministisch gemeint. Für jeden fleißigen Handynutzer d_i erwarten wir also zwingend eine Hirnerweichung nach 10 Jahren. Da die sonst nicht so häufig auftritt, gibt er ihr jeweils die Startwahrscheinlichkeit 1/1000. (Als subjektiver Bayesianist hätte er ihr natürlich auch die Wahrscheinlichkeit 0,9 geben können, was seine Aversionen gegen T besser abgesichert hätte, aber er möchte eben eine ganz seriöse Auswertung seiner Studie durchführen.) Er verfolgt nun über 10 Jahre die Geschicke von 1000 Handynutzern. Leider enden die alle mit Hirnerweichung (E). Was bedeutet das für T? Dazu können wir den bayesianischen Update-Faktor großzügig abschätzen. Wir arbeiten einfach weiterhin mit dem Wert $P(d_i)=1/1000$ für jeden weiteren Handynutzer, obwohl diese Wahrscheinlichkeiten eigentlich steigen sollten. Aber Profesor Wichtig möchte der Theorie T wie gesagt noch eine Chance geben. Damit gilt:

$$P(E|T)/P(E) \leq 1000^{1000} = (10^3)^{1000} = 10^{3000}$$

Was wird damit aus der Wahrscheinlichkeit unserer Theorie T? Auch die lässt sich nun abschätzen. Dazu betrachten wir den Nenner von $P^+(T)$ bei einem Zähler von 1:

$$\text{Nenner} = 10^{(10\hat{}1000)-3000} \geq 10^{(10\hat{}999)},$$
$$\text{denn } 10^{1000}-3000 \geq 10^{999}$$

Damit ist $P^+(T) \leq 10^{-(10\hat{}999)}$, d.h., die Theorie ist weiterhin extrem unwahrscheinlich und wir sollten nach Ansicht der Bayesianer praktisch jede Wette dagegen annehmen, obwohl die experimentellen Daten so eindeutig waren und wir ihre Startwahrscheinlichkeit bereits mit einer unvorstellbar großen Zahl multipliziert haben. 100-mal sechs Richtige im Lotto zu erzielen ist jedenfalls wesentlich wahrscheinlicher, als dass die Theorie T nun wahr sein könnte. Es gibt also praktisch kein Risiko für eine Hirnerweichung selbst bei extremem Handykonsum nach den Angaben von Professor Wichtig. Das kann er gleich so der Presse melden. Die Details des Experiments müssen natürlich nicht

unbedingt veröffentlicht werden, denn ein weniger guter Wissenschaftler als Professor Wichtig (etwa ein klassischer Statistiker) könnte daraus sonst schnell »falsche« Schlüsse ziehen.

Der Auftraggeber Handyflat ist mit dem Ergebnis zufrieden und genehmigt gleich noch eine weitere Studie. Schließlich möchte man ja auf Nummer sicher gehen. Allerdings weisen sie gleich darauf hin (ganz bayesianistisch gedacht), dass man auch die Wahrscheinlichkeiten für eine Hirnerweichung $P(d_i)$ langsam deutlich nach oben anpassen sollte aufgrund der vielen neuen Erfahrungen. Trotzdem rechnet Professor Wichtig weiterhin mit 1/1000, schließlich will er sich keine Beeinflussung durch die Industrie nachsagen lassen. Wichtig stellt sogar weitere 1000 Studien mit insgesamt 1000 000 Teilnehmern an. Alle versterben leider an Hirnerweichung (E*). Handyflat ist ein wenig besorgt, ob das auf den Handykonsum ein schlechtes Licht werfen könnte, doch die genauen Berechnungen des Professors zeigen, dass es überhaupt keinen Grund zur Sorge gibt. Es gilt zwar zunächst:

$$P(E^*/T)/P(E^*) \leq 1000^{1000000} = (10^3)^{1000000} = 10^{3000000},$$

der Update-Faktor wird inzwischen nur noch durch eine 1 mit 3 Millionen Nullen dahinter abgeschätzt, aber es gilt natürlich trotzdem:

$$\text{Nenner} = 10^{(10\hat{}1000)-3000000} \geq 10^{(10\hat{}999)},$$
$$\text{denn } 10^{1000}-3000000 \geq 10^{999}$$

Das heißt, es gilt weiterhin $P^+(T) \leq 10^{-(10\hat{}999)}$ und auch weitere Studien mit realen Personen können daran nichts ändern. Also hat sich an der Grundaussage von Professor Wichtig, Hirnerweichung aufgrund von Handys sei extrem unwahrscheinlich, ja sogar praktisch ausgeschlossen, nichts geändert. Jetzt ist auch der verantwortliche Politiker A überzeugt, dass Mobiltelefonieren ungefährlich ist, denn das Ergebnis der gründlichen und riesigen Studie war ja schließlich die extrem kleine Wahrscheinlichkeit für die Befürchtung T. Das hatte er ohnehin schon vermutet, schließlich hatte ihm die Firma Handyflat das längst versichert. Da waren die Wahlkampfspenden von der Firma Handyflat eigentlich ganz unnötig, aber natürlich trotzdem willkommen.

So beruhigend die Auskünfte unseres bayesianischen Wissenschaftlers Professor Wichtig auch sind, so gibt es doch immer Miesmacher aus anderen Lagern (etwa der klassischen Statistik), die Wichtig seine Aufträge nicht gönnen. Sie kommen mit folgendem miesmacherischen Argument: Alle harten Daten, die auf dem Tisch liegen, geben keinen Grund zur Beruhigung. Schließlich sind eine Million Versuchspersonen alle an Hirnerweichung erkrankt. Das *allein* sollte doch für uns zählen. Einzig die bayesianistische Einschätzung von T durch Professor Wichtig hätte die beruhigende Note ins Spiel gebracht. Die sei aber nur seine persönliche Sache und hätte nichts mit den objektiven Daten und ihrer wissenschaftlichen Auswertung zu tun. Die Daten gäben viel mehr Anlass zu größter Sorge. Wir sollten hier die Daten im Hinblick auf die Theorie bewerten, und Professor Wichtig sollte seine persönlichen Einschätzungen dazu für sich behalten.

Auch Politiker B zeigt sich beunruhigt. Er geht ganz naiv an die Sache heran, weil er eben nichts vom Bayesianismus versteht. Er sagt: »Lassen sie uns auf die Daten schauen. 1 Million Handy-Telefonierer wurden untersucht und alle erlitten eine Hirnerweichung. Gibt das nicht doch Grund zur Sorge? Außerdem gebe ich zu bedenken, dass auch Politiker A ein langjähriger Vieltelefonierer ist.« Nun müssen wir B wohl in die Feinheiten der bayesianischen Datenauswertung einweihen, um ihn wieder zu beruhigen. Die Firma Handyflat bezahlt ihm und seiner Frau eine Ausbildung im bayesianischen Updaten auf den Bahamas, damit er nun auch die Ungefährlichkeit des Handykonsums einsehen kann. Das ist eben alles nur eine Frage der richtigen Einstellung bzw. Startwahrscheinlichkeiten.

Auch schreckliche Umweltfreaks zeigen sich trotz der eindeutigen wissenschaftlichen Belege störrisch. Sie verweisen darauf, dass dieselbe Abschätzung gelten würde, wenn wir 10 Milliarden Menschen getestet hätten (also mehr als momentan auf der Erde vorhanden sind) und alle an Gehirnerweichung gestorben wären. Sollte uns das nicht zu denken geben?

So langsam könnten uns Zweifel beschleichen, ob die bayesianische Herangehensweise in diesem Fall so ganz adäquat ist. Die schlichte Betrachtung der Daten hat auch so ihre Vorzüge. Leider scheinen wir also diese Überlegung der Miesmacher nicht so ganz beiseite wischen

zu können, wenn wir nicht schon strenggläubige (subjektive) Bayesianer sind. Sobald wir im Glauben zweifeln, denken wir immer wieder in ganz naiver Weise: Der *subjektive* Bayesianismus scheint uns keine so große Hilfe in diesem Fall gewesen zu sein. Der klassische Statistiker sagt jedenfalls, dass wir nur auf die Likelihoods etwa im Vergleich zu denen anderer Hypothesen schauen sollten. Hätten wir als Konkurrenzhypothese die Theorie T*, *dass doch zumindest 1% aller Handynutzer keine Hirnerweichung nach ausgiebigem Handykonsum haben würden*, so wäre der Likelihood-Quotient $P(E^*|T)/P(E^*|T^*) = 1{,}6 \cdot 10^{4365}$. Damit sprechen die Daten in gigantischer Weise für T in Relation zu T*, obwohl T* auch schon relativ hohe Quoten von Hirnerweichung prognostiziert. Das ist vielleicht die bessere Art und Weise die Daten auszuwerten. Startwahrscheinlichkeiten sind hier nicht mehr vonnöten. So denken jedenfalls die Likelihoodisten, die nun bald zu Wort kommen sollen.

Oder wir müssen die *objektiven* Bayesianer zu Wort kommen lassen, die hier einwenden, dass die Startwahrscheinlichkeiten zu willkürlich und in unserem Fall sehr einseitig gewählt wurden. Sobald wir mit einem Indifferenzprinzip an die Sache herangehen, kann das nicht mehr passieren. Selbst wenn so ein Indifferenzprinzip keine eindeutigen Ergebnisse liefert und wir hinterher mit einer ganzen Menge von Startwahrscheinlichkeiten arbeiten müssten, würden keine so einseitigen Ergebnisse auftreten. Die Startwahrscheinlichkeiten wären nicht so extrem. Im günstigsten Fall wäre die Wahrscheinlichkeit von T zu Beginn einfach ½ und damit wäre sie nach dem Updaten praktisch 1. In jedem Fall sollte klar werden, dass die Daten unsere Hypothese T stark bestätigen.

Dazu müssen wir uns wieder dem Thema der Bestätigung im bayesianischen Rahmen zuwenden. Die bereits angegebenen Maße der Bestätigung sind allerdings auch nicht ganz leicht zu interpretieren. Die Werte des Differenzmaßes d bleiben hier klein, während die für das Ratio-Maß r und das Likelihoodmaß l groß werden. Genau genommen würde uns bereits ein Vergleich der Bestätigung für die Hypothese T im Vergleich zur Gegenhypothese, wonach Handykonsum nicht zur Hirnerweichung führt, genügen. Womöglich sind hier unterschiedliche Vorstellungen von Bestätigung im Spiel, die es zunächst zu unterscheiden

gilt. Dem widmen wir uns im nächsten Abschnitt und in den weiteren Kapiteln.

5.6.5 Formen der Bestätigung

Nachdem wir inzwischen einige Ansätze kennengelernt haben, können wir wieder auf die Frage zurückkommen, was mit einer Bestätigung von Theorien etwa durch Daten oder anderes Hintergrundwissen gemeint ist. Das entspricht den Hypothesentests in der klassischen Statistik. Dabei lassen sich verschiedene Grundideen unterscheiden und auf unterschiedliche Weise durch Wahrscheinlichkeiten modellieren.

Zunächst haben wir die *relationale Form der Bestätigung* B(H,E), nach der bestimmte Daten D eine Hypothese H (lokal) ein Stück weit bestätigen. Unser Hintergrundwissen K wird dabei normalerweise auch bereits eine Rolle spielen, so dass wir eigentlich eine entsprechende dreistellige Relation vor uns haben: B(H,E;K). Damit ist aber noch nicht gesagt, ob es nicht in unserem Hintergrundwissen viele andere Belege gibt, die gegen H sprechen. Es soll nur die relationale Beziehung zwischen E und H ausgedrückt werden, die allerdings selbst auch von weiterem Hintergrundwissen abhängen kann. Die Fingerabdrücke des Verdächtigen auf der Tatwaffe (E) sprechen dafür, dass er auch der Täter ist (H), aber es mag daneben andere Indizien oder sogar ein Alibi geben, die stark dagegen sprechen. Dass die Fingerabdrücke gegen eine Person sprechen, hängt wiederum von unserem Hintergrundwissen K ab, das besagt, dass Fingerabdrücke nahezu einzigartig für eine bestimmte Person sind. Bayesianer haben oft diese relationale Form der Bestätigung im Auge, wenn sie von *inkrementeller Bestätigung* sprechen und damit meinen, dass $P(H|E\&K) > P(H|K)$ ist. Das Hinzufügen von E zu unserem Hintergrundwissen erhöht die Wahrscheinlichkeit von H. Das ist oft die grundlegende Bestätigungsbeziehung, mit der Bayesianer arbeiten. Und auch beim hypothetisch-deduktiven Bestätigungsbegriff scheint es sich vornehmlich um die relationale Bestätigung zu handeln.

Eine andere Form von Bestätigung finden wir in der *absoluten Bestätigung*. Dabei geht es darum, ob eine Hypothese H insgesamt durch unser Hintergrundwissen bestätigt wird, so dass wir sie akzeptieren dürfen: A(H,K). Auch hierbei geht es wieder um eine Art von relationaler

Bestätigung, aber diesmal ist der Relationspartner gleich das gesamte Hintergrundwissen. Dabei kann man wiederum fragen, was alles zu K gehört. Handelt es sich tatsächlich nur um unser richtiges Wissen im strikten Sinne oder meinen wir damit bloß die Überzeugungen, die von uns zurzeit akzeptiert werden? Diese Debatte möchte ich aber hier nicht weiter verfolgen. Jedenfalls setzt die absolute Bestätigung bereits voraus, dass wir eine Abwägung unterschiedlicher Daten oder anderer Theorien, die für oder gegen eine Hypothese sprechen, in Form einer holistischen Gesamtbewertung durchführen können.

Der Bayesianer wird diese absolute Bestätigung am ehesten durch eine Schwellenwertkonzeption wiedergeben, die er aber nicht so gerne akzeptieren möchte, wie wir oben gesehen haben. Demnach gibt es einen Schwellenwert w>½, ab dem wir H akzeptieren: $P(H|K) \geq w$. Es kann natürlich auch sein, dass der Bayesianer sagt, er benötige eine solche absolute Bestätigungsbeziehung überhaupt nicht. Dann wird er sie ganz durch die Glaubensgrade ersetzen, die eine neue Gesamtbewertung abgeben sollen. Bereits die eliminative Induktion ging einen wichtigen Schritt in Richtung dieser Gesamtbewertung und natürlich vor allem der Kohärenzansatz.

Zu den genannten Beziehungen kann man noch die Frage aufwerfen, ob es eine besondere *Relevanzbeziehung* etwa kausaler Art oder anderer Natur zwischen H und E geben sollte. Wir hatten gesehen, dass die rein logische Ableitbarkeit oder auch die reine Wahrscheinlichkeitserhöhung noch wenig über einen inhaltlichen Zusammenhang aussagen können. Das abduktive Schließen erforderte gegenüber der hypothetisch-deduktiven Bestätigung, dass eine Erklärungsbeziehung vorliegt. Genau das verlangt auch Peter Achinstein (2001) als zusätzliche Bedingung, da wir seiner Meinung nach sonst antiintuitive Ergebnisse zu gegenwärtigen haben, auf die ich im übernächsten Unterkapitel eingehen werde. Hier sind also Zusatzbedingungen denkbar, die u.a. auch darin bestehen könnten, dass die objektive Likelihood P(E|H) einen bestimmten Wert überschreitet.

Des weiteren finden wir eine Reihe *komparativer bzw. relativer Bestätigungsbegriffe*, die etwa besagen, dass E eine bessere Bestätigung für H_1 gegenüber H_2 darstellt: $B(H_1,H_2,E;K)$. Echte Likelihoodisten wie Royall (1997) oder Edwards (1992) argumentieren dafür, dass wir

nur solche vergleichenden Aussagen machen dürfen, da die anderen Formen von Bestätigung nicht objektiv explizierbar sind, sondern etwa auf der Annahme von subjektiven Vorher-Wahrscheinlichkeiten beruhen. Kohärenztheoretiker würden allerdings sagen, dass wir dann zumindest noch eine Gesamtbewertung anschließen müssen, die doch wieder zu etwas Ähnlichem wie einem absoluten Bestätigungsbegriff führt. Jedenfalls hatten wir das abduktive Schließen (und auch die eliminative Induktion) ebenfalls wesentlich auf solche komparativen Bestätigungen gestützt. Likelihoodisten würden also $B(H_1,H_2;E;K)$ dadurch ausdrücken wollen, dass der entsprechende Likelihoodquotient $P(E|H_1)/P(E|H_2)$ größer als 1 ist. Bayesianer würden dagegen eher die Nachher-Wahrscheinlichkeiten miteinander vergleichen und vor allem verlangen, dass $P(H_1|E)/P(H_2|E) > 1$ ist. Das bewertet allerdings nur, welche Veränderungen beim jetzigen Glaubensgrad von H_1 und H_2 durch E noch auftreten und unterliegt somit wieder dem Problem der alten Evidenz (s.u.). Vertreter der Abduktion würden verlangen, dass H_1 E besser *erklärt* als H_2, und bei der Beurteilung dieser Beziehung kann sehr wohl auch der Likelihoodquotient eine wichtige Rolle spielen (vgl. aber Kap. 5.5.18).

Außerdem kann man noch eine andere Form von Vergleich heranziehen, der auch für eine Verrechnung verschiedener Daten hilfreich ist. Wir könnten explizieren, wann ein Datum E_1 eine bessere Bestätigung für H liefert als ein Datum E_2: $B(H;E_1,E_2;K)$. Das können Bayesianer z.B. durch ihre allerdings kontroversen und recht unterschiedlichen Maße für die Bestätigungsbeziehung explizieren. Der Likelihoodist wird dagegen wieder nur auf die objektiven Likelihoods setzen und den Quotienten $P(E_1|H)/P(E_2|H)$ betrachten und fragen, ob der größer als 1 ist.

Man erkennt die unterschiedlichen Formen der Bestätigung, wenn man auf Hempels (1945) Forderungen für Bestätigungsbeziehungen schaut, die wir schon in ausführlicherer Form in Kap. 3.4 im Zusammenhang mit dem hypothetisch-deduktiven Schließen erörtert haben. Hempel trennt hier nicht sauber, welche Art der Bestätigungsbeziehung er explizieren möchte und erhält daher eine »ungesunde« Mischung von Forderungen:

Hempels Prinzipien

(1) A generalization of the form 'All F are G' is confirmed by the evidence that there is an individual that is both F and G.
(2) A generalization of that form is also confirmed by the evidence that there is an individual that is neither F nor G.
(3) The hypotheses confirmed by a piece of evidence are consistent with one another.
(4) If E confirms H then E confirms every logical consequence of H.

Auf welche Bestätigungsbeziehung beziehen sich die Forderungen jeweils? Für Forderung (1) ist offensichtlich, dass nicht die absolute, sondern nur eine relationale und damit eher eine inkrementelle Bestätigung der Hypothesen gemeint sein kann. Dasselbe gilt für (2), wenn wir überhaupt akzeptieren wollen, dass in diesem Fall eine Bestätigung vorliegt. Hempel biss in diesen sauren Apfel, um seine Bestätigungskonzeption mit der Rabenparadoxie zu versöhnen, aber mir scheint diese Lösung nicht wirklich akzeptabel (s.u.). Die Forderung (3) ist dagegen nur sinnvoll für die absolute Bestätigung. Zwei Hypothesen, die im absoluten Sinne (Wahrscheinlichkeit > ½) durch ein E bestätigt wurden, können nicht mehr direkt miteinander inkonsistent sein, für eine inkrementelle Bestätigung ist das dagegen kein Problem. Nehmen wir folgendes einfache Beispiel dafür: Wir haben vier Karten gezogen und entwickeln dazu zwei Hypothesen. H_1: *Es handelt sich um vier Asse.* H_2: *Es handelt sich um viermal Kreuz.* Dann schaue ich mir eine Karte davon an und es handelt sich um ein Kreuzass (E). Dann spricht E für beide inkompatiblen Hypothesen im Sinne der Forderung (1).

Auch die Konsequenzbedingung (4) ist nur für absolute Rechtfertigungen sinnvoll. Es sei $H \equiv H_1 \& H_2$ mit unseren Hypothesen H_1 und H_2 aus dem letzten Beispiel. Dann bestätigt das Herzass (E) zwar die Hypothese H im Sinne der Forderung (1), aber die Konsequenz H_2 aus H wird keineswegs durch E bestätigt. Anders ist das für die absolute Bestätigung. Wenn P(H) > w ist, so gilt das auch für jede logische Konsequenz von H.

Wir erkennen hier leider, dass wir mit »Bestätigung« Unterschiedliches meinen können und das anscheinend auch in wissenschaftstheoretischen Überlegungen wiederfinden, ohne dass die Verwendungsweisen getrennt würden. Es ist schwer zu sagen, ob einer der Begriffe basaler

ist als der andere. Die relationale Bestätigung ist eine stärker lokale Beziehung, auf die wir die absolute Bestätigung vielleicht zurückführen könnten, wenn wir denn wüssten, wie sich die vielen relationalen Bestätigungen (bzw. Abschwächungen) miteinander verrechnen ließen. Der Bayesianer verlangt einfach, dass wir eine gemeinsame Wahrscheinlichkeitsverteilung für alle Hypothesen und alle potentiellen Daten aufweisen, und sagt aber nicht, wie wir die lokalen Beziehungen unabhängig davon betrachten können. Der Holismus wird sozusagen von Anfang an in die Glaubensgrade mit eingebaut, die immer eine Gesamtschau der Wahrscheinlichkeiten für alle in Frage kommenden Aussagen darstellen. Daher passt die relationale Sichtweise der Bestätigung eigentlich nicht so besonders gut in den Bayesianismus hinein. Das zeigte sich bereits im Problem der alten Evidenz. Vielleicht sollte der Bayesianer sich doch stärker mit der absoluten Bestätigung auseinandersetzen, wo dieses Problem nicht so auftritt. Das hieße allerdings, dass fast alle Bayesianer in dieser Frage bisher einer Selbsttäuschung unterlägen.

Jedenfalls hatte ich bereits dafür argumentiert, dass wir die absolute Bestätigung benötigen und auch die Bayesianer ein entsprechendes Konzept entwickeln sollten. Die Schwierigkeiten, zu ermitteln, was wir mit »Bestätigung« meinen und wie sich das mit Wahrscheinlichkeiten in Verbindung bringen lässt, finden sich auch in den Einwänden von Peter Achinstein wieder, der daraus u.a. den Schluss zieht, dass wir mehr Gewicht auf eine Relevanzbeziehung legen müssen. Das hatten wir schon im Falle der hypothetisch-deduktiven Schlussfolgerungen gesehen und Bayesianer halten sich viel darauf zugute, dass sie diese Schlüsse mit ihrem Apparat reproduzieren. Dann erben sie leider auch dessen Probleme und sollten vielleicht ebenfalls zusätzlich eine stärkere Relevanzbeziehung in ihre Bestätigungstheorie einbauen.

5.7 Objektiver Bayesianismus und induktive Logik

5.7.1 Carnaps induktive Logik

Insbesondere Rudolf Carnap (u.a. in 1950) hat versucht, eine relationale Form der Bestätigungsbeziehung durch eine Induktionslogik zu explizieren. Es ging ihm darum, durch c(H,E) das Ausmaß anzugeben, in dem

die Aussage H durch die Aussage E objektiv gestützt wird. Das Projekt wird von vielen inzwischen als gescheitert angesehen, aber von anderen Autoren wie Patrick Maher (2010) doch mit einigen guten Argumenten weiter verfolgt. Da wir auf möglichst weitgehende Objektivierungen im Bayesianismus angewiesen sind, weil wir sonst den Problemen des Bayesianismus zum Opfer fallen können, denen schon Prof. Wichtig unterlag, sollten wir noch einmal nachsehen, wo die Probleme der induktiven Logik liegen, um zumindest das zu retten, was davon zu retten ist.

Carnap wollte das Maß c (für »confirmation«) möglichst a priori bestimmen und die Regeln – ähnlich wie die Regeln der deduktiven Logik – aus einfachen Überlegungen ableiten. Dazu setzte er vor allem auf das *Indifferenzprinzip*, dessen Problematik wir bereits kennengelernt haben. Er relativierte es auf eine bestimmte prädikatenlogische Sprache L mit Prädikaten $F_1,\ldots,F_n$ und Individuenkonstanten $a_1,\ldots,a_r$ und entwarf dazu seine induktive Logik. Die Grundidee ist dabei, dass die Sprache uns die Menge aller möglichen Weltzustände liefert, die sich in diesem Ansatz aus der Menge aller möglichen Kombinationen der Prädikate angewandt auf die Individuen zusammensetzt. Für einstellige Prädikate F_i stellt jede der folgenden komplexen Aussagen (Vollkonjunktionen) einen der $2^{n \cdot r}$ möglichen Weltzustände dar:

$$\pm F_1(a_1) \,\&\ldots\&\, \pm F_1(a_r) \,\&\, \pm F_2(a_1) \,\&\ldots\&\, \pm F_n(a_r)$$

Dann bestimmen wir für c(H,E), wie hoch der Anteil der Weltzustände ist, in denen H&E gilt, innerhalb der Menge der E-Weltzustände. Das heißt, wir betrachten nur noch die E-Zustände und fragen uns, wie viele davon auch H-Zustände sind, und notieren das als Quotienten:

$$c(H,E) := m(H\&E) \,/\, m(E)$$

Dabei soll m(E) die Anzahl der Elemente in E bezeichnen bzw. im allgemeineren Fall ein Maß für den Inhalt der Menge E angeben. Wenn also die Beschränkung auf E-Zustände die Quote für H-Zustände erhöht, dann bestätigt E die Hypothese H. Dies lässt sich vermutlich am einfachsten an einem simplen Beispiel verstehen. Meine Hypothese H sei, dass der Würfelwurf eine 5 ergibt. Wenn ich dann erfahre (E),

dass eine ungerade Zahl gewürfelt wird, so erhöht sich für einen fairen Würfel die Wahrscheinlichkeit für H von 1/6 auf 1/3, weil unser Datum E nur noch die drei Möglichkeiten übrig lässt, von denen eine gerade H darstellt.

Für eine recht einfache Sprache können wir das Projekt beispielhaft durchführen. Wählen wir eine einfache Welt mit nur einem Prädikat F und drei Objekten a, b, c, dann gibt es 8 verschiedene Zustände, die durch die 8 möglichen Vollkonjunktionen unserer atomaren Aussagen gegeben sind:

Weltzustände

1. Fa & Fb & Fc	2. ¬Fa & Fb & Fc
3. Fa & ¬Fb & Fc	4. Fa & Fb & ¬Fc
5. ¬Fa & ¬Fb & Fc	6. ¬Fa & Fb & ¬Fc
7. Fa & ¬Fb & ¬Fc	8. ¬Fa & ¬Fb & ¬Fc

Auf den Weltzuständen wird als Erstes eine Gleichverteilung angenommen, so dass jedem Zustand das Gewicht 1/8 gegeben wird. Damit können wir alle Wahrscheinlichkeiten, die uns in diesem Rahmen interessieren, ausrechnen. Zunächst ist klar, dass c(Fa) = 1/2, denn das entspricht der Wahrscheinlichkeit, dass Fa bei einer leeren Information bzw. Tautologie t wie etwa Fb∨¬Fb auftritt. Für c(Fa) = c(Fa|t) ist Fa in 4 von 8 Zuständen wahr. Unsere Hypothese h sei nun: h = *Alle Objekte sind F.* Da h nur im Zustand 1 gilt, ist c(h) = 1/8 und c(h|Fa) = 1/4, da bei den verbliebenen 4 Zuständen (1,3,4,7) nur einer ein h-Zustand ist. Entsprechend ist c(h|Fa&Fb) = 1/2. Das sieht alles recht intuitiv aus, wenn wir das dabei eingesetzte Indifferenzprinzip einmal akzeptiert haben.

Leider erweist sich diese einfache Konstruktion als nicht wirklich geeignet, um *induktiv* zu schließen oder Hypothesen induktiv zu rechtfertigen. So gilt in jedem Fall c(Fa|Fb) = c(Fa), d.h., die induktive Logik weist nicht die *Induktionseigenschaft* auf, über die wir in Kapitel 1.6.2 und 5.3.10 bereits gesprochen haben, die sich in 5.3.10 für bestimmte Hypothesen als Konsequenz aus dem Dogmatismusverbot zu leicht ableiten ließ. Wenn wir bestimmte Instanzen unserer Hypothese h bestätigen, wird diese dadurch immer nur deduktiv ein Stück weit bestätigt; und zwar wird nur der Teil der Hypothese, der über genau diese Instanzen spricht,

deduktiv bestätigt, die Hypothese erscheint aber im Übrigen nicht plausibler als vorher. Selbst wenn sich also bereits mehrere Objekte als F erwiesen haben (oder auch gerade nicht erwiesen haben), bleibt die Wahrscheinlichkeit für ein noch nicht getestetes Objekt dafür F zu sein gerade 1/2. Induktives Lernen aus der Erfahrung wäre dann nicht möglich.

Carnap (1950) musste die Induktionseigenschaft deshalb durch weitere Überlegungen sicherstellen. Er nahm als basal nun nicht mehr die einzelnen Weltzustände an, sondern bestimmte *Strukturbeschreibungen*, die sich daraus ergeben, dass diejenigen Zustände identifiziert werden, die durch Permutation der Basisobjekte auseinander hervorgehen. Solche Zustände stimmen dann in ihrer *Struktur* überein, nur dass diese durch andere Objekte ausgefüllt wird bzw. dass wir den Objekten andere Namen gegeben haben. In gewisser Weise stellen also diese Zustände dieselbe Welt dar und sollen daher nicht mehr unterschieden werden.

Das passte besonders gut zu seinen Annahmen im *Logischen Aufbau der Welt* von 1928, wonach nur *Wissen über Strukturen* wissenschaftliche Objektivität besitzt, weil es weniger subjektiv gefärbt ist als das materiale Wissen, das wir auf den unteren Ebenen unserer Erkenntnis finden. Dieser erkenntnistheoretischen Idee kann ich hier nicht nachgehen. Sie stellt eine interessante Überlegung dar, die neben anderen Überlegungen eine Grundlage für moderne Ansätze wie den Strukturenrealismus bildet. Jedenfalls finden sich bei anderen Autoren immer wieder Ansätze, die auf ähnlichen Ideen beruhen. Der Schritt zu Strukturbeschreibungen war also nicht nur eine Ad-hoc-Anpassung, um die Induktionseigenschaft zu erhalten, sondern hatte eine darüber hinaus gehende Basis in der Erkenntnistheorie. Allerdings verlässt der Schritt die einfache Anwendung des Indifferenzprinzips schon deutlich. Hier wäre also durchaus Raum für weitere Debatten, wenn wir das Programm in dieser Form fortführen wollten.

Angewandt auf unser Beispiel ergibt sich nun:

Es gibt vier Strukturbeschreibungen

{1}	»Alles ist F.«
{2, 3, 4}	»Zwei Fs, ein ¬F.«
{5, 6, 7}	»Ein F, zwei ¬Fs.« und

{8} »Alles ist ¬F.«

Diese 4 Strukturbeschreibungen erhalten nun jeweils das Gewicht 1/4, das sich dann weiter gleichmäßig auf die dazugehörigen Unterzustände verteilt. Es wird durch das Maß m* ausgedrückt.

Zustandsbeschreibung	Strukturbeschreibung	Gewichtung	m*
1. Fa & Fb & Fc	I. Alles ist F	1/4	1/4
2. ¬Fa & Fb & Fc			1/12
3. Fa & ¬Fb & Fc	II. Zwei Fs, ein ¬F	1/4	1/12
4. Fa & Fb & ¬Fc			1/12
5. ¬Fa & ¬Fb & Fc			1/12
6. ¬Fa & Fb & ¬Fc	III. Ein F, zwei ¬Fs	1/4	1/12
7. Fa & ¬Fb & ¬Fc			1/12
8. ¬Fa & ¬Fb & ¬Fc	IV. Alles ist ¬F	1/4	1/4

Tabelle 5.9: Das Maß m* für Strukturbeschreibungen

Unser neues Maß für die induktive Bestätigung ergibt sich dann zu:

c*(H,E) := m*(H&E)/m*(E)

Tatsächlich weist dieses Maß wieder die gewünschte Induktionseigenschaft auf. So können wir ausrechnen, dass c*(Fa|Fc) = 2/3 ist, während c*(Fa) = 1/2 ist. Das liegt daran, dass c*(Fa|Fc) = m*(Fa&Fc)/m*(Fc) = (1/4+1/12)/(1/4+1/12+1/12+1/12) = 2/3 ist. Der Weltzustand 1 nimmt hier eine Sonderstellung ein und erhält so ein besonderes Gewicht, das zu der Induktionseigenschaft führt.

Das Grundproblem des Indifferenzprinzips ist allerdings nicht wirklich gelöst. Die Gleichverteilung der Gewichte auf bestimmte Zustände bzw. Strukturen ist relativ zu einer ganz bestimmten Sprache bzw. eine ganz bestimmte Beschreibung der Situation und würde etwas anderes ergeben, wenn wir mit anderen Grundprädikaten (und eventuell sogar anderen Namen) arbeiten würden. Das wird in diesem ersten Ansatz zur induktiven Logik recht deutlich und bleibt weiterhin ein gravierendes Problem für jede entsprechende induktive Logik. Allerdings haben die Vertreter etwa eines objektiven Bayesianismus wie Jon Williamson darauf durchaus bestimmte Antworten parat, die wir bereits diskutiert haben.

5.7.2 Mahers Explikation von Carnaps induktiver Logik

Carnap hat neben diesem konstruktiven Weg vor allem einen *axiomatischen Ansatz* in der induktiven Logik anzubieten, den Maher (2010, 2006) ausführlich und wohlwollend diskutiert. Maher unterscheidet zwischen induktiver und physikalischer Wahrscheinlichkeit. Für die Erstere scheint nach Maher nur eine induktive Logik die angemessene Explikation zu bieten. Die subjektiveren Lesarten als vernünftige Glaubensgrade scheitern dagegen aus den oben bereits angeführten Gründen. Insbesondere betont Maher von Beginn an, dass nicht für jede Aussage eine zahlenmäßig angebbare Wahrscheinlichkeit existiert. Dort, wo sie existiert, möchten wir eine zweistellige Funktion p(A|B&K) definieren, die uns angibt, wie stark A durch B induktiv gestützt wird bei einem gewissen Hintergrundwissen K, auf dessen explizite Angabe im weiteren aber verzichtet wird, da es sich dabei um eine Konstante für die folgenden Überlegungen handelt. Dazu werden nun eine Reihe von Axiomen formuliert, die die klassischen Wahrscheinlichkeitsaxiome enthalten, aber auch deutlich darüber hinausgehen. Maher bezieht sich dabei vor allem auf die späteren Arbeiten Carnaps (vgl. Carnap 1971, 1980), die seiner Meinung nach die besseren Axiomatisierungen bieten.

Axiome der induktiven Logik
Für alle Aussagen A, B, C und D gilt:
Axiom 1: $p(A|B) \geq 0$.
Axiom 2: $p(A|A) = 1$.
Axiom 3: $p(A|B) + p(\neg A|B) = 1$, wenn B&K konsistent ist.
Axiom 4: $p(A\&B|C) = p(A|C)\ p(B|A\&C)$.
Axiom 5: Wenn gilt: A&K ≡ C&K und B&K ≡ D&K
dann ist: $p(A|B) = p(C|D)$.
Für alle Aussagen über Stichproben E gilt:
Axiom 6: (Regularität) $p(E) > 0$.
Axiom 7: (Symmetrie) p(E) bleibt konstant bei einer Permutation der Individuen.
Axiom 8: (Instanzenrelevanz) $p(F_i a_n|E\&F_i a_m) > p(F_i a_n|E)$, wenn E weder a_m noch a_n enthält.
Axiom 9: (λ-Bedingung) Wenn a ein Individuum ist, das nicht in E vor-

kommt, dann hängt $p(F_i a|E)$ nur davon ab, über wie viele Individuen E spricht und wie viele davon nach E F_i aufweisen.

Dazu müssen wir nun einige Erläuterungen abgeben. Die Basisaxiome 1-5 entsprechen ungefähr den Standardaxiomen der Wahrscheinlichkeitsrechnung. Dazu gibt es eine abzählbare Menge von Individuen a_1, a_2,... und für jede Familie von Eigenschaften F_1, F_2,...,F_n die alle Eigenschaften eines Typs (einer Modalität) darstellen (etwa Farben) eine Menge von Grundprädikaten, aus denen sich die Basisaussagen $F_i a_m$ bilden lassen. Aus endlichen Konjunktionen solcher Basisaussagen bestehen dann unsere Stichprobenaussagen E.

Das *Regularitätsaxiom* 6 gibt einen unproblematischeren Teil des Dogmatismusverbots wieder. Es ist hier beschränkt auf endliche Datenaussagen und verlangt nur, dass keine Daten durch unser Hintergrundwissen bereits ganz ausgeschlossen werden. Die *Symmetrieforderung* 7 verlangt, dass nicht bereits bei unserer Bezugnahme auf die Gegenstände schon Annahmen über die Eigenschaften dieser Gegenstände involviert sind. Wir lernen über die Objekte erst anhand unserer Beobachtungen und nehmen vorher noch keine Favorisierungen vor. Das wird auch als *Austauschbarkeit* der Individuen bezeichnet. Man könnte das so ausdrücken, dass wir keine Vorannahmen über die Eigenschaften der Objekte in unserem Hintergrundwissen gestatten, sondern erst durch induktives Lernen zu ersten Annahmen dieser Art gelangen.

Diese Axiome haben aber bereits weitergehende Konsequenzen, die u.a. bei Sandy Zabell (2009) genauer dargestellt werden. Das Austauschbarkeitsaxiom nimmt an, dass es keine speziellen zeitlichen Entwicklungen in den Daten gibt, sondern dass sie sich als Teilfolgen einer unendlichen Folge darstellen lassen, die selbst einfach eine Mischung von von *unabhängigen und identisch* verteilten Ereignissen darstellt (Zabell 2009, 283).

Das Axiom 8 fordert die *Induktionseigenschaft,* wonach wir aus dem Vorliegen von mehr Instanzen eines Prädikats F sofort und immer darauf schließen dürfen, dass die Wahrscheinlichkeit ansteigt, dass weitere Gegenstände ebenfalls F sind. Hier kommt auch tatsächlich eine neue Forderung ins Spiel. So erfüllt z.B. das Ziehen aus einer (endlichen) Urne mit Kugeln von möglicherweise verschiedenen Farben zwar die

Forderung der Austauschbarkeit, aber nicht mehr die Forderung der Induktionseigenschaft, denn die Wahrscheinlichkeit (etwa für weitere rote Kugeln) sinkt mit jeder neuen roten Kugel.

Axiom 8 stellt tatsächlich schon eine recht starke Forderung dar. Man sollte sie z.B. sinnvollerweise auf gesetzesartige Prädikate beschränken, um bestimmte antiintuitive Konsequenzen mit seltsamen Prädikaten zu vermeiden.

Hier können wir etwa an Prädikate denken, die zwar nicht in unserer Sprache enthalten sind, aber durchaus in einer anderen vorhanden sein könnten, wie das goodmansche »grue« (aus *green* und *blue* gebildet). Dabei ist grue etwa ein Prädikat, das auf Smaragde zutrifft und besagt, dass sie grün sind, wenn sie schon untersucht wurden und dass sie Blau sind, wenn sie noch nicht untersucht wurden. Das neue seltsame Prädikat tritt in Konkurrenz zum alten Prädikat grün. Wir würden etwa schließen, dass alle Smaragde grün sind, weil alle bisher untersuchten Smaragde sich als grün herausstellten. Aber mit einem ebensolchen Schluss könnten wir dann schließen, dass alle Smaragde grue sind. Beide Schlüsse geraten aber in einen Konflikt und wir sind außerdem der Meinung, dass unser erster Schluss der sinnvollere ist. Über Prädikate wie grue, die nicht zu gesetzesartigen Mustern führen, sollten wir keine Extrapolationen vornehmen. Doch es ist schwer, genau zu sagen, was die guten vor den schlechten Prädikaten auszeichnet. Nichtsdestotrotz sollten wir es versuchen, und in Axiomen wie dem Axiom 8 sollten wir nicht für die schlechten Prädikate eine Induktionseigenschaft verlangen.

Ähnliches gilt schon für andere seltsame Prädikate. Besagt F etwa »Gegenstände, die in den Werken von Conan Doyle erwähnt werden«, so möchten wir keineswegs, dass wir mit ihrer Hilfe entsprechend induktiv schließen können. Haben wir etwa gerade mehrere Gegenstände beobachtet, die in den Werken von Conan Doyle erwähnt wurden, wird das im Normalfall nicht die Wahrscheinlichkeit für alle anderen Objekte erhöhen, dass sie auch dort Erwähnung finden werden. Welche Prädikate aber projizierbar sind und welche nicht, ist eine schwierige Frage an unser Hintergrundwissen, bei der es u.a. darum geht, welche Prädikate sich in bestimmten Theorien bewähren bzw. welche Prädikate zu gesetzesartigen Aussagen führen und für welche das nicht der Fall ist. Diese Frage ist jeweils im Umfeld der konkreten Theorienbildung zu beurteilen,

wobei uns die Frage leiten kann, ob sie in einem nomischen Muster in relevanter Form auftreten oder nicht. Ein Ansatz wie der carnapsche, der mehr oder weniger auf der syntaktischen Ebene verbleibt, kann das von sich aus nicht leisten. Er benötigt hier einen Input aus unserem Hintergrundwissen, der für bestimmte Prädikate bzw. die von ihnen ausgedrückten Eigenschaften, besagt, inwieweit sie projektierbar sind.

Axiom 9 (die λ-Bedingung) besagt letztlich, dass wir es nur mit einer einfachen Art von aufzählender Induktion zu tun haben. Was für den Bestätigungsgrad an empirischer Information zählt, ist nur die bisherige relative Häufigkeit der F-Objekte. Damit kann ein Vertreter des abduktiven Schließens sich natürlich nicht anfreunden. Diese Konsequenz wird auch in dem sogenannten λ-Theorem deutlich, das aus den Axiomen abgeleitet werden kann (vgl. Maher 2010, 601).

λ-Theorem: Für $n > 2$ existiert ein $\lambda > 0$ und $k_1,\ldots,k_n \in (0,1)$ so dass gilt: Wenn E eine Stichprobe von s Individuen ist, von denen s_i die Eigenschaft F_i aufweisen und a nicht in E enthalten ist, dann gilt:
(*) $p(F_ia|E) = (s_i+\lambda\cdot k_i)/(s+\lambda)$

Zunächst lässt sich die Funktion der k_i verstehen, indem wir den Fall s = 0 betrachten, also den Fall ohne weitere Daten. Dafür gilt $p(F_ia) = k_i$, d.h., die k_i drücken unsere Vorher-Wahrscheinlichkeit dafür aus, dass a gerade die Eigenschaft F_i aus unserer Familie von Eigenschaften aufweist. Es handelt sich, wie bereits erwähnt, alles um Eigenschaften eines Typs oder einer *Modalität* wie Carnap sich ausdrückte, die sozusagen einen bestimmten Bereich unter sich aufteilen. Nehmen wir an, wir hätten etwa die drei Farbprädikate rot, blau und grün als Familie von Eigenschaften und betrachten dazu ein einfaches Beispiel. Nehmen wir etwa $\lambda = 2$ und als Vorher-Wahrscheinlichkeiten jeweils 1/3. Dann ergibt sich für E = $rot(a_1)$ & $rot(a_2)$ & $grün(a_3)$ & $blau(a_4)$:

$$p(rot(a_5)|E) = (2+2/3)/(4+2) = 4/9$$

In unserem Beispiel ist also die Wahrscheinlichkeit dafür, dass das nächste Objekt wieder rot ist, fast 1/2, weil die Hälfte der bisher beobachteten Objekte rot war. Allerdings geht auch die Vorher-Wahrscheinlichkeit mit

ein, so dass eben doch nicht ganz 1/2 erreicht wird. Wie stark sie mit eingeht, wird durch den Parameter λ bestimmt. Dessen genaue Rolle wird durch ein Umschreiben von (*) noch deutlicher (vgl. Maher 2010, 601):

(**) $p(F_i a|E) = (s_i/s)[s/(s+\lambda)] + k_i\ [\lambda/(s+\lambda)]$

Das belegt, dass $p(F_i a|E)$ ein gewichtetes Mittel der relativen Häufigkeit der F_i (also von s_i/s) und der Vorher-Wahrscheinlichkeit k_i ist. Man erkennt hieran, dass sich $p(F_i a|E)$ eher an der Vorher-Wahrscheinlichkeit orientiert, wenn λ sehr groß ist. Im Grenzfall $\lambda = \infty$ bleiben wir bei der Vorher-Wahrscheinlichkeit und lernen nichts mehr aus der Erfahrung. Wird λ dagegen sehr klein, orientieren wir uns mehr an den aufgetretenen relativen Häufigkeiten. Im Grenzfall $\lambda = 0$ bestimmt sie allein unsere Einschätzung, ob das nächste Objekt wieder ein F_i ist.

Maher (2010) argumentiert dafür, $\lambda = 2$ zu wählen, doch das ist keineswegs zwingend. Man könnte sich auch stärker an der jeweiligen Familie von Prädikaten orientieren und λ etwa anhand von empirischen Ergebnissen kalibrieren. Für homogenere Bereiche von Eigenschaften sollten wir λ kleiner wählen, während in inhomogeneren Gegenstandsbereichen λ größer gewählt werden sollte. Das entspricht den Bemerkungen für das einfache konservative Extrapolieren. Hier bleibt jedenfalls noch Spielraum für weitere Ideen dieser Art. Insgesamt ergibt sich hier das sogenannte *Carnapsche λ-Kontinuum* $c_\lambda(H,E)$ von Bestätigungsfunktionen, in denen die vorherige logische Wahrscheinlichkeit mit den Daten auf unterschiedliche Weise verrechnet wird, je nachdem wie wir λ wählen.

Einen interessanten Zusammenhang möchte ich noch erwähnen. Wenn wir $\lambda = 2$ wählen und $k_i = ½$, so erhalten wir gerade die Laplacesche Nachfolgerregel, die wir schon in Kapitel 1 eingeführt haben:

Laplacesche Nachfolger-Regel: $p(F_i a|E) = \dfrac{s_i + 1}{s + 2}$

Für den Spezialfall, dass sich bisher alle Objekte als F_i erwiesen haben, also s_i=s ist, finden wir als Wahrscheinlichkeit für ein weiteres Objekt

a gerade F_i aufzuweisen: (s+1)/(s+2). Diese Regel scheint hier nicht schlecht gewählt zu sein und gibt für kleine Stichproben eine erste sinnvolle Verrechnung von Vorher-Einschätzungen mit unseren Daten an. Für größere Zahlen nähert sich der Wert dann einfach der beobachteten relativen Häufigkeit, also in unserem Fall der 1. Wir werden uns langsam immer sicherer, dass das nächste Objekt auch die Eigenschaft F_i hat.

Maher (2010) diskutiert noch eine Reihe von Kritikpunkten an der induktiven Logik, die etwa bei Hajek (2009, Kap. 3.2) zusammengestellt werden. Der erste ist, dass die Wahl von λ willkürlich sei. Das stimmt nicht ganz, da wir für die Wahl von λ schon bestimmte Gesichtspunkte nennen können, aber sich schlicht auf $\lambda = 2$ festzulegen, ist auch keineswegs unbestreitbar. Insbesondere bleibt die Offenheit bei der Wahl der Grundprädikate und der logischen Startwahrscheinlichkeiten bestehen.

Auch an Carnaps Idee, dass die Axiome wie das Symmetrieaxiom tatsächlich einen analytischen Charakter haben, wurde Kritik geübt. Maher zieht sich einfach darauf zurück, dass es sich um Teile der Definition von p handelt und damit um analytische Aussagen. Doch wir suchen schließlich nach einer Explikation von induktiver Wahrscheinlichkeit, die unsere epistemischen Ziele optimal umsetzt und das bleibt natürlich eine z.T. empirische Frage. Wir können also fragen, welche Axiome zu einem p führen, das dann zu möglichst intuitiven Einschätzungen von Bestätigungsbeziehungen führt und das lässt sich nicht a priori bestimmen. Jedenfalls genügen die eher vagen intuitiven Begründungen der Axiome, die wir bisher haben, alleine dafür noch nicht.

Maher gibt hingegen zu, dass wir für E möglichst unsere gesamten (empirischen) Belege zu berücksichtigen haben (etwa in unserem Hintergrundwissen K), und es bleibt z.T. vage, was damit genau gemeint ist. Doch diese Vagheit muss uns nicht davon abhalten, mit einem etwas unbestimmten Hintergrundwissen K an unserer Konstruktion festzuhalten. Wir stoßen an vielen Stellen in unserer Erkenntnistheorie auf Vagheiten, ohne die betreffenden Begriffe deswegen gleich als unbrauchbar zu verwerfen. Dem kann ich mich durchaus anschließen.

Problematischer ist da schon, dass wir bestimmte Aussagen K und E als felsenfest und irrtumssicher gegeben ansehen, aus denen dann all unsere Einschätzungen ableitbar sind, und wir damit einer fundamentalisti-

schen Erkenntnistheorie folgen. Maher (2010) versucht das als *kontextuelle Position* zu beschreiben, wonach wir einfach in einem bestimmten Kontext auf K und E vertrauen, aber damit nicht ausschließen, dass wir in anderen Kontexten dieses Hintergrundwissen auch in Frage stellen können. Das sollte natürlich weiter ausgeführt werden und führt in einige der durchaus ernstzunehmenden Probleme des Kontextualismus, die wir jedoch ebenfalls nicht weiter verfolgen werden. Man könnte es so darstellen, dass wir eben nur eine eingeschränkte Fragestellung behandeln: *Wenn wir einmal annehmen, dass wir E und K sicher wissen, was sollte dann unsere induktive Wahrscheinlichkeit dafür sein, dass Fa zutrifft?* Es ist sicher legitim, zunächst einmal mit etwas vereinfachenden Fragen zu beginnen, wenn ein Gebiet so komplex und umstritten ist, wie das der (aufzählenden) Induktion.

Weitere Probleme sind darin zu sehen, dass Allhypothesen normalerweise immer die Wahrscheinlichkeit 0 bei jeder realistischen Datenlage E haben. Außerdem ist der ganze Ansatz auf ein recht einfaches Modell festgelegt und daher nicht leicht auf komplexere praktische Beispiele anwendbar. Dessen ungeachtet mag er natürlich für diese eingeschränkte Fragestellung hilfreiche Einsichten zur induktiven Wahrscheinlichkeit anbieten. Die Hauptkritikpunkte bleiben dabei sicherlich, dass die Anwendung des Indifferenzprinzips eine Abhängigkeit von bestimmten Beschreibungen aufweist und dass die Induktionseigenschaft ohne weitere Qualifikationen für alle Prädikate gefordert wird.

5.7.3 Williamsons objektiver Bayesianismus

In den letzten Jahren war es vor allem Jon Williamson, der in einer Reihe von Aufsätzen und zwei Büchern (Williamson 2005, 2010) für einen objektiven Bayesianismus eingetreten ist, der einige wichtige Grundideen der induktiven Logik aufrechterhält. Im Unterschied zur induktiven Logik wird zunächst genauer bestimmt, wie wir unser empirisches Wissen bzw. Hintergrundwissen in bestimmter Form einbringen können. Diesen Schritt nennt Williamson Kalibrierung.

Dabei werden alle möglichen Anforderungen an unsere Glaubensgrade berücksichtigt, die unser Hintergrundwissen zu bieten hat. Etwa, dass bestimmte Aussagen Wahrscheinlichkeit 1 erhalten sollten, oder dass

bestimmte Aussagen eine vorgegebene Wahrscheinlichkeit in einem vorgegeben Bereich erhalten sollen, weil wir bestimmte Grenzen kennen, innerhalb derer bestimmte physikalische Wahrscheinlichkeiten oder relative Häufigkeiten zu finden sind, die unsere umfassendstes Wissen über die betreffenden Aussagen darstellen. So wird eine Menge P_e von Glaubensgraden ermittelt, die gemäß unserem Hintergrundwissen zulässig erscheinen. Dazu bilden wir noch die konvexe Hülle $<P_e>$ aller dieser Funktionen (s. etwa Williamson 2010, Kap. 3.3). Aus der wählen wir aber dann diejenige aus (und die ist oft eindeutig bestimmt, da sie aus einer konvexen abgeschlossenen Menge gewählt wird), die der völligen Gleichverteilung $P_=$ am nächsten kommt.

Diesen letzten Schritt nennt Williamson *Äquivokation* und er wählt dafür den Kullback-Leibler-Abstand als Maßstab. Hier wird wieder eine Grundidee der induktiven Logik aufgegriffen, wonach wir zumindest so indifferent sein sollten, wie es uns unsere empirischen Evidenzen erlauben. Damit sollen insbesondere Extremwerte (wie sie Professor Wichtig schamlos für seine Zwecke nutzte) ausgeschlossen werden, und die Wahl der Ausgangswahrscheinlichkeiten soll völlig objektiv gestaltet werden.

Die Äquivokation kann auch einfach so beschrieben werden, dass in der Menge $<P_e>$ die Funktion P mit der maximalen Informationsentropie gewählt (MaxEnt) wird: $H(P) = -\sum_{\omega\in\Omega} P(\omega)\cdot\log P(\omega)$, wobei über die Elementarereignisse (oder in unserem Fall die Vollkonjunktionen) $\omega\in\Omega$ aufsummiert wird (s. etwa Williamson 2010, Kap. 2.3). Williamson (2011) gibt eine Reihe von Beispielen an, in denen die klassische Konditionalisierung und MaxEnt nicht übereinstimmen. Das sind normalerweise Fälle, in denen die klassische Konditionalisierungsregel keine Ergebnisse liefert und nicht richtig anwendbar ist, wie im sogenannten Judy-Benjamin-Beispiel. In diesen Fällen argumentiert er dafür, dass MaxEnt die besseren Resultate liefert und daher zu bevorzugen sei. Allerdings wendet sich z.B. James Joyce (2009a) in seiner Diskussion des Judy-Benjamin-Beispiels gegen die MaxEnt-Lösung und bevorzugt andere Abstandsmaße zur Bestimmung der nächsten Lösung aus Sicht der Gleichverteilung (vgl. Kap. 5.5.5 und 5.5.6).

Da wir hier zunächst meistens Fälle betrachten werden, in denen MaxEnt und die klassische Konditionalisierung übereinstimmen, soll das

nicht das Hauptthema in der Debatte des objektiven Bayesianismus sein. Ein grundsätzlicheres Problem, das Williamson schon in (2007) diskutiert, ist ein Grundproblem für alle Ansätze der induktiven Logik, nämlich zu bestimmen, inwiefern wir tatsächlich aus der Erfahrung lernen können. Das scheint schon in einfachen Situationen zu belegen, dass der Bayesianismus bzw. eine induktive Logik dabei eher ein Buchhaltungssystem als ein echter Problemlöser ist. Doch schauen wir erst einmal, wie der williamsonsche Bayesianismus dabei vorgeht.

Der Standardeinwand, von dem wir starten, lässt sich etwa am folgenden Pfirsichbeipiel erläutern. Nehmen wir an, die ersten 100 untersuchten Pfirsiche hätten einen Kern gehabt, symbolisiert durch k_1, ..., k_{100}, und wir fragen uns, wie hoch wohl die Wahrscheinlichkeit ist, dass der 101-te Pfirsich wieder einen Kern hat. Dabei sorgt MaxEnt zunächst dafür, dass gilt: $P(k_{101}|k_1\&...\&k_{100}) = ½$. Auf das Problem war Carnap schon in seiner induktiven Logik gestoßen. Kann der objektive Bayesianer also nicht aus seiner Erfahrung lernen? Carnap hatte die erforderliche Induktionseigenschaft in seinem axiomatischen Ansatz einfach postuliert (s. letztes Unterkapitel).

Williamson (2007, 2008) untersucht dieses Problem ebenfalls und kommt zu dem Schluss, dass wir uns hier auf weiteres Hintergrundwissen stützen müssen, dass die Kritiker des objektiven Bayesianismus bisher nicht beachtet haben. Es wird außer Acht gelassen, dass es sich jeweils um Instanzen desselben Prädikats handelt. Williamson (2008, 344) beschreibt das am Beispiel schwarzer Raben Ba_i:

> To derive the problem it is assumed that initially there are no constraints, and that, once the ravens have been observed, there is a single constraint induced by the evidence. This overlooks important knowledge that is implicit in the language, namely that Ba_1,...,Ba_k are all related inasmuch as they are all applications of the same predicate. If this information is not taken into account then no connection between the observations can be made.

Wenn wir diesen Zusammenhang berücksichtigen, kommen wir nach Williamson zu einer entsprechenden Anforderung für unsere Glaubensgrade, die darin besteht, dass wir einen induktiven Einfluss τ_n der

vorhergehenden Beobachtungen bestimmen können. Diesen Einfluss können wir auch in einem bayesschen Netz visualisieren, aber letztlich müssen wir vor allem bestimmte Anforderungen an die Glaubensgrade formulieren.

Nehmen wir an, für ein beliebiges vorgegebenes n sei ε die Anzahl der Kerne, die sich in n untersuchten Pfirsichen ergeben haben. Weiterhin sei k^{ε} eine Abkürzung für eine Konjunktion von Aussagen der Art: $\pm k_1 \& ... \& \pm k_n$, wobei das »±« gerade ε-mal positiv ausfällt. Um Williamsons Konzeption vorstellen zu können, vereinfache ich hier seine Notation bereits und hoffe, dass sie so etwas leichter nachvollziehbar ist. Wir definieren außerdem:

$$p_{\varepsilon} := P(k_{n+1}|k^{\varepsilon})$$

Dann können wir nach Williamson einen konkreten induktiven Schwellenwert $\tau_n \geq 0$ angeben, so dass gilt:

$$p_{\varepsilon} \geq p_{\varepsilon'} + \tau_n, \text{ wenn } \varepsilon > \varepsilon' \text{ ist.}$$

Wir nennen dann τ_n den *n-ten induktiven-Einfluss-Schwellenwert* (»n-th inductive influence threshold«). Er gibt uns an, um welchen Betrag die Wahrscheinlichkeit für k_{n+1} mindestens wachsen muss, wenn wir jeweils mehr Kerne beobachten gegenüber Daten mit weniger beobachteten Kernen in Pfirsichen. In unserem Fall liefert dann MaxEnt:

$$p_{\varepsilon} = ½ + ½\ \tau_n\ (2\varepsilon - n)$$

Daraus wird im konkreten Fall:

$$P(k_{101}|k_1 \& ... \& k_{100}) = ½ + 50 \cdot \tau_{100}$$

Und damit liegt zumindest für $\tau_n > 0$ eine positive induktive Bestätigung vor. Es ergeben sich sogar noch weitere Zusammenhänge. Zunächst ist klar, dass $\tau_n < 1/n$ bleiben muss, damit wir wieder eine Wahrscheinlichkeit erhalten. Daher können wir also ansetzen: $\tau_n = 1/(n+\lambda_n)$ mit einem λ_n zwischen 0 und Unendlich. Nehmen wir nun noch vereinfachend an, dass die λ_n konstant sind, erhalten wir wieder Carnaps λ-Kontinuum für den einfachen Fall eines Grundprädikats:

$$p_\varepsilon = \frac{\varepsilon + \frac{1}{2}\lambda}{n + \lambda}$$

Wir erhalten somit bestimmte einfache Regeln für bestimmte Werte von λ. Für λ = 1 schätzen wir die Wahrscheinlichkeit für einen weiteren Kern direkt anhand der bisher beobachteten relativen Häufigkeit der Kerne. Für λ = 1 erhalten wir die sogenannte Jeffreys-Perks-Regel, für λ = 2 die Laplace-Regel, die wir bereits besprochen haben, und für λ = ∞ findet kein Lernen aus der Erfahrung mehr statt: p_ε = ½.

Wir kehren hier also wieder zur Frage zurück, wie plausibel die Vorgaben des carnapschen λ-Kontinuums für unser Lernen aus der Erfahrung in diesem einfachen Fall einer konservativen Induktion sind? Eine ausführliche Debatte zu den carnapschen Axiomen findet sich bei Sandy Zabell (2009). Williamson ist aber nicht direkt festgelegt auf Annahmen wie Austauschbarkeit, denn die λ_n müssen nicht konstant gewählt werden, und er gibt zu, dass das carnapsche λ-Kontinuum nur unter bestimmten Voraussetzungen sinnvolle Ergebnisse liefert.

Außerdem plädiert Williamson (2007, Abschnitt 9 und S. 705) dafür, dass wir in dem Fall, dass kein weiteres Hintergrundwissen vorliegt, mit λ=0 arbeiten und die Wahrscheinlichkeit anhand der relativen Häufigkeit schätzen, wie es auch der statistische Syllogismus vorschlägt. Das ist ganz im Sinne der MaxEnt-Regel.

Insbesondere vergleicht er noch die carnapsche λ-Regel mit dem δ-Kontinuum von Christopher Nix und Jeff Paris, das die Autoren und verschiedene Mitstreiter in einigen Aufsätzen entwickelt haben (s. etwa Nix & Paris 2006). Einen entscheidenden Vorteil des Carnap-Ansatzes sieht er darin, dass dieser das gewünschte Konvergenzverhalten aufweist, was für den Nix-Paris-Ansatz nicht gilt. Es sollte nämlich für die geschätzten Wahrscheinlichkeit gelten:

$$\lim_{n\to\infty} (P(k_{n+1}|k^\varepsilon) - \tfrac{\varepsilon}{n}) = 0$$

Das ist sicher eine wichtige Eigenschaft für das induktive Schließen. Zumindest im Grenzwert sollte sich die geschätzte Wahrscheinlichkeit der relativen Häufigkeit im Grenzwert annähern (in fast allen Fällen). Doch damit allein wird die Debatte um die beste induktive Logik noch nicht zu entscheiden sein.

Mir ist aber ein anderes Thema wichtiger: Inwiefern verhilft uns nun die induktive Logik oder der objektive Bayesianismus zu unseren Induktionsschlüssen? Wir haben gesehen, dass wir auf die eine oder andere Weise immer darauf angewiesen sind, die Induktionseigenschaft extra zu fordern und dann die entsprechenden Parameter selbst extern zu schätzen und in den Rahmen des objektiven Bayesianismus einzubringen. Der Bayesianismus selbst liefert dann nur das Buchhaltungsverfahren, um damit weiter rechnen zu können. Es bleibt dagegen für uns die Aufgabe bestehen einzuschätzen, wie sehr die vorangehenden Instanzen eines Prädikates es begründen können, zukünftige entsprechende Instanzen zu erwarten. Dafür liefern uns die hier diskutierten Ansätze genaugenommen keine weitere Hilfestellung.

Woran sollen wir uns dabei aber orientieren? Hier sind wir m.E. wieder auf unsere informellen Überlegungen zum Schluss auf die beste Erklärung angewiesen. Wir müssen uns etwa fragen, was die beste Erklärung dafür ist, dass bisher alle Pfirsiche einen Kern hatten oder warum ein Medikament in 70% der Fälle wirksam war. Liegen hier nomische (kausale) Muster zugrunde oder handelt es sich nur um einen Zufall? Um diese grundlegenden Debatten kommen wir nicht herum, wenn wir etwa sinnvolle Werte für den n-ten induktiven-Einfluss-Schwellenwert bestimmen wollen. Im Falle der Pfirsiche gehen wir davon aus, dass es sich um eine natürlich Art handelt und da die Kerne eine wichtige Funktion bei der Fortpflanzung haben, handelt es sich vermutlich um ein wesentliches Merkmal der Art, solche Kerne entweder aufzuweisen oder nicht. Also ist der Wert relativ hoch anzusetzen.

Allerdings wissen wir auch, dass mit den Methoden der modernen Genetik selbst solche Merkmale verändert werden können und sie daher nicht völlig stabil sind. Grundlegender sind da schon Merkmale von Elementarteilchen. Ehe wir mit unseren bayesianischen Methoden zu genauen induktiven Schlüssen gelangen, sind wir wiederum auf Überlegungen zum (metaphysischen) gesetzesartigen Hintergrund unserer Erscheinungen mit angewiesen. Der Traum der induktiven Logiker von einer reinen (apriorischen) induktiven Logik scheint jedenfalls nicht in Erfüllung zu gehen. Sie liefert uns ein schönes Buchhaltungssystem für induktiven Schließen, aber zumindest einige entscheidende Parameter müssen wir anhand externer Überlegungen beisteuern.

5.7.4 Hawthornes induktive Logik

James Hawthorne (2011) hat in verschiedenen Aufsätzen eine moderne induktive Logik vorgestellt, die nicht mehr davon ausgeht, dass es *eindeutige* Vorher-Wahrscheinlichkeiten gibt. Wir können durchaus Vagheiten und Mehrdeutigkeiten in der Verteilung von Startwahrscheinlichkeiten zulassen und unseren epistemischen Zustand dann durch eine *Menge* von Wahrscheinlichkeitsfunktionen wiedergeben. Wichtig ist vor allem, dass die Likelihoodanbindung angenommen wird, und durch die entsprechenden Konvergenztheoreme für die Likelihoods (s. Kap. 5.6.3) findet schließlich eine Angleichung der Startwahrscheinlichkeiten statt, wie wir das oben bereits erörtert haben, solange die Ausgangsmengen jedenfalls nicht zu heterogen sind.

Die Wahrscheinlichkeiten werden dabei im Sinne eines *objektiven Bayesianismus* nicht einfach nur als beliebige Glaubensgrade interpretiert. Vielmehr sollen sie dann verstanden werden als *Bestätigungsfunktionen* P(A|B), die darüber Auskunft geben, in welchem Ausmaß A durch B gestützt wird. Man könnte auch sagen: P(A|B) beschreibt den Anteil der A-Welten innerhalb der B-Welten, also den Anteil der Welten, in denen A wahr ist innerhalb der Menge der Welten, in denen B wahr ist. Diese Bestätigungsfunktionen funktionieren daher ähnlich wie Wahrheitswertbelegungen in der Logik und können in analoger Weise verstanden werden.

Hawthorne formuliert für derartige Bestätigungsfunktionen P_α die minimalen grundlegenden Axiome, die jede solcher Funktionen erfüllen sollte. Das α ist dabei ein Parameter, der für die jeweilige Person mit einem bestimmten Hintergrundwissen steht oder eben für eine bestimmte solcher Bestätigungsfunktionen in einer größeren Menge. Wir nutzen selbstverständlich wieder das bayessche Theorem, aber damit muss eben keineswegs gleich ein subjektives Verständnis der Bestätigungsfunktionen oder Wahrscheinlichkeiten einhergehen. Aus den oben bereits genannten Gründen (insbesondere dem, dass bestimmte konditionale Wahrscheinlichkeiten P(A|B) durchaus sinnvoll sein können, auch wenn P(B) = 0 ist) sind die grundlegenden Axiome direkt für bedingte Wahrscheinlichkeiten formuliert. Und so erhalten wir:

Axiome von Hawthorne (2009)

(1) $P_\alpha(D|E) < 1$ für einige Sätze D und E
Für alle Sätze A, B, C, und D gilt:
(2) Wenn gilt $B \models A$, dann ist $P_\alpha(A|B) = 1$
(3) Wenn gilt $\models (B \equiv C)$, dann ist $P_\alpha(A|B) = P_\alpha(A|C)$
(4) Wenn gilt $C \models \neg(B\&A)$, dann ist entweder $P_\alpha((A \vee B)|C) = P_\alpha(A|C) + P_\alpha(B|C)$ oder für jeden Satz D: $P_\alpha(D|C) = 1$
(5) $P_\alpha(A\&B|C) = P_\alpha(A|B\&C) \cdot P_\alpha(B|C)$
(6) Wenn A ein Axiom der Mengenlehre oder reinen Mathematik ist bzw. wenn A eine analytische Wahrheit ist, dann sei für alle Aussagen C: $P_\alpha(A|C) = 1$
(7) Wenn für alle C gilt: $P_\alpha(A|C) = 1$, dann ist A ein Axiom der Mengenlehre oder reinen Mathematik oder ist analytisch wahr im Sinne von P_α.

Das Axiom (1) ist eine Form von Nichttrivialitätsforderung. Zumindest einige Sätze bestätigen einige andere Sätze auf nichttriviale Weise. Die maximale Bestätigung wird bei logischen Folgebeziehungen erreicht (Axiom 2). Dabei geben logisch äquivalente Aussagen jedem Satz dieselbe Stützung (Axiom 3) und diese Bestätigung ist in der bekannten Weise additiv (Axiom 4). In Axiom (5) finden wir die allgemeine Multiplikationsregel bzw. Kettenregel für zweistellige Wahrscheinlichkeitsfunktionen. Damit haben wir die Axiome, aus denen alle üblichen Theoreme der Wahrscheinlichkeitsrechnung wieder folgen. Da es sich um Bestätigungsfunktionen für empirische bzw. kontingente Sätze handeln soll, erhalten die analytischen Wahrheiten in Axiom 6 und speziell 7 eine Sonderstellung als die *einzigen* Aussagen, die immer den Bestätigungsgrad 1 durch alle anderen Aussagen erhalten. Diese Regularitätsannahme geht über die Standardaxiome hinaus und entspricht dem Dogmatismusverbot im klassischen Bayesianismus.

Außerdem wird vor allem eine strikte Likelihoodanbindung sehr ernst genommen, wodurch auch die »old evidence« Probleme gelöst werden sollen, die wir im nächsten Abschnitt behandeln werden. Hawthorne stützt sich außerdem ganz auf die oben angeführten Likelihood-Konvergenzsätze, um zu zeigen, dass die ursprüngliche Vagheit und Mehrdeutigkeit der Menge von Bestätigungsfunktionen, die unsere epis-

temischen Zustände ausdrücken, durch wiederholtes Updaten schnell reduziert wird (s.o.). Damit haben wir einen Ansatz innerhalb der induktiven Logik vor uns, der nicht auf *syntaktische* Überlegungen zur Bestimmung von Startwahrscheinlichkeiten setzt, wie das noch bei Carnap der Fall war. Hawthorne hält die Probleme durch Prädikate vom »grue« Typ für überzeugende Einwände gegen die syntaktisch orientierten induktiven Logiken. Das geht in dieselbe Richtung wie meine Auffassung, dass insbesondere Generalisierungen, die einen nomischen Charakter haben, induktiv stützbar sind, wir in anderen Fällen aber viel zurückhaltender sein sollten. Dieser Unterschied wird auf einer bloß syntaktischen Ebene aber noch nicht sichtbar.

5.8 Probleme bayesianischer Bestätigungskonzeptionen

5.8.1 Das Versagen der Likelihoodanbindung

Ein Problem des klassischen subjektiven Bayesianismus (auch in Bezug auf die Likelihoods) ist das der »alten Evidenz« (»old evidence«) bzw. der bereits *bekannten Daten*. Wenn E erst einmal zu unserem Hintergrundwissen B gehört, erhält es die Wahrscheinlichkeit 1 (und eigentlich meint der Bayesianer mit P(E) immer P(E|B)), dann ist klar, dass P(E) bereits den Wert 1 erhalten muss. Überhaupt erhält E beim Updaten mit E den Wert 1: P(E|E) = 1. Aus P(E) = 1, folgt aber sogleich auch, dass P(E|H) = 1 sein muss für beliebige Hypothesen H. Darauf kommen wir unten wieder zurück. Zunächst möchte ich aber noch auf ein weitergehendes Problem hinweisen, nämlich dass die Likelihoods nicht nur durch alte Daten mit P(E) = 1 verändert werden, sondern ebenso durch andere Informationen, die etwas über E aussagen, bereits deutlich verändert werden können.

Andere Erkenntnisse als E selbst können bereits die Likelihoods verändern, so dass der Bayesianer jedenfalls nicht mehr mit den ursprünglichen Likelihoods arbeiten kann. Mit P(E|H) = r ist dann eben nicht mehr gemeint, was die Theorie H über E aussagt, sondern vielmehr, was die upgedatete Glaubensfunktion P mit Hintergrundwissen B inzwischen alles über E zu sagen hat. Also genau genommen gilt eben: P(E|H) = P(E|H&B). Dabei können sogar schon kleinste Informationen in unserem

Hintergrundwissen, die E nur peripher betreffen, unsere Likelihoods verfälschen und wir können schließlich nicht mehr die Likelihoodanbindung einhalten, wenn wir bayesianisch updaten. Nehmen wir zur Illustration dieses Zusammenhangs ein schönes Beispiel von Hawthorne (2011b).

Ein Arzt möchte mit einem Patienten einen Laufbandtest durchführen, um herauszufinden, ob der Patient eine Herzkrankheit aufweist (h). Der Arzt weiß aus zahlreichen medizinischen Studien, dass die Wahrscheinlichkeit 10% dafür ist, dass der Test ein negatives Resultat (e) liefert (also fälschlicherweise keinen Hinweis auf eine Herzkrankheit angibt), wenn der Patient eine Herzkrankheit aufweist: P(e|h) = 0,1. (Die Sensitivität P(¬e|h) für das Auftreten einer Herzkrankheit ist bei dem Test also nur 0,9.) Das sollte dann im Sinne der Likelihoodanbindung unsere entsprechende Likelihood für das Updaten von h mit e sein.

Außerdem wissen wir aber noch, dass die routinierte Krankenschwester des Arztes sich durch ein positives Testergebnis nur selten aus der Ruhe bringen lässt (d), nämlich nur in 5% aller Fälle: P(d|¬e) = 0,05. Weiterhin sei noch erwähnt, dass die Erschütterung der Krankenschwester unabhängig davon ist, ob die Herzkrankheit tatsächlich auftritt, denn das weiß sie normalerweise nicht: P(d|¬e&h) = P(d|¬e). Das ist etwa die Beschreibung eines kleinen bayesschen Netzes H⇛E⇛D, mit Variable H mit jeweils zwei Werten {h, ¬h} und den entsprechenden Variablen E und D. Das Vorliegen einer Herzkrankheit (H) beeinflusst das Testergebnis (E) und das wiederum die Befindlichkeit der Krankenschwester (D).

In dem Netz können wir nun fragen, wie sich die Information e auf unsere Vermutungen bzgl. der Herzkrankheit auswirkt und uns dazu auf die Likelihood P(e|h) = 0,1 stützen. Doch bevor der Arzt etwa die Information e oder ¬e erhält, erhält er noch die neue Information j ≡ ¬e→d ≡ e∨d, wonach ein positives Testergebnis definitiv die Krankenschwester erschüttert, falls es denn auftritt. Die Information besagt anders formuliert, dass unser Testergebnis negativ ausfällt oder unsere Krankenschwester erschüttert ist. Eins von beidem wird zumindest eintreten. Dann muss der Arzt als guter Bayesianer zunächst mit j updaten. Dabei erhält er aber leider die neue Likelihoodfunktion: $P^+(e|h)$ = P(e|j&h) = 0,69. Das zeigt die folgende Berechnung:

Voraussetzungen im Beispiel:
h: Herzkrankheit liegt vor; e: negatives Testresultat;
d: Erschütterung der Krankenschwester und j ≡ e∨d.
P(e|h) = 0,1; P(d|¬e) = 0,05; P(d|¬e&h) = P(d|¬e) also gilt:
P(j|¬e&h) = P(e∨d|¬e&h) = P(d|¬e&h) = P(d|¬e) = 0,05 und
P(j|e&h) = P(e∨d|e&h) = 1,

dann erhalten wir beim Updaten mit j in Kurzschreibweise:
$P^+(e|h) = P(e|jh) = P(ejh)/P(jh) = P(ejh)/[P(ejh) + P(jh\neg e)] =$
$P(j|eh)\cdot P(e|h) / [P(j|eh)\cdot P(e|h) + P(j|h\neg e)\cdot P(\neg e|h)] =$
0,1 / [0,1+0,05·0,9] = 0,69

Dabei wurde die Kettenregel angewandt und mit dem Satz der totalen Wahrscheinlichkeit der Nenner bestimmt, wobei durch P(h) gekürzt werden konnte. Das Resultat sollte uns jedenfalls ein wenig erschüttern, denn damit ist die ursprüngliche Likelihoodanbindung verlorengegangen. Wenn wir nun mit e updaten würden, müssten wir die neue Likelihood $P^+(e|h) = 0{,}69$ anwenden. Die Likelihoodanbindung ist ähnlich wie im Falle des Hauptprinzips, das auch nicht einzuhalten ist, wenn wir bayesianisch updaten, ebenso wenig aufrechtzuerhalten, wenn wir zwischendurch mit recht harmlosen aussehenden Informationen wie ¬e→d updaten. Damit geht für solche Fälle schon der Aspekt der Konvergenzidee verloren, dass zwar die Startwahrscheinlichkeiten subjektiv sind, aber wir doch die fest vorgegebenen Likelihoods haben, die intersubjektiv bestehen bleiben.

Allerdings sollte uns das Resultat andererseits auch nicht allzu sehr überraschen. Denken wir dazu an unser obiges Netz. Die Herzkrankheit beeinflusst das Testergebnis und dieses seinerseits die Erschütterung der Krankenschwester. Erhalten wir nun die spezielle Information j, dass das Testergebnis negativ war (e) oder die Krankenschwester erschüttert (d), so können wir zwei Fälle unterscheiden. Zum einen könnte j wahr sein, weil e vorliegt, dann ist $P^+(e) = 1$ und damit $P^+(e|h) = 1$. Zum anderen könnte j wahr sein, weil d der Fall ist. Dann sagt uns aber unser kleines Netz, dass das selbst bei Vorliegen von ¬e kaum vorkommt: P(d|¬e) = 0,05. Also ist wohl zu erwarten, dass vermutlich doch e vorliegt. Damit steigt die Wahrscheinlichkeit von e und damit die von $P^+(e|h) > P(e|h)$. Doch das ist eben höchst unerwünscht, wenn wir eigentlich doch ermitteln

wollen, inwieweit das Auftreten von e unsere Hypothese h stützt oder schwächt.

Betrachten wir etwa den Fall von zwei Ärzten. Der erste ist unser bisheriger Arzt, der die doch relativ irrelevante Information erhält, dass die Krankenschwester erschüttert ist, wenn sie von einem positiven Testresultat erfährt, während der andere Arzt diese Information nicht erhält. Dann werden sie die Wahrscheinlichkeit für eine Herzkrankheit des Patienten unterschiedlich beurteilen, wenn e auftritt, obwohl die Information j darüber inhaltlich eigentlich nicht viel aussagt. Wenn die Likelihoodanbindung nicht mehr gilt, geht auch die wissenschaftliche Objektivität oder zumindest die intersubjektive wissenschaftliche Übereinstimmung verloren. Daher plädiert Hawthorne in (2011c) überzeugend dafür, nur solche Wahrscheinlichkeitsfunktionen als Bestätigungsfunktionen zuzulassen, für die die Likelihoodanbindung (bei ihm »direct inference likelihoods«) gilt. Sie wählen als Likelihoods nur das, was uns die Theorien und ihre Hilfshypothesen darüber sagen, wie wahrscheinlich bestimmte Daten zu erwarten sind. Stimmen diese Werte nicht mehr überein, verfolgen und testen die Wissenschaftler genau genommen bereits verschiedene Theorien, obwohl sie weiterhin dieselben Sätze verwenden, um sie auszudrücken (vgl. Hawthorne 2011c).

Das Problem des Bayesianismus liegt darin begründet, dass allgemein P(E|H) eigentlich der Wahrscheinlichkeit P(E|H&B) mit dem Hintergrundwissen B entspricht. Betrachten wir etwa den Abstand d(H,E) = P(H|E)–P(H) als mögliches Bestätigungsmaß, dann betrifft der nicht die Aussagekraft, die E bzgl. H hat, sondern gibt vielmehr an, was E nun relativ zum bisherigen Hintergrundwissen zur Stützung von H beizutragen hat. Gehört E jeodch bereits zum Hintergrundwissen, dann liefert die Information E keinen zusätzlichen Schub für H mehr und scheint damit nach der bayesianischen Bestätigungstheorie ebenfalls keine Bestätigung für H zu liefern. Doch das entspricht keineswegs unserem normalen Verständnis von Bestätigung. Überhaupt gilt: Wenn z.B. H⇒E gilt, so hängt der Update-Faktor nur noch an P(E). Je besser E bereits in unserem Überzeugungssystem verankert ist, desto weniger Stützung bietet E für H. Doch das hat nicht mehr viel mit unserem üblichen Verständnis von Bestätigung zu tun. Unser Hintergrundwissen

soll zwar Einfluss auf die Bestätigungsbeziehung haben, aber dabei soll es hauptsächlich um den Einfluss unserer theoretischen Annahmen gehen und nicht so sehr um historische Zufälligkeiten, wie die, ob wir E schon wissen, bevor wir H entwickelt haben, oder erst danach.

Wenn etwa jemand behauptet, eine Münze habe eine physikalische Wahrscheinlichkeit von 0,6 bei einer bestimmten Wurfanordnung auf Kopf zu fallen (H), dann wird diese Hypothese H und unsere Annahmen über die Randbedingungen des Werfens zusammen unser Modell M bilden, das somit die Wahrscheinlichkeit für bestimmte Daten D bestimmt. Dieser Zusammenhang ist zeitlos und unabhängig von unserem Informationsstand und wird daher durch unsere Glaubensgradfunktionen nicht richtig wiedergeben. Hawthorne (2005) schlägt deshalb vor, dass wir neben den Glaubensgraden noch eine *reine Bestätigungsfunktion* benötigen, die diesen Zusammenhang darstellt. Doch wie ist dann der Zusammenhang zu unseren Glaubensfunktionen? Hawthorne plädiert dafür, dass hier die Bestätigungsfunktion P den Ton angibt und die Glaubensfunktion Bel sich daran orientieren sollte:

$$\mathrm{Bel}(h_i) = \mathrm{P}(h_i|b\&c^n\&e^n)$$

Dabei handelt es sich um das Hintergrundwissen b und die Bedingungen c^n unter denen die Daten e^n gewonnen wurden. Das lässt sich auch auf den Fall von unsicherem Wissen anwenden:

$$\mathrm{Bel}(h_i) = \sum_{b,c,e} \mathrm{P}(h_i|b\&c^n\&e^n)\ \mathrm{Bel}(b\&c^n\&e^n)$$

Hierbei wird über das Hintergrundwissen b und die c und die e aufsummiert. Auch Patrick Maher (2006) analysiert sorgfältig, welche Interpretation für das Konzept der induktiven Wahrscheinlichkeit angemessen ist und hält die rationalen Glaubensgrade der Bayesianer dafür für ungeeignet. Wenn es uns mehr um die Frage der möglichst objektiv zu bestimmenden Theorien*bestätigung* in der Wissenschaft geht, sollten wir uns mehr darum bemühen uns den hawthorneschen Bestätigungsfunktionen zu nähern und die Likelihoodanbindung als Forderung möglichst weitgehend in unsere probabilistischen Überzeugungssysteme einzubauen. Wie weitgehend das gelingen kann und welche Art von Informationen wir vielleicht sogar als irrelevant ausschließen müssen (wie

im oberen Beispiel die Information j), das wird die weitere Forschung zeigen müssen.

5.8.2 Das Problem der alten Evidenz und die Intransparenz der Bestätigungsbeziehung

Das Problem der alten Evidenz tritt überall dort in Erscheinung, wo wir das Ausmaß der Bestätigung einer Theorie H durch Daten D bestimmen möchten. Das war schließlich eine klassische Frage der Erkenntnis- und der Wissenschaftstheorie, mit der wir in diesem Buch gestartet sind. Betrachten wir wiederum das klassische Differenzmaß d(H,E) = P(H|E) – P(H). Es scheint dazu zu dienen, die relationale Bestätigung von H durch das Datum E anzugeben (und wird wohl auch von den meisten Bayesianern so eingesetzt), aber das Problem ist, dass sich P durch mehrfaches Updaten langsam verändert und eigentlich d(H,E) = P(H|E&B) – P(H|B) damit gemeint ist. P(H) gibt uns also immer den *momentanen Plausibilitätsgrad* bzw. rationalen Glaubensgrad von H an, bei einem bestimmten vorgegebenen Hintergrundwissen B. Gehört dann E bereits zu diesem Hintergrundwissen, ist seine Wahrscheinlichkeit 1 und d(H,E) = 0. Das liegt schon daran, dass bei P(E) = 1 der bayesianische Update-Faktor P(E|H)/P(E) auch gleich 1 wird. Doch das sagt wenig darüber aus, ob E einen Grund für die Annahme von H darstellt. Das Differenzmaß ändert sich mit der Entwicklung unseres Hintergrundwissens und ist daher nicht in der Weise relational zu verstehen, wie Bayesianer das gerne sehen würden. Das Ratiomaß hilft hier offensichtlich nicht weiter und selbst das Likelihoodmaß hilft uns nur dann, wenn die Likelihoods gerade die objektiven Likelihoods sind, die durch die Hypothesen selbst gegeben sind, und nicht die bloß upgedateten Wahrscheinlichkeiten, die der Bayesianer verwendet.

Also sagen uns die normalen Glaubensgrade P(H|E) zwar, wie plausibel H im Lichte von E und unserem gesamten Hintergrundwissen ist, aber sie geben dann nicht mehr unbedingt an, wie stark E unsere Hypothese H bestätigt. Das kann verschüttet sein unter anderen Daten und Zusammenhänge und sollte wohl eher durch Vergleiche der Art $P_H(E)/P_{H^*}(E)$ zu anderen Hypothesen H* oder durch Differenzen der

Form $P_\alpha(H|E)–P_\alpha(H)$ für reine Bestätigungsfunktionen P_α angegeben werden.

Das heißt, der übliche Glaubensgrad P(H|E) = P(H|E&B) stellt schon die gesamte Bestätigung dar, die H durch unser ganzes Hintergrundwissen B einschließlich E erfährt. Es hat sich so entwickelt, indem wir mit einer bestimmten gemeinsamen Vorherverteilung P_1 bei einem Hintergrundwissen B_1 für all unsere Überzeugungen gestartet sind, und hat sich dann nur durch Updaten zu unserer heutigen Verteilung P_n mit dem neuen Hintergrundwissen B_n entwickelt. Dabei entstand jede neue Verteilung P_{i+1} aus P_i durch das Updaten mit einem Datum E_i, das wir zu diesem Zeitpunkt beobachtet haben. Genauso mechanisch ging (jedenfalls für den Bayesianer) B_{i+1} aus B_i hervor, indem wir E_i zu B_i hinzugefügt haben. Das ist ein wirklich einfaches mechanisches Modell, das allerdings durch seine Einfachheit schon gewisse Schwierigkeiten mit sich bringt. Führen wir etwa zu einem bestimmten Zeitpunkt neue Theorien ein, müssen wir wieder von vorne mit dem Updaten beginnen.

Da es sich um eine ausschließlich dynamische Auffassung von P handelt, die jeweils die neu auftretenden Daten mit einbezieht, stellt P(H) = $P_{t,S}(H)$ jeweils die Gesamtbeurteilung einer Person S zum Zeitpunkt t dar. Das passt eher zur Konzeption einer *absoluten Bestätigung*, die allerdings keinen Hinweis mehr darauf gibt, inwiefern nun E dazu beiträgt, H zu stützen. Man könnte das die *Intransparenz der bayesianischen Bestätigungstheorie* nennen. Alle Bestätigungen sind bereits in $P_{t,S}(H)$ in irgendeiner Form enthalten und sind dann nicht mehr im Einzelnen aufzuschlüsseln. Alle Bestätigungen durch Daten und möglicherweise anderes Hintergrundwissen ist in $P_{t,S}$ »versteckt« und damit in intransparenter Weise enthalten. Das zeigen auch die eher kläglichen Versuche der Bayesianer, diese relationalen Bestätigungsbeziehungen wieder aus $P_{t,S}$ herauszulesen. Die schauen wir uns gleich kurz an.

Doch zunächst betrachten wir ein historisches Beispiel, das an dieser Stelle gern genannt wird. Die Perihelbewegung des Merkurs beschreibt die Bewegung des sonnennächsten Punktes auf der Ellipsenbahn des Merkurs um die Sonne herum (das Aphel bezeichnet den sonnenfernsten Punkt dieser Bahn). Leider verhält sich das Perihel des Merkurs nicht so, wie es von der newtonschen Gravitationstheorie vorausgesagt wurde. Daher spricht man von der *Perihelanomalie des Merkurs*. Die sprach

gegen die newtonsche Gravitationstheorie und war ein Grund für die Entwicklung der allgemeinen Relativitätstheorie, denn dieser gelang es, diese Bewegung korrekt zu erklären. Als Einstein die allgemeine Relativitätstheorie entwickelte, war die Perihelanomalie des Merkurs aber schon bekannt. Dann hätte nach unserem Differenzmaß diese Anomalie keinen Bestätigungseffekt mehr für die einsteinsche Theorie gehabt. Doch das entspricht überhaupt nicht unserer üblichen Auffassung dieser Zusammenhänge. Die Physiker gingen natürlich davon aus, dass die Anomalie einen besonders gewichtigen Grund für eine Annahme der allgemeinen Relativitätstheorie darstellt. Über deren genaue Stärke können wir weiter diskutieren, aber sicher nicht über das Faktum einer Bestätigung selbst. Das kann der Bayesianer nicht richtig rekonstruieren. Er hat dazu einige Vorschläge entwickelt (Epizyklen des Bayesianismus?), mit denen er unsere Intuitionen zur Bestätigung doch noch einfangen möchte.

Eine erste Idee findet sich in der obengenannten Beschreibung der Entwicklung unserer Wahrscheinlichkeitsfunktion P und unseres Hintergrundwissens B. Gehen wir zunächst davon aus, dass die Theorie H vor dem Auftreten des Datums E bekannt war. Dann gab es einen Entwicklungsschritt in der Vergangenheit unseres Hintergrundwissens von B_i zu B_{i+1}, bei dem gerade E zu B_i hinzugefügt wurde. Dann könnten wir als Maß für die relationale Bestätigung gerade das *historische Maß* annehmen:

$$B_{hist}(H,E) = P_{i+1}(H) - P_i(H) = P_i(H|E) - P_i(H)$$

Doch auch das trifft unsere Wünsche nicht wirklich. Es könnte vielleicht dazu dienen, festzustellen, welche Bestätigung E für H zum damaligen Zeitpunkt lieferte, aber das sagt nicht unbedingt etwas darüber aus, wie die heutige Bestätigung von H durch E aussieht mit unserem heutigen anderen theoretischen Hintergrundwissen. Überhaupt müssen wir uns dazu nur den Update-Faktor noch einmal ansehen und überlegen, wie darin $P_i(E)$ jeweils bestimmt wird. Normalerweise gehen wir davon aus, dass wir über eine vollständige Menge sich ausschließender Hypothesen $\{H_1,\ldots,H_n\}$ verfügen, mit deren Hilfe sich P(E) berechnen lässt: $P(E) = \sum_i P(E|H_i)\cdot P(H_i)$. Dabei gehen die Wahrscheinlichkeiten $P(H_i)$ ein, die sich durch jedes Updaten langsam verändern werden. Welchen Wert

der Update-Faktor annimmt, hängt also von unserer jeweiligen Wahrscheinlichkeitsverteilung auf die Hypothesen ab. Damit ist $B_{hist}(H,E)$ nicht nur ein bloß historischer Wert, sondern auch noch dort relativ zu dem Zeitpunkt, an dem man auf E gestoßen ist. Finden wir zunächst viele Daten, die für die Hypothesen sprechen, die E einen hohen Wert geben, so dass deren Wahrscheinlichkeiten hoch sind, dann stellt demnach E keine so starke Bestätigung für H mehr dar, als wenn die Reihenfolge andersherum wäre.

Der Bayesianer mag das rechtfertigen und eben davon ausgehen, dass Bestätigung ein derart *historisch relatives Konzept* darstellt. Für das Hintergrundwissen, das die Hypothesen wie H schon gut bestätigt hat, ist dann eben kein Platz mehr für eine weitere gute Bestätigung von H durch E. Doch das widerspricht deutlich unserer üblichen Auffassung von Bestätigung. Die Bestätigung von H durch E mag durch ganz neu entwickelte Theorien anders ausgehen, weil diese alternative Erklärungen von E anbieten können, sie sollte sich aber nicht durch bloß neue Daten entscheidend verändern. Die Daten mögen unser Vertrauen in die Theorie stärken, sie entwerten aber dadurch doch nicht andere Informationen – vor allem nicht vollständig, wie das im Falle des Problems der alten Daten der Fall zu sein scheint. Jedenfalls bietet $B_{hist}(H,E)$ nicht das gewünschte relationale Maß, weil es sich auf altes Hintergrundwissen bezieht, und womöglich ist es schon deshalb nicht so gut geeignet, weil es die Bestätigung zu stark auf die jeweils vorliegenden Daten relativiert.

Einige Bayesianer gehen daher einen anderen Weg und sagen, wir müssten zu unserem Hintergrundwissen B, das E enthält, ein anderes *kontrafaktisches Hintergrundwissen* B* entwickeln, das gerade E nicht mehr enthält (das um E *kontrahiert* wurde). Dann könnten wir das neue kontrafaktische Maß für unsere Bestätigung entwickeln:

$$B_{kontraf}(H,E) = P(H|E\&B) - P(H|B^*)$$

Allerdings zeigte schon das Hawthorne-Beispiel (und die Debatte um die Glaubensdynamiken), dass wir weitere Aussagen neben E aus B herausnehmen müssen, um zu B* zu gelangen. Wir müssen geradezu alle Spuren von E in B tilgen, also z.B. auch alle konditionalen Aussagen

mit E beseitigen, sonst erhalten wir die Probleme wie in Hawthornes Diagnose-Beispiel. Dann ist aber auch keineswegs mehr klar, wie wir die Wahrscheinlichkeiten für alle unsere Überzeugungen, also insbesondere für unsere Hypothesen $\{H_1,\ldots,H_n\}$ neu vergeben sollen. Das müssen wir aber wissen, um $B_{kontraf}(H,E)$ berechnen zu können.

Das Ganze ist sicher eine kreative Idee, die unseren Intuitionen von Bestätigung entgegenkommt, aber wir fallen nun endgültig aus dem bayesianischen Verfahren heraus, denn der Schritt von B zu B* ist kein Schritt, der im klassischen Bayesianismus angelegt wird. Wir müssen nun ganz neue Erwägungen unternehmen, für die uns der bayesianische Apparat nicht weiterhilft. Viele Wissenschaftstheoretiker sehen auch die Schwierigkeiten als zu groß an, zunächst B* zu entwickeln und dann völlig aus der Luft auch noch P(H|B*) festzulegen. Für P(H|B*) müssten wir einen Neustart mit neuen Vorher-Wahrscheinlichkeiten beginnen, und der klassische Statistiker würde zu Recht einwenden, dass diese wiederum subjektiv seien und noch nicht einmal durch wiederholtes Updaten geadelt und objektiviert wurden.

Es gibt zwar noch andere Lösungsideen, aber letztlich bleibt das Problem der alten Evidenz für den klassischen Bayesianismus bestehen und besitzt bisher keine klare Lösung. Das hieße im Endeffekt, dass der Bayesianismus keine wirklich überzeugende Konzeption eines relationalen Bestätigungsbegriffs anzubieten hat. Das wäre für ein erkenntnistheoretisches und wissenschaftstheoretisches Grundprogramm schon ein arges Manko. Den besten Vorschlag für ein Maß der Bestätigungsstärke finden wir dann m.E. in dem komparativen Vergleich der objektiven direkten Likelihoods zweier Hypothesen H und H*, das wir schon oben genannt haben $B(H,H^*;E) = P_H(E)/P_{H^*}(E)$. Dabei handelt es sich um objektive Werte und wir finden das entsprechende Maß im Rahmen der klassischen Statistik in Kapitel 6 wieder und ebenfalls dort in dem Bayes-Faktor der bayesianischen Statistik. Außerdem stellt es einen wesentlichen Maßstab für das abduktive Schließen dar. Hier kommen viele Entwicklungen zusammen, und wir können damit dem Problem der alten Evidenz entkommen.

Man wirft damit allerdings erneut die alte Frage auf, was wir mit den bayesianischen Glaubensfunktionen innerhalb der Erkenntnistheorie in der Hand haben? Im Kapitel 5.8.8 wird Peter Achinstein diese gesamte

Problematik verstärken und dann eine Konzeption von absoluter Bestätigung vorschlagen, die den Bayesianismus allerdings ein gutes Stück weit verlässt und damit ebenfalls eher an einen objektiven Bayesianismus anknüpft.

5.8.3 Die Asymmetrie von Vorhersage und Retrodiktion

Ein Thema kam schon weiter oben kurz zur Sprache und lauert ebenso im letzten Kapitel im Hintergrund, nämlich das der Reihenfolge von Theorien und Daten. Spielt die Reihenfolge der Entdeckung für die Bestätigung eine Rolle? Das wird manchmal behauptet und könnte dem Bayesianer ein Stück weit helfen, in seiner Bewältigung des Problems der alten Daten. So wird etwa behauptet, dass Daten, die eine Theorie vorhersagt, diese deutlich besser stützen, als solche Daten, die sie nur nachträglich ableiten oder reproduzieren kann, die aber zum Zeitpunkt ihrer Aufstellung schon bekannt waren. Dazu wird z.B. angenommen, dass wir eine neue Theorie natürlich immer so »hinbasteln« können, dass sie zu den schon bekannten Daten passt. Man kann hier von der These der *Bestätigungsasymmetrie von Prognose und Retrodiktion* sprechen.

Betrachten wir dazu ein besonders unschönes Beispiel: Unsere Theorie T besagt, dass alle Gegenstände zur Erde fallen, sobald wir loslassen und sie nicht durch eine der bekannten Kräfte gehalten werden, bis auf den einen Fall, als Thomas am 13.2.1995 einen Ball losgelassen hat und der in der Luft stehen blieb. Der eine Fall wird schlicht ausgenommen, weil wir schon wissen, dass sich dieser Ball nicht an die Fallregeln in unserer Theorie gehalten hat. Unsere Theorie T ist offensichtlich sehr hässlich und ließ sich nur an die Daten anpassen, weil wir zum Zeitpunkt ihrer Aufstellung schon von diesem Ball-Ereignis E am 13.2.1995 wussten. Wir möchten kaum sagen, dass durch dieses Ball-Ereignis E unsere Theorie bestätigt wird, obwohl sie auch zu diesem Datum passt. Das sähe schon etwas anders aus, wenn wir unsere Theorie bereits 1990 aufgestellt hätten und dabei das spezielle Ball-Ereignis prognostiziert hätten. Unsere Theorie würde sicher mehr Aufmerksamkeit erfahren und man würde fragen, wie uns diese korrekte Prognose gelingen konnte.

Allerdings bliebe das Problem, dass unsere Theorie T das spezielle Ball-Ereignis E nicht wirklich *erklären* kann. Das wird nur als spezielle

Ausnahme beschrieben, ohne dafür spezielle Gesetze zu haben. Daher würde auch ein Vertreter des abduktiven Schließens davon ausgehen, dass T nicht wirklich bestätigt würde durch E. Ein Teil der Aussage von T erweist sich durch E zwar als wahr, aber der Rest von T wird durch E nicht gestärkt – auch intuitiv nicht. Es bleibt aber zumindest ein psychologischer Effekt zurück, dass gerade überraschende Vorhersagen einer Theorie diese besonders stützen, wenn sie denn eintreffen. Ein Beispiel dafür hatten wir oben schon erwähnt, nämlich die Lichtablenkung im Schwerefeld der Sonne, die die allgemeine Relativitätstheorie korrekt vorhergesagt hatte. Sollten wir also die Asymmetriethese unterschreiben?

Ich bin i.A. nicht dieser Meinung. Das hieße nämlich, dass der Zeitpunkt, zu dem wir Daten finden bzw. Theorien entwickeln zu entscheidender erkenntnistheoretischer Bedeutung erhoben würde. Nehmen wir zwei Zeitpunkte 1 und 2, die nacheinander liegen und betrachten Wissenschaftler A und B. Wissenschaftler A beobachtet zunächst E zu 1 und entwickelt dann T zum Zeitpunkt 2. Für Wissenschaftler ist die zeitliche Reihenfolge genau andersherum. Intuitiv soll E eine starke Stützung für H darstellen. Außerdem gleichen sich beide Wissenschaftler in all ihrem weiteren Hintergrundwissen völlig. Dann wären sie zum Zeitpunkt 2 eigentlich auf genau demselben Kenntnisstand, aber durch die unterschiedliche Reihenfolge wäre etwa T für B gut bestätigt und für A nicht so gut bestätigt. Da es für A und B keine weiteren Unterschiede im Hintergrundwissen zu 2 mehr gibt, sehe ich nicht, wie wir diese Asymmetrie sinnvoll verteidigen könnten.

Die obigen Beispiele sind also eher irreführend oder nicht wirklich relevant, weil es sich nicht um eine intuitive Rechtfertigung unserer abstrusen Falltheorie handelt. Dazu müssten wir schon bessere Beispiele und Überlegungen finden, die mir aber für solch einfache Fälle nicht bekannt sind. Sicher bietet der Überraschungseffekt einer erfolgreichen Prognose intuitiv eine besondere Bestätigung, aber eine nüchterne Analyse sollte das mehr als einen psychologischen Effekt betrachten und nicht unsere Bestätigungstheorien daran orientieren. Hier bleibt jedoch sicher Raum für weitere Debatten, die sogar im Rahmen der klassischen Statistik zu finden sind. So hat etwa Gerhard Schurz (2014a) etwas andere Fälle betrachtet, in denen eine Theorie durch die Daten erst fertiggestellt wird, indem die Daten dazu genutzt werden, dort bestimmte Parameter

erst festzulegen. Diese neuen Theorien können dann kaum durch die parameterfestlegenden Daten gestützt werden (vgl. Kap. 3.1).

5.8.4 Neue Theorien und Innovationsfeindlichkeit

Ein weiteres Thema im Umfeld des Problems der alten Daten ist die Aufstellung ganz *neuer* Theorien im Laufe des Forschungsprozesses. Gerade dadurch werden oft besondere Fortschritte in der Wissenschaft erzielt. Das zeigte sich auch in den letzten Beispielen wie der speziellen und der allgemeinen Relativitätstheorie. Doch im Bayesianismus ist in mehrfacher Hinsicht kein Platz für neue Theorien. Der ganze Prozess des Updatens ist so gestaltet, dass die Menge der Aussagen, deren Wahrscheinlichkeiten sich dabei ändern, insgesamt konstant bleiben. Kommen neue Aussagen hinzu, müssen wir jedenfalls ganz neu starten und neue Ausgangswahrscheinlichkeiten festlegen. Insbesondere beziehen wir uns immer wieder auf eine vollständige Menge von Hypothesen $\{H_1,\ldots,H_n\}$, um mit ihrer Hilfe etwa P(E) zu bestimmen, aber auch um jeweils den Bereich der relevanten Hypothesen für einen bestimmten Bereich im Blick zu haben, für den sich die Wahrscheinlichkeiten jeweils zu 1 aufsummieren. Darin ist kein Platz vorgesehen für ganz neue Hypothesen.

Als Notlösung könnten wir vielleicht zunächst daran denken, dass eine Catch-all-Hypothese C in unserer Liste von Hypothesen ist, die wir wiederum aufspalten können, in die neue Hypothese und eine neue Catch-all-Hypothese C*. Doch auch dieses Verfahren stößt irgendwann an seine Grenzen und ist eigentlich nicht unbedingt sinnvoll. Die neue Hypothese mag uns gerade im Lichte der bisherigen Daten besonders plausibel erscheinen und sollte dann vielleicht schon eine höhere Wahrscheinlichkeit erhalten als C. Sie führt im Normalfall eher zu einer kompletten Neubewertung der bisherigen Hypothesen. Dafür gibt es jedoch beim bayesianischen Updaten keinen Platz. Außerdem verlässt man mit dieser Idee natürlich längst den klassischen Bayesianismus. Wir sollten also bei Einführung einer neuen Hypothese einen vollständigen Neustart vornehmen, und erst danach können wir wieder mit dem bayesianischen Updaten einsetzen.

Das deckt sich mit Kuhns Ideen darüber, wie das Auftreten eines neuen Paradigmas unsere bisherigen Bewertungen über den Haufen werfen kann. Wir kennen das ebenfalls aus dem abduktiven Schließen und aus dem Vorgehen der Kohärenztheoretiker. Dahinter stand der Gedanke, dass die Daten nur im Lichte unseres Hintergrundwissens für bestimmte Theorien sprechen und dazu gehört vor allem, über welche alternativen Theorien wir verfügen. Die Fingerabdrücke des Herrn X auf der Mordwaffe sprechen zunächst sehr stark für Herrn X als Täter. Sollte allerdings die Theorie an Plausibilität gewinnen, dass Herr Y am Vortag ganz bewusst die Mordwaffe Herrn X in die Hand gedrückt hat (wobei er überraschenderweise Handschuhe trug, was mehreren Beobachtern aufgefallen ist), so richten sie sich plötzlich stärker gegen Y und nicht mehr so stark gegen X.

Neue Theorien (oder in unserem Beispiel: neue Verdächtige) können solche epistemischen Beziehungen also komplett verschieben. Das kann auch mit einer Uminterpretation der Beobachtungen einhergehen. Während für den Phlogistontheoretiker die Zunahme des Gewichts bei einer Verbrennung dafür spricht, dass Phlogiston ein negatives Gewicht hat, eröffnete die Sauerstofftheorie eine ganz neue Möglichkeit, nämlich dass eine Verbrennung die Vereinigung mit dem Sauerstoff darstellt, der seinerseits natürlich ein normales positives Gewicht aufweist.

Neue Theorien eröffnen zugleich neue Möglichkeiten, die Daten bestimmten Hintergrundphänomenen zuzuordnen, wodurch sich die Gewichte der Daten anders auf die Hypothesen verteilen. Dafür kennt der Bayesianismus allerdings kein Verfahren. Man kann sagen, er ist in diesem Punkt stark konservativ und eher *innovationsfeindlich*. Den Punkt, dass Daten nur im Lichte der Konkurrenzhypothesen mehr oder weniger für bestimmte Hypothesen sprechen, greifen z.B. die Likelihoodisten wieder auf und behaupten, wir könnten immer nur eine komparative Theorienbewertung durchführen, wonach ein Datum E im Lichte unseres Hintergrundwissens H_1 gegenüber H_2 in einem bestimmten Maße plausibler macht, ohne dass wir damit etwas über die Bestätigung von H_1 sagen könnten, was sich nicht auf H_2 bezieht. Darauf kommen wir im nächsten Abschnitt zurück.

Vorher möchte ich noch eine Bemerkung zu einem verwandten Thema machen, nämlich der *Theorienfeindlichkeit* des Bayesianismus. Die

haben schon verschiedene Autoren wie Popper und Glymour betont. Eine Frage, die Bayesianer meist nicht so sehr beschäftigt, ist, was wissenschaftliche Theorien als solche kennzeichnet und wofür wir sie benötigen. Meine Antwort auf die zweite Frage war bereits, dass Theorien vor allem zur Erklärung und auch zu Prognosen dienen sollten, die wir ihrerseits benötigen, um gezielt in die Welt eingreifen zu können. Außerdem sollten akzeptierte Theorien natürlich wissenschaftliches Wissen darstellen bzw. dem so nahe wie möglich kommen. Auf die erste Frage biete ich hier keine ausführliche Antwort an. Jedenfalls sollten Naturgesetze bzw. nomische Muster im Zentrum der wissenschaftlichen Theorien stehen. Aber diese haben noch andere Komponenten (vgl. Bartelborth 1996, Balzer & Moulines & Sneed 1987).

Popper (1984) wies schon darauf hin, dass wir Theorien vor allem für ihren empirischen Gehalt schätzen sollten und daher nach möglichst »tiefen« Theorien suchen. Das hat vor allem mit ihrer Erklärungskraft zu tun. Da Popper allerdings jede Form von induktiver Bestätigung von Theorien für unmöglich hielt, hat er unter den noch nicht falsifizierten Theorien den empirischen Gehalt als alleiniges Kriterium für die Theorienwahl angenommen. Das ist sicherlich zu kurz gegriffen. Die induktive Bestätigung von Theorien darf natürlich nicht außen vor bleiben. Die ist für einen Bayesianer vor allem in der Wahrscheinlichkeit P(H) der Theorie zu sehen. Doch das steht in einem Konflikt mit dem Gehalt und der Tiefe der Theorie. Auch das hatte Popper schon klar gesehen. Bayesianer scheinen nun den zu Popper entgegengesetzten Fehler zu machen. Sie setzen ausschließlich auf die induktive Stützung der Theorien und beziehen ihre jeweilige Erklärungsstärke nicht in die Theorienwahl mit ein. Man geht einfach davon aus, dass man schon über eine geeignete Liste von konkurrierenden Theorien verfügt und nur noch auf deren Plausibilität zu achten hat.

Doch wenn die Wahrscheinlichkeit der Theorien unser einziges Kriterium für die Auswahl der von uns zu akzeptierenden Theorien bleiben sollte, dann setzen sich immer die weniger gehaltvollen Theorien gegenüber den substantiellen, tiefen Theorien durch. Nehmen wir an, wir hätten eine tiefe Theorie T, aus der zahlreiche Beobachtungen oder sogar Phänomene (als Typen von Beobachtungen) $E_1,\ldots,E_n$ logisch folgen. Dann gilt zunächst: $P(T) \leq P(E_1\&\ldots\&E_n)$. Sind die Phänomene

statistische und bisher unabhängig voneinander (in unserer Glaubensfunktion P), so wird dieser Wert vermutlich recht klein sein, da die rechte Seite sich in ein entsprechendes Produkt aufspaltet. In jedem Fall sollten wir erwarten, dass die Theorie T einen Überschussgehalt gegenüber der bloßen Konjunktion der Daten hat. Das hieße aber, dass die »Theorie« $T^* \equiv E_1 \& \ldots \& E_n$ nach dem Updaten mit den Daten $E_1, \ldots, E_n$ die Wahrscheinlichkeit 1 hätte und damit der Theorie T überlegen wäre. T* ist allerdings nur eine Konjunktion der Daten und keine substantielle Theorie. Insbesondere folgen zwar die vorliegenden Daten aus T*, aber T* hat überhaupt keine Erklärungskraft oder Prognosefähigkeit. Dazu fehlt ihr die Angabe eines nomischen Musters, das das Auftreten der Daten erklären könnte und womöglich die kausalen Mechanismen angibt, die zu den Daten $E_1, \ldots, E_n$ geführt haben. Es gilt also $P(T) < P(T^*) = 1$.

Allgemein müssen die oberflächlichen Theorien, die vielmehr Zusammenfassungen von Daten darstellen als tieferliegende Mechanismen zu benennen, vom Bayesianer die höhere Wahrscheinlichkeit erhalten. Schauen wir dann nur auf die Wahrscheinlichkeit und nicht mehr auf die Erklärungskraft der Theorien, werden wir nur die »billigen« und flachen Theorien auswählen und selbst simple Konjunktionen wie T* stehen dann blendend dar. Darin zeigt sich die *Theorienfeindlichkeit des Bayesianismus*, die nicht die tatsächliche Praxis der Theorienwahl in der Wissenschaft widerspiegelt. Wollen wir mit dem Bayesianismus ein Instrument finden, dass uns bei der tatsächlichen Theorienwahl in der Wissenschaft anleiten kann oder sie rekonstruieren kann, so müssen wir den einfachen Bayesianismus dementsprechend ergänzen.

Ein Bayesianer könnte etwa antworten, dass in puncto Wahrscheinlichkeit nur noch die Theorien miteinander verglichen werden sollten, die dieselbe Erklärungskraft haben. Das wäre allerdings ein völlig neues Element im Bayesianismus. Wir müssten zunächst feststellen, welche Theorien dieselbe Erklärungskraft haben, und danach erst unsere Liste ausrichten. Doch die Erklärungskraft von Theorien ist oft nicht so einfach vergleichbar und außerdem wollen wir natürlich weiterhin Theorien mit unterschiedlicher Erklärungskraft weiterhin miteinander vergleichen. Das ist nicht einfach, aber der Kohärenztheoretiker hat dafür bereits Ansätze entwickelt, wie das geschehen kann. Hier muss der Bayesianismus jedenfalls dringend verbessert werden, um seine

Theorienfeindlichkeit abzulegen und das zweite Ziel der Theorienwahl (informative Theorien) in irgendeiner Form berücksichtigen (vgl. Bartelborth 2005). Sonst drohen die bayesianischen Wahrscheinlichkeiten für Theorien zu einer theoretischen Größe ohne praktische Bedeutung zu verkümmern. Einen neuen Weg werden wir später in Achinsteins Zwei-Komponenten Konzeption von Bestätigung finden, die Erklärungszusammenhänge und Wahrscheinlichkeiten miteinander kombiniert. Einen besseren Weg haben wir aber schon im Schluss auf die beste Erklärung kennengelernt, der beide Ziele im Blick behält und die Likelihoods und weitere bayesianische Größen zur Bestimmung der Erklärungsstärke u.a. mit heranziehen kann. Dem kommt der Likelihoodismus noch am nächsten, den wir als Nächstes anschauen werden.

5.8.5 Der Likelihoodismus und das Favorisieren von Hypothesen

Die Likelihoodisten bieten einen Ansatz, der bestimmte Einsichten des Bayesianismus erhält, aber gerade die subjektiven Elemente ausschließen möchte und ebenso wenig anfällig ist, für das Problem der alten Evidenz oder das der Theorienfeindlichkeit. Er argumentiert wie folgt: Startwahrscheinlichkeiten mag jemand zu seinem persönlichen Vergnügen einsetzen, aber wieso sollten die in die *wissenschaftliche* Theorienbewertung eingehen? Sollten wir dafür die Meinung eines bestimmten Experten hernehmen oder soll jeder seine eigene wählen? Dann verhelfen uns die Daten jedenfalls nicht zu einer wissenschaftlichen Übereinstimmung oder einem wissenschaftlichen Fortschritt. Die Gegenseite der Umweltschützer in unserer obigen Glosse wird etwa die Startwahrscheinlichkeit $P(T) = 1 - 10^{-(10\hat{}10000)}$ für korrekt halten, und es wird so nie eine Annäherung zwischen beiden Seiten stattfinden, obwohl man sich über die realen Likelihoods eigentlich einig ist, die sich aus den Theorien ergeben, und man diese im Sinne der Likelihoodanbindung ebenso für wissenschaftlich relevant hält. Der Bayesianer wird durchaus zugestehen, dass die Likelihoods uns darüber Auskunft geben sollen, in welchem Ausmaß spezielle Daten bestimmte Theorien stützen. Vor allem die Likelihoods sind daher das Objektive an der probabilistischen Herangehensweise. Darin stimmen uns ebenfalls die klassischen Statistiker zu. Der Likelihoodist stellt sich daher auf

den Standpunkt, wir dürften nur die objektiven, direkten Likelihoods heranziehen, um Theorien epistemisch zu bewerten. Das äußert sich zunächst in einigen Prinzipien, mit deren Hilfe wir den Likelihoodismus charakterisieren können. Da ist zunächst das Likelihoodgesetz:

Das Likelihoodgesetz: Für zwei einander ausschließende Hypothesen h_1 und h_2 und Daten e gilt:
(a) e *favorisiert* h_1 gegenüber h_2 gdw. $P(e|h_1\&b) > P(e|h_2\&b)$ und
(b) der *Likelihoodquotient* $P(e|h_1\&b)/P(e|h_2\&b)$ misst, in welchem *Maße* e bei gegebenem Hintergrundwissen b h_1 gegenüber h_2 favorisiert.

Manchmal wird auch nur die Behauptung (a) als Likelihoodgesetz bezeichnet. Als Likelihoodprinzip wird dann die tendenziell stärkere Behauptung bezeichnet, dass *alles, was wir über die Begründung von Theorien sagen können, in den Likelihoods enthalten ist* (vgl. MacKay 2005, 32). Jedenfalls sind an dieser Stelle für einen echten Likelihoodisten (wie etwa Edwards 1992 und Royall 1997) nur objektive Likelihoods zugelassen, die sich nicht in irgendeiner Weise auf subjektive Wahrscheinlichkeiten stützen müssen. Wir werten im Wesentlichen nur aus, was uns die Theorien über die Daten sagen und ziehen noch weiteres objektives Hintergrundwissen mit dazu heran.

Die erste substantielle Konsequenz aus dem Likelihoodgesetz ist, dass die Theorienbewertung nur noch einen *komparativen Charakter* hat. Wir können jeweils zwei Hypothesen daraufhin vergleichen, wie gut sie zu den Daten passen, aber nicht mehr bestimmen, wie gut sie absolut gerechtfertigt sind. Das harmoniert wieder mit der Kohärenzauffassung und dem abduktiven Schließen, denn auch dort liegt das Schwergewicht der Auswertung von Hypothesen in dem *Vergleich*, wie gut die Daten zu den konkurrierenden Theorien passen.

Dabei ergeben sich zwei Hauptargumente, die für das Likelihoodgesetz (oder die Likelihood-Intuition) sprechen. Für den Bayesianer finden wir den folgenden Zusammenhang für einen Vergleich der Wahrscheinlichkeiten der beiden Hypothesen vor und nach dem Updaten mit E:

$$\frac{P(H_1|E)}{P(H_2|E)} = \frac{P(H_1)}{P(H_2)} \cdot \frac{P(E|H_1)}{P(E|H_2)}$$

Die Nachher-Wahrscheinlichkeiten der Hypothesen verschieben sich also beim Updaten mit E genau dann zugunsten von H_1, wenn gilt: $P(E|H_1) > P(E|H_2)$.

Aber wir sind nicht auf subjektive baysianische Wahrscheinlichkeiten angewiesen. Der Schluss auf die beste Erklärung und unsere entsprechenden Intuitionen sprechen ebenfalls für das Likelihoodgesetz. Prima facie erklärt H_1 das Datum E besser als H_2, wenn die Ungleichung $P(E|H_1) > P(E|H_2)$ gilt. Das zeigt sich auch in vielen Beispielen. Vermutet der Arzt, dass die Masern oder eine Grippe vorliegen, wird er das Auftreten roter Hautflecken als Indiz für die Masern ansehen, weil es bei Masern viel wahrscheinlicher ist, dass sie auftreten, als bei Grippe. Oder nehmen wir an, wir hätten zwei Hypothesen über eine Münze. Die erste besagt P(Kopf) = 0,5 und die zweite P(Kopf) = 0,8. Wenn nun von 100 Würfen mit der Münze 85-mal Kopf fällt, spricht das für die zweite Hypothese, weil unser Datum viel wahrscheinlicher ist bei Wahrheit der zweiten Hypothese. Die Likelihood-Intuition ist also sehr gut nachvollziehbar. Trotzdem gibt es natürlich auch Probleme für den Likelihoodismus.

Für disjunktive Hypothesen erhalten wir normalerweise so keine (objektiven) Likelihoods mehr und können uns etwa nur durch Abschätzungen für die möglichen Likelihoods behelfen, um damit noch bestimmte Aussagen treffen zu können oder aber wir sind auf bayesianische subjektive Wahrscheinlichkeiten für die einzelnen Hypothesen angewiesen. Die Idee ist aber klar, dass der objektive Anteil beim bayesianischen Updaten von den objektiven Likelihoods herrührt und man sich eben auf diesen Anteil in der Theorienbewertung beschränken möchte.

Oftmals geht man noch zum natürlichen Logarithmus des Likelihoodquotienten über, um eine symmetrische Skala um die Null herum zu erhalten. Außerdem können unabhängige Log-Likelihoods dadurch aufaddiert werden. Doch das sind nur noch technische Feinheiten, die ich nur erwähne, damit sich der Leser schon darauf einstellen kann, wenn er in der Literatur darauf stößt.

Da keine subjektiven Wahrscheinlichkeiten für den Likelihoodisten zugelassen sind, tritt nun das Problem der alten Evidenz nicht mehr in Erscheinung. Ist ein Datum E im Lichte der Theorie H_1 wahrscheinlicher als aus Sicht der Theorie H_2, dann spricht das Auftreten dieses Datums

mehr für H_1 als für H_2, ganz gleich, wann uns das Datum bekannt wurde. Hier wird sozusagen eine perfekte Likelihoodanbindung umgesetzt und man verzichtet auf die weiteren subjektiven Bestandteile, die uns im Bayesianismus begegnen. Allerdings bleiben dann z.B. auch die theoretischen Zusammenhänge außen vor. Wie können wir unsere Theorie T etwa bewerten, wenn wir z.B. andere Theorien akzeptieren, die nicht gut mit T zusammenpassen? Über diese holistischeren Aspekte der Theorienbewertung weiß uns der Likelihoodist nichts zu sagen.

Außerdem wurden auch hier Gegenbeispiele vorgeschlagen (vgl. Fitelson 2007, 2013), von denen wir eines genauer betrachten wollen. Es wird eine Spielkarte aus einem gewöhnlichen Kartenspiel (etwa mit 52 Karten) nach guter Durchmischung gezogen. Dann erfahren wir etwas über die Spielkarte und haben zwei Hypothesen:

Gegenbeispiel zum Likelihoodismus (Fitelson 2007, 2013)
h_1 = die Karte ist Kreuzass; h_2 = die Karte ist schwarz;
e = die Karte ist ein Kreuz;
Dann gilt: $P(e|h_1) = 1 > P(e|h_2) = ½$

Favorisiert nun unsere Information e, dass es sich bei der Karte um ein Kreuz handelt, tatsächlich die Hypothese h_1, dass es ein Kreuzass ist, gegenüber der Hypothese h_2, dass die Karte schwarz ist? Das scheint nicht der Fall zu sein, denn aus e folgt schließlich deduktiv h_2, aber nicht h_1. Fitelsons Beispiel scheint demzufolge das Likelihoodgesetz sogar in beide Richtungen in Frage zu stellen. Wir werden das Gesetz wohl durch zusätzliche Forderungen abschwächen müssen. Von einer komparativen Bestätigung kann man danach nur dann sprechen, wenn weitere Anforderungen erfüllt sind.

Als Verfechter der Abduktion liegt es natürlich nahe zu verlangen, dass die Hypothesen die Daten auch *erklären* müssen, um durch die Daten bestätigt zu werden. Für wissenschaftliche Beispiele, die der Likelihoodist vor allem im Blick hat, scheint das eine recht plausible Forderung zu sein. Fitelson sucht dagegen eher nach einer reinen Wahrscheinlichkeitskeitsbedingung etwa aus dem bayesianischen Lager, um zumindest ein schwaches Likelihoodgesetz formulieren zu können.

Er betrachtet daher zunächst, was der Bayesianer über das Favorisieren sagen kann. Dabei eignet sich die folgende simple Regel leider nicht:

SR: e *favorisiert* h_1 gegenüber h_2 gdw. $P(h_1|e) > P(h_2|e)$

Die Regel SR hat den offensichtlichen Nachteil, dass sie nicht verlangt, dass e überhaupt eine besondere Relevanz für h_1 oder h_2 besitzen muss. Es könnte sogar der Fall sein, dass die Wahrscheinlichkeit von h_1 durch Updaten mit e sinkt und die von h_2 steigt, nur dass hinterher immer noch die Wahrscheinlichkeit von h_1 höher ist als die von h_2, weil sie vorher schon deutlich höher war. Der Bayesianer wird also für die komparative Bestätigung wieder ganz auf die Maße der Bestätigung setzen müssen, die wir in Kapitel 5.5.8 bereits diskutiert haben. Nehmen wir an, wir haben einen Bestätigungsbegriff B(h,e), wonach h durch e in einem bestimmten Maße B(h,e) bestätigt wird, dann lässt sich daraus auch das Favorisieren ableiten:

Bayesianisches Favorisieren
e favorisiert h_1 gegenüber h_2 gdw. $B(h_1,e) > B(h_2,e)$

Das wirft allerdings wieder das Problem auf, mit welchem Maß wir arbeiten sollen. Setzen wir für B z.B. das Ratio-Maß ein, erhalten wir genau den Teil (a) des Likelihoodgesetzes. Fitelson knüpft dagegen an sein Likelihood-Maß an und erhält so das folgende schwache Likelihoodgesetz:

Schwaches Likelihoodgesetz (Fitelson)
e favorisiert h_1 gegenüber h_2 gdw.
(1) $P(e|h_1) > P(e|h_2)$ und
(2) $P(e|\neg h_1) \leq P(e|\neg h_2)$

Damit wird sich der strikte Likelihoodist allerdings nicht wirklich anfreunden können, weil er die »Catch-all-Likelihoods« ($P(e|\neg h_1)$ und $P(e|\neg h_2)$ in der zweiten Ungleichung nicht für objektiv hält. Doch wenn wir nicht so strikt sind, haben wir zunächst eine Konzeption des Favorisierens gefunden, die die bekannten Gegenbeispiele zurückweist. In unserem obigen Gegenbeispiel ist sie offensichtlich nicht erfüllt, denn es gilt dort: $P(e|\neg h_1) = 13/52$ und $P(e|\neg h_2) = 0$. Das schwache Likelihoodgesetz stellt so einen Kompromiss dar. Fitelson betont, dass wir für die zweite

Ungleichung oft nur eine Plausibilitätsabschätzung benötigen und nicht unbedingt genaue Werte. Außerdem benötigen wir dann keine Vorher-Wahrscheinlichkeiten vom Typ P(h). Das macht diese Explikation des Favorisierens selbst für Nicht-Bayesianer annehmbar. Es soll jedenfalls im Folgenden immer wieder einmal als Richtschnur für die komparative Theorienbestätigung dienen.

Einen noch einfacheren Weg, Gegenbeispiele dieser Art zu vermeiden schlägt Jake Chandler (2013) vor, nämlich das Likelihoodgesetz dadurch abzuschwächen, dass nur solche Hypothesen miteinander verglichen werden, die einander ausschließen (vgl. dazu Fitelson (2013)). Das passt auch sehr gut zur Anwendung des Likelihoodgesetzes im Rahmen des Schlusses auf die beste Erklärung, denn im Normalfall betrachten wir dort nur tatsächlich konkurrierende Hypothesen, so dass diese Bedingung automatisch erfüllt ist.

Schwaches Likelihoodgesetz (Chandler)
e favorisiert h_1 gegenüber h_2 gdw.
(1) $P(e|h_1) > P(e|h_2)$ und
(2) $h_1$und h_2 schließen einander aus

Man kann geradezu sagen, dass die Grundidee der Likelihoodisten in allen Ansätzen zum induktiven Schließen in der einen oder anderen Form wiederzufinden ist. Danach bestätigt ein Datum e eine Hypothese h ceteris paribus umso mehr, je stärker die Hypothese h dieses Datum vorhersagt, je größer also die objektive Wahrscheinlichkeit P(e|h) ist. Den Grenzfall kennen wir aus der hypothetisch-deduktiven Theorienbestätigung, aber wir finden diesen Gedanken auch im abduktiven Schließen wieder und sogar im Bayesianismus. Sogar die klassische Statistik setzt ganz auf diese objektiven Likelihoods. Es verwundert also nicht, dass die Likelihoodisten das dann zum alleinigen Maßstab erheben wollen. Allerdings hatte die Debatte um das abduktive Schließen schon deutlich herausgestellt, dass es gute Gründe gibt, noch weitere Aspekte bei der Theorienwahl zu berücksichtigen. Das wird noch einmal bestätigt durch das Gegenbeispiel von Fitelson.

Halpern & Pucella (2006) haben sogar eine Logik für die Auswertung von Daten e für eine Liste $L = \{h_1,\ldots,h_n\}$ von Hypothesen angegeben, bei

der man sich zunächst nicht auf zwei Hypothesen beschränken muss und außerdem die Likelihoods normiert werden, damit die Daten wieder ein Gewicht zwischen 0 und 1 für die Hypothesen aufweisen. Dabei wählen sie als naheliegendes Maß $w(e,h_i)$ für das Gewicht der Daten e etwa für h_i:

Normiertes Gewicht der Daten: $w(e,h_i) = P(e|h_i) \;/\; [\sum_{1\leq j\leq n} P(e|h_j)]$

Dieses Maß wäre natürlich auch geeignet, um die Kohärenzwerte in einem größeren Rahmen mit mehr als zwei Szenarien zu bestimmen und könnte so meinen Ansatz aus Kapitel 4.4 verallgemeinern.

5.8.6 Bayesianismus für unendlich viele Hypothesen

In der bayesianischen Statistik arbeitet man typischerweise mit einer Parametrisierung von Hypothesen (vgl. etwa Held 2008 und Kap. 6.4.2). Nehmen wir etwa eine Münze, die wir werfen, und gehen davon aus, dass sie eine feste, aber unbekannte Wahrscheinlichkeit θ dafür aufweist, dass Kopf kommt. Dann können wir die Menge der möglichen Hypothesen $h_{\theta\in\Theta}$ als die Hypothesen betrachten, für die $P(\text{Kopf}|h_\theta) = \theta$ ist mit $\Theta = [0,1]$. Wir haben damit ein Kontinuum von Hypothesen, die wir normalerweise direkt durch den Modellparameter θ kennzeichnen können, ohne jeweils h_θ zu schreiben. Für die Hypothesen benötigen wir eine Wahrscheinlichkeitsdichte $p(\theta)$, die die Wahrscheinlichkeit bestimmt, dass eine der Hypothesen aus einem Intervall $I = [a,b]$ die richtige Hypothese ist:

$$P(\theta \in I) := \int_I p(\theta)\, d\theta \;\text{ mit }\; \int_\Theta p(\theta)\, d\theta = 1$$

Auch hier können wir wieder mit Daten x updaten. Dazu benötigen wir die entsprechenden Likelihoods $f(x|\theta)$ und erhalten zunächst wieder die bayessche Formel in der Proportionalschreibweise, die es uns oft ermöglicht, die Art der resultierenden Nachher-Verteilungen zu bestimmen, ohne uns um die meist schwer zu berechnenden Proportionalitätskonstanten Gedanken machen zu müssen:

Bayessches Theorem für Dichten 1
$p^+(\theta) := p(\theta|x) \propto p(\theta)\cdot f(x|\theta)$

Auch hier muss die neue Dichte p^+ wieder eine Wahrscheinlichkeitsdichte sein und somit finden wir den entsprechenden Proportionalitätsfaktor und erhalten dann das klassische bayessche Theorem:

Bayessches Theorem für Dichten 2

$$p^+(\theta) = p(\theta|x) = \frac{p(\theta) \cdot f(x|\theta)}{\int_\Theta p(\theta) \cdot f(x|\theta)\, d\theta}$$

Die bayesianische Statistik versucht i.A., einen möglichst objektiven Bayesianismus zu vertreten, und zu diesem Zweck die Ausgangswahrscheinlichkeiten möglichst objektiv zu gestalten. Wissen wir etwa nur wenig über den gesuchten Parameter θ, so sollte die Vorher-Dichte möglichst uninformativ sein. Es sollen letztlich nur die Daten sprechen. Dazu werden wieder Überlegungen und Hilfsmittel aus der Informationstheorie eingesetzt (vgl. Bernardo 2009 zur »reference analysis«). In einfachen Fällen kommt etwa das Maximum-Entropie-Prinzip zur Anwendung, das wir bereits geschildert haben. Allerdings haben wir schon auf das Problem der Anwendung von Indifferenzprinzipien verwiesen, das sich immer auf eine ganz bestimmte Beschreibung der Situation bezieht. In der Statistik spricht man von dem statistischen Modell, mit dem man eine Situation repräsentiert. Dazu gehören etwa die grundlegenden Konzepte, mit denen wir die Situation beschreiben und zumindest eine Likelihoodverteilung für die Daten. Wie schwierig es ist, eine Situation zunächst gut zu modellieren, hat sich beim berühmten Ziegenparadox gezeigt, das wir in Kapitel 5.9.1 kurz besprechen werden. Das belegt zugleich, wie viel bereits auf einer sehr basalen Ebene von einer angemessenen Beschreibung abhängen kann. In diesem Fall sind sogar namhafte Statistiker in die Irre gegangen.

Andererseits wissen wir in unserem Münzenbeispiel schon, dass die Wahrscheinlichkeit dafür, dass die Münze extreme Wahrscheinlichkeiten für Kopf aufweist, eher gering sein muss. Wie sollte denn eine Münze konstruiert sein, die in einer normalen Wurfumgebung praktisch immer Kopf ergibt? Das heißt, wir werden normalerweise in einer Vorher-Verteilung schon eine gewisse Bevorzugung der mittleren Werte einbauen, da sie unserem Hintergrundwissen über diesen Vorgang am besten entspricht.

Das lässt sich am besten erkennen, indem wir uns ersten Beispielen zuwenden. Ein Beispiel findet sich in Bernardo (2009). Nehmen wir an,

unsere Daten wären n Beobachtungen eines Bernoulli-Experiments mit dem unbekannten Parameter θ und r positiven Ergebnissen. Es könnte sich also um n Würfe einer Münze handeln mit r-mal Kopf und der unbekannten Wahrscheinlichkeit θ für Kopf. Nehmen wir außerdem an, dass unser Wissen über θ durch eine Betaverteilung Be(α,β) gegeben sie, so dass also gilt:

$$P(\theta|\alpha,\beta) \propto \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}$$

Dann ist die Nachher-Verteilung leicht zu bestimmen mit Hilfe des bayesschen Satzes in Proportionalform. Wir müssen dazu nämlich nur mit der Likelihood multiplizieren, dass bei Vorliegen von θ r-mal ein positives und (n-r)-mal ein negatives Resultat auftritt und das ist gerade: $\theta^{r} \cdot (1-\theta)^{n-r}$. Also entsteht eine neue Betaverteilung mit:

$$P(\theta|n,r,\alpha,\beta) \propto \theta^{r+\alpha-1} \cdot (1-\theta)^{n-r+\beta-1}$$

Schwierig zu berechnen sind allerdings die genauen Normierungsfaktoren. Man kann zumindest schon erahnen, was mit der ursprünglichen Glockenkurve dabei passiert. Sie wird im Normalfall deutlich steiler und konzentriert damit ihre Hauptmasse auf ein kleineres Gebiet.

Im konkreten Beispiel sei θ der Anteil der positiven Stimmen in einem Referendum und unsere Daten D bestehen in einer Umfrage (vgl. Bernardo 2009). Hier muss man allerdings schon erste Idealisierungen akzeptieren, um das als ein Beispiel für ein Bernoulli-Experiment zu betrachten. Ausgangspunkt sei eine Betaverteilung Be(50,50), bei der die Wahrscheinlichkeit für einen Sieg des Referendums bei 0,5 liegt. Nun werden 1500 Personen befragt, von denen 720 sich für das Referendum aussprechen und 780 dagegen sind. Damit hat die Nachher-Wahrscheinlichkeit eine Be(770,830) Verteilung und es wird:

$$P(\theta<0{,}5|D) = 0{,}933$$

Das heißt, dass nun die Wahrscheinlichkeit für einen Misserfolg des Referendums bereits auf 93% angewachsen ist.

Auf diese Art setzt der Bayesianer die Nachher-Verteilungen ein, um die induktiven Auswirkungen der Daten zu beschreiben. Viele Aussagen sind

nun möglich, die einem klassischen Statistiker verwehrt bleiben. Sehr deutlich wird das für die Kredibilitätsintervalle, die das bayesianische Gegenstück zu den Konfidenzintervallen des klassischen Statistikers sind. Wenn wir für den Parameter θ wissen möchten, ob er in ein bestimmtes Intervall I fällt, und wir verfügen über eine Vorher-Verteilung f und besitzen bereits Daten x darüber, dann gilt:

$$P(\theta \in I|x) = \int_I p(\theta|x)\, d\theta$$

So können wir auch für θ bestimmte (1–α)-*Kredibilitätsintervalle* I konstruieren, indem wir verlangen, dass $P(\theta \in I|x) = 1-\alpha$ gilt. Für die drückt unsere Gleichung direkt aus, dass der gesuchte Parameter θ mit der Wahrscheinlichkeit 1–α in I liegt. Allerdings gibt es in aller Regel natürlich viele entsprechende (1–α)-Intervalle. Dazu können wir weitere Anforderungen formulieren, um das weiter einzugrenzen. Naheliegend wäre etwa, dass das Intervall möglichst kurz sein sollte, denn dort konzentriert sich dann die Hauptmasse der Verteilung. Ähnliche Probleme finden sich aber ebenso für Konfidenzintervalle (vgl. Kap. 6.6.2).

Der große Vorteil der Kredibilitätsintervalle ist die einfachere Interpretation. Wir werden in Kapitel 6 sehen, dass der klassische Statistiker nicht sagen darf, dass das Konfidenzintervall zum Niveau 95% mit einer Wahrscheinlichkeit von 95% den wahren Wert enthält. Zumindest wird dafür der statistische Syllogismus benötigt, der selbst eher ein Grundprinzip der induktiven Logik darstellt und dem klassischen Statistiker eigentlich nicht zur Verfügung steht. Ohne den statistischen Syllogismus können wir dann aus Sicht der klassischen Statistik (bzw. der Nicht-Probabilisten) eigentlich keine speziellen Aussagen über einzelne Konfidenzintervalle mehr treffen. Für den Bayesianer sieht das mit den Kredibilitätsintervallen anders aus. Er behauptet ganz direkt, dass gegeben bestimmte Daten und ein bestimmtes Wahrscheinlichkeitsmodell das 95%-Kredibilitätsintervall den gesuchten Parameter mit einer Wahrscheinlichkeit von 95% enthält.

Es bleibt natürlich das Problem der Vorher-Verteilungen. Bayesianer sind sogar bereit, um zu möglichst uninformativen Verteilungen zu kommen, auch uneigentliche Dichtefunktionen zuzulassen. Das sind Dichten, deren Integrale über den gesamten Raum unendlich werden.

Eine Gleichverteilung etwa auf den positiven reellen Zahlen liefert keine eigentliche Dichte mehr. Allerdings sollte zumindest nach dem Updaten mit den Daten wieder eine echte Dichtefunktion entstehen, damit wir damit etwas anfangen können, und daher sind auch für die uneigentlichen Dichten weitere Restriktionen zu beachten. Dazu diskutieren Bayesianer unterschiedliche Ansätze, aber das geht dann stark in technische Gefilde, die wir hier nicht weiter verfolgen können.

Dazu gehört auch, dass man mit sogenannten Nuisance-Parametern bzw. störenden Parametern im Bayesianismus im Prinzip einfach umgehen kann, weil man über die entsprechenden Wahrscheinlichkeitsverteilungen verfügt, um sie »wegzuintegrieren«. Nehmen wir als einfaches Beispiel an, wir hätten für jede Körpergröße y eines amerikanischen Manns eine Wahrscheinlichkeitsverteilung für sein Gewicht x. Aber leider wissen wir nicht, wie viele Amerikaner jeweils eine bestimmte Größe y aufweisen. Wenn wir nun etwa abschätzen wollen, wie schwer amerikanische Männer im Durchschnitt sind, erweist sich der Parameter Körpergröße y als störender Parameter. Ich könnte etwa sagen, dass die 1,85m großen Amerikaner im Durchschnitt 88 kg wiegen, aber wenn auch noch die Wahrscheinlichkeiten für bestimmte Körpergrößen unbekannt ist, hilft mir das nicht recht weiter. Wie werden wir den Parameter y nun los? Bayesianer starten einfach mit einer Vorherdichte p(x,y), die uns eine gemeinsame Verteilung für Größe und Gewicht liefert. Eine entsprechende Wahrscheinlichkeitsverteilung für das Gewicht x erhalten wir dann durch Integration über y:

$$f^*(x) = \int p(x,y)\, dy$$

Entsprechend erhalten wir dann für weitere vorliegende Daten D auch noch eine Nachherdichte p(x,y|D). Jedenfalls können wir nun durch Aufsummieren über die verschiedenen Größen gewichtet mit ihren jeweiligen Auftretenshäufigkeiten zu einer reinen Dichte für die Gewichte gelangen:

$$p^*(x|D) = \int p(x,y|D)\, dy$$

Damit haben wir eine neue Dichte ausschließlich für das Gewicht gefunden, die auch noch Daten D mit einbeziehen kann. Was ist nun die

zu erwartende mittlere Größe? Hierfür können wir als *Punktschätzung* einfach den Erwartungswert für p* bestimmen:

$$E(p^*(x|D)) = \int x \cdot p^*(x|D)\, dx$$

So hat der Bayesianer ein Problem im Prinzip schnell und einfach gelöst, das den klassischen Statistiker immer wieder plagt und zu allerlei Tricks greifen lässt.

Betrachten wir ein einfaches Zahlenbeispiel dazu mit nur wenigen diskreten Werten für die Größe (1,65 m und 1,85 m) und das Gewicht (70 kg und 80 kg und 90 kg). Andere Werte werden in unserem Beispiel einfach nicht angenommen. Nehmen wir weiter an, dass wir die Propensitäten kennen, mit denen jemand, der etwa 1,65 m groß ist, die jeweiligen Gewichtsstufen einnimmt: zu 50%: 70 kg, zu 33%: 80kg und zu 17%: noch 90 kg. Entsprechendes wissen wir auch für die 1,85 m großen Männer: zu 7% 70 kg, zu 57% 80 kg und zu 36% 90 kg. Wenn wir nun aber nicht wissen, wie viele Männer 1,65 m groß sind und wie viele die Größe von 1,85 m erreichen, dann wissen wir z.B. nicht, was das durchschnittliche Gewicht der amerikanischen Männer ist. Diese Werte ergänzt der Bayesianer nun durch subjektive Wahrscheinlichkeiten. Er geht etwa davon aus, dass 30% der Männer 1,65 m und 70% dann 1,85 m groß sind. Um das Ganze übersichtlich zu gestalten, nehmen wir an, wir hätten eine idealisierte Stichprobe von 1000 Männern, die sich exakt nach den objektiven und den subjektiven Wahrscheinlichkeiten richtet, so dass unsere gemeinsame Verteilung in Häufigkeiten wie folgt aussieht:

	1,65 m	1,85 m	$\sum$
70 kg	150 (50%)	50 (7%)	200
80 kg	100 (33%)	400 (57%)	500
90kg	50 (17%)	250 (36%)	300
$\sum$	300	700	1000

Tabelle 5.10: Beispiel für Randwahrscheinlichkeiten

Daran können wir nun direkt erkennen, welche Anteile den unterschiedlichen Gewichtsklassen zukommen: 20% der Männer sind 70 kg schwer, 50% 80 kg schwer und 30% wiegen sogar 90 kg. Das ist die Randverteilung

(bzw. marginale Wahrscheinlichkeit) für unser Gewichtsproblem und wir können auch ausrechnen, was das Durchschnittsgewicht der Männer ist: 0,2·70 + 0,5·80 + 0,3·90 = 81. Im Durchschnitt ist der amerikanische Mann demnach 81 kg schwer. Man sieht auch leicht, dass dieses Ergebnis erheblich von der Verteilung auf die beiden Größen 1,65 m und 1,85 m abhängt. Die Verteilung in der rechten Spalte ist dann jedenfalls die Verteilung ohne den störenden Parameter *Größe*. Die oben angegebenen Integrale stellen die Aufsummierung für den Fall einer feineren Aufteilung auf die beiden Parameter dar. Wir haben dieses Verfahren bereits im Umgang mit den bayesianischen Netzen kennengelernt.

Der klassische Statistiker wird dieses Verfahren allerdings trotz der schönen Berechnung kaum als genuinen Vorteil des Bayesianismus akzeptieren, denn der einfache Weg, die störenden Parameter loszuwerden, kommt hier nur dadurch zustande, dass die fehlenden Daten einfach subjektiv geschätzt werden. Dazu kann er entgegnen, dass er das natürlich auch könnte, aber so eine skrupellose Ersetzung fehlender Daten durch subjektive Größen für die Wissenschaft nicht akzeptieren möchte.

Trotzdem punktet der Bayesianismus in all den genannten Fällen zunächst einmal. Auf den Einwand der zu subjektiven Schätzungen antwortet er mit der erwähnten Konvergenz und damit, dass auch der klassische Statistiker an manchen Stellen auf subjektive Einschätzungen angewiesen ist. Trifft man diese aber wie es der Bayesianer vorgibt, dann lösen sich viele unserer Interpretations- und Verständnisprobleme auf recht einfache Weise. Einfach jedenfalls im Prinzip, da die konkreten Berechnungen oft nur numerisch durchzuführen sind und schon erhebliche Kalkulationsprobleme bieten können. Wir werden in Kapitel 6 sehen, wie schwierig es an entsprechender Stelle die klassischen Statistiker haben, ihre Methoden zu rechtfertigen und ihre Resultate zu interpretieren. Allerdings kommen vorher auch noch einige Fragen auf den Bayesianer zu.

Welche Hypothesen werden übrigens durch den Bayesianer in diesem Verfahren ausgewählt? Typischerweise werden nun nicht mehr einzelne Hypothesen ausgewählt, sondern im Sinne der Kredibilitätsintervalle ganze Mengen von benachbarten Hypothesen, über die wir nur sagen

können, dass die wahre Hypothese nach unserer Überzeugung (unserem P bzw. der Dichte f) mit hoher Wahrscheinlichkeit in dieser Menge zu finden sein wird. Dabei erwarten wir normalerweise eine immer stärkere Konzentration der Wahrscheinlichkeitsmasse auf einen immer kleineren Bereich, der die richtige Theorie also immer genauer spezifiziert. Das wirft wieder neue Fragen auf, wie wir das in unseren Entscheidungen umsetzen können und wie der Zusammenhang zu unseren klassischen Überzeugungssystemen aussieht. Die bayesianische Statistik ist aber technisch zu anspruchsvoll, um diesen Zweig des Bayesianismus hier in seinen Details weiter zu verfolgen. Wir können aber zumindest noch einige konkrete Beispiele ansehen und noch darauf aufmerksam machen, dass die bayesianische Statistik zugleich ein altes Problem induktiven Schließens lösen kann.

5.8.7 Das Gewicht der Daten

Es gab schon frühzeitig den Einwand, dass unsere Wahrscheinlichkeiten nicht alle Formen der Unsicherheit, die wir über bestimmte Größen haben, richtig zum Ausdruck bringen. Betrachten wir dazu zunächst drei Situationen: In S_1 haben wir für eine bestimmte Münze keinen speziellen Grund für die Annahme, dass sie verfälscht sei. Also vermuten wir, dass die Wahrscheinlichkeit für Kopf gerade ½ ist. In S_2 haben wir dazu noch bestimmte Daten für unsere Münze, nämlich 10 Würfe mit 5-mal Kopf. Auch hier würden wir vermuten, dass P(Kopf) = ½ ist. In S_3 verfügen wir über sehr viel stärkere Daten, nämlich 10000 Würfe mit 5000-mal Kopf als Resultat. Schlussfolgerung wiederum: P(Kopf) = ½. In unseren Wahrscheinlichkeiten spiegelt sich hier nicht unsere gesamte Unsicherheit wieder, die sich von Situation zu Situation schließlich stark verändert. Man könnte sagen, dass das *Gewicht der Daten* hier nicht repräsentiert wird. In S_1 ist das noch sehr gering, da wir bestenfalls einen symmetrischen Eindruck der Münzen haben oder sogar über überhaupt keine Daten verfügen, während es dann bis hin zu S_3 zunimmt. Doch wo findet sich das in unserer Darstellung der Unsicherheiten wieder? Hierauf hat der bayesianische Statistiker eine gute Antwort: Wenn wir mit einer relativ uninformativen Dichte starten, dann zeigt sich das Gewicht der Daten in der zunehmenden Konzentration der upgedateten Dichten.

Man könnte sagen, die kleiner werdende Streuung der neuen Dichten wäre ein Maß für das Gewicht oder die Stärke der Daten.

Das lässt sich in einem konkreten Beispiel gut erkennen, das zugleich etwas über das Konvergenzverhalten in solchen Beispielen sagt. Nehmen wir wieder an, wir suchten nach der Wahrscheinlichkeit für eine Münze Kopf zu ergeben. Nun starten wir mit drei unterschiedlichen Vorherdichten, die auf verschiedene Weise unsere Anfangsunsicherheit ausdrücken sollen (vgl. dazu Büchter & Henn 2006 465 ff.):

$f_1(x) := 2 - |4 \cdot x - 2|$
$f_2(x) := 140 \cdot (1-x)^3 \cdot x^3$
$f_3(x) :=$ 1 für alle $x \in [0;1]$ sonst 0.

Das sind drei Arten unsere Vorher-Unsicherheit zum Ausdruck zu bringen. Die erste Funktion bildet ein symmetrisches Dreieck über dem Einheitsintervall, die zweite einen entsprechenden Hügel und die dritte einfach eine Gleichverteilung. Die ersten beiden Vorherdichten tragen der Tatsache Rechnung, dass es recht unplausibel erscheint, dass die Münze (bei unseren normalen Arten zu werfen) so konstruiert ist, dass nur noch Kopf oder nur noch Zahl erscheint. Im dritten Modell gehen wir einfach einmal davon aus, dass wir nichts konkretes über unsere Vorherdichte aussagen können und setzen nach dem reinen Indifferenzprinzip auf eine völlige Gleichverteilung. Nun wird mit dem Datum D upgedatet: D ≡ Bei 100 Würfen ist 58-mal Kopf gekommen. Damit erhalten wir als Nachher-Dichten in allen drei Fällen relativ spitz verlaufende Berge:

$f_1(x|D) \approx [(2 - |4 \cdot x - 2|) \cdot 100!/0{,}0166 \cdot 58! \cdot 42!] \cdot x^{58} \cdot (1-x)^{42}$
$f_2(x|D) \approx (10^{33}/4{,}95) \cdot (1-x)^{45} \cdot x^{61} f_3(x|D) \approx (10^{31}/3{,}5) \cdot (1-x)^{42} \cdot x^{58}.$

Das zeigt schon eine deutliche Annäherung der Dichten bei nur wenigen Daten. Wie stark sich die Daten nun auf einen bestimmten Bereich konzentrieren, erkennen wir, indem wir ein bestimmtes Intervall I = [0,52;0,64] betrachten und die Dichten darüber integrieren: $P_i(x \in I|D) = \int_I f_i(x|D)\, dx$. Dabei zeigt sich das Gewicht der Daten, denn die Daten haben diese Konzentration bewirkt. Dazu betrachten wir die jeweiligen Vorher- und Nachher-Wahrscheinlichkeiten:

Zur Konzentration der Nachher-Wahrscheinlichkeitsdichten

$P_1(x \in I) \approx 0{,}20$ $\quad P_1(x \in I|D) \approx 0{,}78$
$P_2(x \in I) \approx 0{,}24$ $\quad P_2(x \in I|D) \approx 0{,}79$
$P_3(x \in I) = 0{,}12$ $\quad P_3(x \in I|D) \approx 0{,}78$

Die Werte zeigen wiederum, dass die genauen Vorher-Wahrscheinlichkeiten nicht so bedeutsam sind, wenn wir erst genügend Daten gesammelt haben. Wir sehen eine Konzentration der Dichtefunktionen auf dasselbe Intervall, wobei das Ausmaß der Konzentration zugleich ein gutes Maß für das Gewicht der Daten abgibt. Diese Repräsentation des Gewichts der Daten ist allerdings auf die bayesianische Darstellung mit Dichtefunktionen angewiesen. Der klassische Bayesianer kann das nachbilden, indem er mit Wahrscheinlichkeitsintervallen (bzw. mit Mengen von Glaubensgradfunktionen wie bei Hawthorne) statt Punktwahrscheinlichkeiten arbeitet. Bei nur wenigen Daten oder uneinheitlichen Daten können größere Intervalle das schwächere Gewicht dieser Daten zum Ausdruck bringen – ähnlich wie im Falle der Dichten die größere Streuung. Dieser Ansatz wird von Bayesianern immer wieder einmal erwogen und ein Stück weit ausgeführt, doch er stellt eher eine Minderheitenposition dar. Für den einfachen Bayesianismus bleibt daher die Darstellung des Gewichts der Daten weiterhin ein Problem.

5.8.8 Achinsteins Einwände

Peter Achinstein (2001, Kap. 4) entwickelt eine Reihe von Argumenten hauptsächlich in Form von Gegenbeispielen gegen eine zu einfache Gleichsetzung von relationaler oder absoluter Bestätigung und ihren Explikationen mit Hilfe von Wahrscheinlichkeiten. Das betrifft nicht nur die subjektiven Bayesianer, sondern eigentlich alle Probabilisten. Dieser Punkt wird von Probabilisten gern vernachlässigt. Man unterstellt geradezu, dass diese Zusammenhänge zwischen epistemischer Rechtfertigung und Wahrscheinlichkeit offensichtlich sind und keiner längeren Debatte bedürfen. Haben wir demnach erst einmal Glaubensgrade erhalten, ist dann eigentlich klar, dass sie erkenntnistheoretisch relevant sind, und wir damit einfach bessere, weil feinere Unterscheidungen zur Verfügung haben als der klassische Erkenntnistheoretiker. Gegen diese Nachlässigkeit sind Achinsteins Beispiele eine gute Medizin.

Kommen wir zunächst zur *inkrementellen Bestätigung*. Die ersten beiden Beispiele sollen zeigen, dass eine Wahrscheinlichkeitserhöhung allein nicht hinreichend für eine Bestätigung einer Theorie ist. Achinstein spricht oft von Evidenz und denkt dabei daran, dass uns e einen guten Grund dafür liefern sollte, an H zu glauben, damit wir sagen können, dass e h bestätigt. Mit b wird unser Hintergrundwissen beschrieben.

Lotteriebeispiel 1

b = Am Montag werden 1000 Lotterielose verkauft und John kauft davon 100 Lose und Bill kauft 1 Los.
Und es gewinnt bei der Lotterie am Mittwoch genau ein Los.
e = Am Dienstag werden alle Lose bis auf diejenigen von Bill und John zerstört.
h = Bill gewinnt

In solchen Beispielen dürfen wir unser Hintergrundwissen natürlich nicht ausblenden. In dieser Situation wird die Wahrscheinlichkeit dafür, dass Bill gewinnt, zwar von P(h|b)=1/1000 durch die Information e auf p(h|e&b)=1/101 angehoben, aber nach Achinstein können wir trotzdem nicht sagen, dass h durch e bestätigt wird. Dem stehen seiner Meinung nach mehrere Gründe entgegen:

1. Ist die Gewinnwahrscheinlichkeit von 1/101 immer noch sehr gering. Wir sagen schließlich auch nicht, wenn jemand einen normalen Fahrstuhl besteigt, das sei ein Grund für die Annahme, dass er jetzt bald sterben würde, obwohl eine leichte Wahrscheinlichkeitserhöhung dafür gegeben ist. Doch die ist zu gering, als dass wir sie bereits einen *Grund* für seinen Tod nennen würden. So ist das auch im Falle von Bill. Wir verfügen bisher über keine guten Gründe dafür, dass Bill gewinnt. Das Argument finde ich recht überzeugend. Wir bezeichnen winzige Veränderungen der Wahrscheinlichkeit nicht unbedingt schon als Gründe dafür, etwas anzunehmen. Aber vielleicht könnte der Probabilist entgegnen, dass man diese Fälle als Grenzfälle einer gewöhnlichen Praxis betrachten darf, die echte Wahrscheinlichkeitserhöhungen als hinreichend für Bestätigung ansieht.

2. Nach Achinstein würde e zwei konkurrierende Hypothesen *zugleich* bestätigen. Johns Wahrscheinlichkeit für einen Gewinn (h*) steigt nämlich von 1/10 auf 100/101. Das scheint mir für die relationale Bestätigung

aber keineswegs so überraschend zu sein, wie das Kartenbeispiel im letzten Abschnitt zeigen sollte. Allerdings kommt hier ins Spiel, dass h* vor allem gestützt und gut begründet wird und die Steigerung bei h nur ein minimaler Nebeneffekt ist. Aber es scheint auch nicht zwingend erforderlich zu sein, dass damit eine gewisse Bestätigung von h ausgeschlossen würde.

Das Beispiel ist also noch nicht sehr überzeugend, zeigt aber schon gewisse Problemstellen der inkrementellen Bestätigung auf. Achinsteins zweites Beispiel betrifft einen Profi-Schwimmer, der ins Wasser geht. Ist das ein Grund für die Annahme, dass er ertrinken wird? Achinstein setzt wieder darauf, dass winzige Wahrscheinlichkeitserhöhungen keinen echten Grund liefern. Da ist zwar etwas dran, aber es wird den Probabilisten nicht gleich umhauen.

In weiteren Beispielen versucht er zu zeigen, dass die Wahrscheinlichkeitserhöhung auch nicht erforderlich ist für eine Bestätigung. Das gelingt ihm m.E. schon etwas besser.

Lotteriebeispiel 2

e_1 = Die New York Times (NYT) berichtet, dass Bill Clinton 999 von 1000 Lotterielosen besitzt.
e_2 = Die Washington Post (WP) berichtet, dass Bill Clinton 999 von 1000 Lotterielosen besitzt.
b = Die Lotterie ist fair und es wird ein Los wahllos gezogen, das gewinnt. NYT und WP berichten beide zu 100% zuverlässig.
h = Bill Clinton gewinnt die Lotterie.

Die Situation ist etwas unrealistisch, was auch von Roush (2004) bemängelt wurde, aber nicht unmöglich. Man sollte vielleicht ein etwas besseres Beispiel suchen. Jedenfalls gilt hier:

$$P(h|e_1\&b) = P(h|e_2\&e_1\&b) = 999/1000$$

Das heißt, dass e_2 die Wahrscheinlichkeit für h nicht weiter erhöht. Trotzdem stellt bei unserem Hintergrundwissen e_2 natürlich einen sehr guten Grund dar, an h zu glauben. Man könnte sagen, unsere Gründe seien überdeterminiert. Gleichwohl entwertet e_1 damit nicht die Qualität

von e_2. Sollten wir e_1 bereits kennen, benötigen wir e_2 nicht mehr, und es trägt nichts Neues zu unserem Wissen mehr bei. Außerdem bleibt die Tatsache bestehen, dass aus e_2 und unserem Hintergrundwissen h logisch folgt. Natürlich ist also e_2 eine Bestätigung von h.

Man kann sagen, dass es sich dabei um einen Verwandten des Problems der alten Evidenz handelt. Wenn wir schon wissen, dass Bill Clintons Gewinnwahrscheinlichkeit 999/1000 ist, so werden weitere Belege dafür vom inkrementellen Ansatz einfach nicht mehr gezählt. Der Ansatz gibt uns eben keine Auskunft darüber, wie stark e_2 für h spricht, sondern nur darüber, was e_2 bei unserem bisherigen Wissen über h noch dazu beisteuern kann. Die Bestätigungswirkung von e_2 geht dabei leider völlig verloren. In unserem Beispiel würde e_1 als bestätigend betrachtet, aber e_2 nicht mehr. So sollte das Hintergrundwissen nicht in die Bestätigungsbeziehung eingehen. Der Bayesianer beschreibt hier nicht die objektiven Beziehungen zwischen Daten und Hypothesen, sondern nur welchen zufälligen Beitrag bestimmte Daten jeweils noch für unseren Kenntnisstand haben.

Geben wir die unrealistische Annahme, dass beide Zeitungen völlig irrtumssicher berichten, zugunsten der Annahme auf, dass sie sehr zuverlässig berichten (Quote 99%), so könnte das Beispiel immer noch eine unerwünschte Asymmetrie nach sich ziehen: e_1 würde unsere Wahrscheinlichkeit für h stark erhöhen und e_2 dann nur noch wenig.

Achinstein hat aber sogar noch ein Beispiel mit einer Wahrscheinlichkeitsverringerung, das trotzdem einen Fall von guten Gründen darstellt.

Medikamentenbeispiel

e_1 = Am Montag nimmt David am Vormittag um 10:00 Uhr ein Medikament M ein, welches Davids Symptom S beseitigen soll.

e_2 = Am Montag um 11:00 Uhr am Vormittag nimmt David Medikament M', um S zu beseitigen.

b = Medikament M hat 95% Effektivität Symptom S innerhalb von 2 Stunden zu beseitigen; Medikament M' ist zu 90% effektiv S innerhalb von 2 Stunden zu beseitigen. M' hat die Nebenwirkung, dass M' die Wirkung von M vollständig blockiert, ohne in der eigenen Wirkung beeinträchtigt zu sein.

h = Davids Symptom S ist um 13:00 Uhr weg.

Beide Medikamente M und M' sind also recht effektiv und M' hebt die Wirkung von M auf. M ist zwar effektiver (95%) als M', aber M' (90%) hat vielleicht weniger Nebenwirkungen. Daher nimmt David dann ebenfalls noch M', als er es erhält. Zu dem Zeitpunkt gehören zu unserem Hintergrundwissen bereits b und e_1. Die Wahrscheinlichkeit für eine Genesung war also 95%. Durch e_2 sinkt die Wahrscheinlichkeit für eine Genesung auf 90%. Dennoch bildet e_2 natürlich einen sehr guten Grund für die Annahme, dass David genesen wird, denn es hat ja selbst auch eine Heilungsquote von 90%. Hier senkt e_2 also die Wahrscheinlichkeit für h und stellt trotzdem eine Bestätigung für h dar. Das ist für einen Probabilisten sicherlich ein größeres Problem, denn hier bedeutet eine h-bestätigende Information e_2 eine Wahrscheinlichkeitsabsenkung. Er könnte sich allerdings auf das Likelihoodmaß l beziehen. Die Betrachtung von $P(e_2|h)/P(e_2|\neg h)$ kann uns womöglich noch darüber Auskunft geben, dass e_2 Relevanz für die Bestätigung von h besitzt.

Das Beispiel stellt natürlich insbesondere ein Problem für eine probabilistische Konzeption der Kausalität dar, nach der eine Ursache normalerweise das Auftreten der Wirkung erhöhen sollte. Im Beispiel ist die Gabe von M' zwar offensichtlich eine Ursache der Heilung, verringert aber deren Wahrscheinlichkeit. Diese Problematik überträgt sich nun auf unsere epistemische Situation.

Achinstein wendet sich dann der *absoluten Bestätigung* zu und versucht sie allerdings so zu verstehen, dass auch hier ein einzelnes Datum e die Bestätigung für h erbringen soll: $P(h|e\&b) > ½$. Doch Achinstein verweist zu Recht darauf, dass in dieser Form die Bestätigungsfunktion nicht selektiv ist und wir daher keine wirklich *relationale Bestätigung* (meine Redeweise) erhalten. Sollte also bereits unser Hintergrundwissen zu hohen Wahrscheinlichkeiten für h führen, können wir anhand von (*) $P(h|e\&b) > ½$ nicht sagen, dass e nun h bestätigen würde. Gehört etwa zu unserem Hintergrundwissen, dass Michael Jordan ein männlicher Basketballspieler ist, so mag für die Information (e) »Michael Jordan hatte Cornflakes zum Frühstück« und für die Hypothese (h) »Michael Jordan wird nicht schwanger«, zwar die Gleichung (*) gelten, aber e ist trotzdem irrelevant für h. Das zeigt allerdings vor allem, dass man mit der absoluten Bestätigungsbeziehung keine relationalen Bestätigungen auszeichnen kann, sondern dass es eher wie oben schon beschrieben

eine Beziehung zwischen einer Hypothese und unserem gesamten Hintergrundwissen ist. Es wird dabei keine Aussage e speziell hervorgehoben aus unserem Hintergrundwissen.

Tatsächlich gibt es also unterschiedliche Verwendungsweisen unserer Konzepte von Bestätigung und Begründung und die Frage bleibt weiter zu diskutieren, inwiefern diese Verwendungsweisen durch bestimmte Wahrscheinlichkeitsungleichungen rekonstruiert werden können.

Peter Achinstein (2001) plädiert selbst dafür, objektive epistemische Wahrscheinlichkeiten zu verwenden – ähnlich wie die objektiven Bayesianer. Allerdings bleibt eher unklar, welcher Art diese Wahrscheinlichkeit genau sein soll. Darauf aufbauend gehören für Achinstein zu einer Bestätigung von H durch E zwei Dinge: Erstens muss im Sinne der Schwellenwertkonzeption die Wahrscheinlichkeit P(H|E) > 0,5 sein und zweitens muss es als eine *Relevanzbeziehung* zwischen E und H eine Erklärungsbeziehung geben, die allerdings verschiedene Formen annehmen kann, wonach etwa H nun E erklärt oder H und E beide durch eine weitere Theorie T erklärt werden.

Achinsteins hybride Bestätigungskonzeption: E bestätigt H gdw.
(1) P(H|E) > 0,5
(2) Es besteht eine direkte oder indirekte Erklärungsbeziehung zwischen H und E.

Die Grundideen dieser Konzeption sind sicherlich richtig, dass wir zum einen versuchen müssen, mit objektiven Wahrscheinlichkeiten zu arbeiten, dass wir eine Art von Schwellenwertkonzeption benötigen und dass wir für eine Relevanzbeziehung andererseits noch auf eine weitere Verbindung zwischen den Daten und den Hypothesen angewiesen sind, die in einer Erklärungsbeziehung zu suchen ist. Das unterstützt die entsprechende Forderung des Erklärungskohärenzansatzes. Leider werden die oben genannten offenen Fragen etwa nach der Art der objektiven Wahrscheinlichkeit nicht zufriedenstellend beantwortet. Hier kommen wir um eine Bezugnahme auf den Bayesianismus also nicht herum.

5.9 Weitere spezielle Probleme und Anwendungen des Bayesianismus

Es sollen im Folgenden einige weitere Anwendungen des Bayesianismus diskutiert werden, die uns dabei helfen, den Bayesianismus besser kennenzulernen und zu verstehen. Dabei wird es u.a. um schwierige Probleme der Bestätigungsproblematik gehen.

5.9.1 Bayesianische Entscheidungstheorie

Bayesianer verweisen gerne darauf, dass ihre Konzeption über die bayesianische Entscheidungstheorie auf vielfältige Art mit unserer Handlungspraxis verknüpft ist. Das gilt natürlich ebenfalls für die klassische Erkenntnistheorie und wir haben das bereits in Kapitel (5.3) angesprochen. Tatsächlich sind die Beziehungen aber für den bayesianischen Ansatz besonders eng und gut ausgearbeitet. Leider können wir in diesem Rahmen diese Verbindung nicht ausführlich analysieren, sondern nur kurz die Anbindung schildern. Zunächst sind wir gezwungen, eine (kardinale) Nutzenfunktion einzuführen, die über einen rein erkenntnistheoretischen Rahmen hinausgeht. Die Werte, um die es dabei jeweils geht, können einen sehr praktischen Charakter haben, der über unsere bisherige erkenntnistheoretische Fragestellung weit hinausführt. Dann wird der *erwartete Nutzen* für die verschiedenen Handlungsmöglichkeiten bestimmt. Dazu benötigen wir zusätzlich die subjektiven Wahrscheinlichkeiten, die angeben, mit welcher Stärke wir erwarten, dass bestimmte Umweltsituationen S_i eintreffen, und die Handlungsanweisung der bayesianischen Entscheidungstheorie besagt dann, dass es rational ist, die Handlung auszuwählen, für die der erwartete Nutzen maximal ist.

Diese Anbindung des Bayesianismus an eine Theorie rationalen Entscheidens ist sicher ein Pluspunkt für den Bayesianismus. Allerdings gibt es auch zahllose praktische und theoretische Probleme, die damit verknüpft sind und die Entscheidungstheorie ist sicher nicht auf den Einsatz subjektiver Wahrscheinlichkeiten festgelegt. Natürlich würden wir dort objektive bevorzugen, nur dass wir die oft nicht zur Verfügung haben. In den kritischeren Darstellungen zur Entscheidungstheorie

werden diese Fragen umfangreich diskutiert, worauf ich hier nicht entsprechend ausführlich eingehen kann. Doch an drei Problemen möchte ich ein wenig beleuchten, welcher Art die Probleme dabei sein können. Das ist das Briefumschlagsparadox, das Ziegenproblem und das St. Petersburg Paradox. Der Leser findet dazu eine Fülle von Material z.B. im Internet.

Das *Briefumschlagsparadox* ist schnell erzählt. Sie befinden sich in einer Gameshow und dürfen einen von zwei Briefumschlägen auswählen und dann aufmachen. In beiden Umschlägen ist Geld enthalten, aber in dem einen doppelt so viel, wie in dem anderen. Sie wissen eben nur nicht, in welchem sich die höhere Summe befindet. Sie wählen zunächst einen aus und schauen hinein. Es sind 100 Euro enthalten. Nun dürfen Sie aber noch wechseln und könnten statt des ersten Umschlags auch den anderen wählen. Wie sollten Sie sich rationalerweise verhalten?

Es scheint klar zu sein, dass Ihre neuen Informationen über die Summe in dem Umschlag hier nicht weiter führen. Es ist völlig gleichgültig, welchen der Umschläge Sie wählen, denn Sie haben keine Ahnung, in welchem die größere und in welchem die kleinere Summe enthalten ist. Sind Sie für ein wichtiges Projekt auf zusätzliche 200 Euro zwingend angewiesen, sollten Sie natürlich den anderen Umschlag wählen, um die Chance auf die 200 Euro beizubehalten. Aber wir gehen davon aus, dass es solche speziellen Wünsche nicht gibt, und es Ihnen schlicht darum geht, möglichst viel Geld mitzunehmen.

Nun kommt die bayesianische Entscheidungstheorie ins Spiel, um uns doch noch zu helfen. Sie legt uns folgende Überlegung nahe: Es gibt zwei gleichberechtigte Möglichkeiten. Erstens könnten im anderen Brief 50 Euro sein und zweitens 200 Euro. Was würde uns dann bei einem Wechsel statt der erhaltenen 100 Euro erwarten? Lassen Sie uns den sogenannten Erwartungswert bestimmen:

$$\text{Erwartungswert(anderer Umschlag)} = ½\cdot 50 + ½\cdot 200 = 125$$

Also wäre danach ein durchschnittlicher Gewinn von 25 Euro durch einen Wechsel zu erwarten. Wir sollten demnach wechseln. Und das Ergebnis wäre für jeden positiven Eurobetrag entsprechend ausgefallen.

Doch das erscheint uns ganz unplausibel angesichts der völligen Symmetrie der beiden Umschläge. Irgendetwas ist schiefgegangen. Was

das ist, dazu gibt es unterschiedliche Vorschläge, die ich nicht diskutieren werde. Dagegen möchte ich für folgende Einsicht argumentieren: Nicht immer bietet eine solche entscheidungstheoretische Analyse Vorteile gegenüber der simplen intuitiven Betrachtung. Sie kann uns sogar auf komische Abwege führen. Damit wir eine vernünftige Entscheidung treffen können, müssen wir zunächst ein sinnvolles probabilistisches Modell aufstellen, und das ist uns in diesem Fall wohl nicht gelungen, obwohl unsere Modellierung hier keineswegs subjektiv erscheint. Vielmehr haben wir uns auf ein Indifferenzprinzip für die Vorher-Wahrscheinlichkeiten gestützt und damit sogar einem objektiven Bayesianismus Genüge getan.

Ähnliche Probleme, zunächst ein sinnvolles Wahrscheinlichkeitsmodell zu finden, kennen wir aus dem berühmten *Ziegenproblem* (und seinen Varianten), das ich bereits im ersten Kapitel dargestellt habe und nun wenigstens noch einmal kurz erwähnen möchte: Sie sind Kandidat in einer Gameshow und sehen sich drei Türen gegenüber. Hinter einer ist ein Ferrari, hinter den anderen beiden eine Ziege. Sie möchten den Ferrari. Nun dürfen Sie sich zunächst völlig frei für eine Tür entscheiden. Nehmen wir etwa einmal Tür 1. Der Moderator weiß, wo der Ferrari steht und muss Ihnen nun einen Tipp geben, indem er von den verbleibenden zwei Türen (2 und 3) eine öffnet, hinter der jedenfalls eine Ziege steht. Nehmen wir an, es sei Tür 2. Dann haben Sie noch einmal die Möglichkeit zu wechseln und statt Tür 1 nun auf Tür 3 umzuschwenken. Sollten Sie das tun? Selbst gestandene Statistiker waren sich darin unsicher und schrieben z.T. wütende Briefe gegen die erste Auflösung des Problems. Ja, Sie sollten wechseln. Die Gewinnwahrscheinlichkeit von Tür 1 ist 1/3, während die von Tür 3 doppelt so hoch ist und 2/3 beträgt. Der Moderator verweist mit seinem Hinweis, dass hinter Tür 2 eine Ziege steht, indirekt darauf, dass dann ein Ferrari, der nicht hinter Tür 1 zu finden ist, in jedem Fall nur noch hinter Tür 3 stehen kann. Es gibt unterschiedliche Wege, sich das zu verdeutlichen und man kann es sogar experimentell überprüfen, indem man das Spiel mehrfach wiederholt und die Gewinnraten bestimmt.

Ein kurzer Hinweis dazu: Wenn wir das Spiel etwa 99-mal wiederholen und uns dabei strikt an die Strategie halten, immer an unserer ersten Wahl festzuhalten, wie viele Ferraris werden wir wohl mitnehmen? Dass wir sogleich richtig liegen, hat eine Wahrscheinlichkeit von 1/3, also

werden es ca. 33 Ferraris sein. Das bedeutet aber, dass jemand mit der Gegenstrategie (oder ein zweiter Spieler), der also immer die verbliebene andere Tür wählt, die restlichen 66 Ferraris gewinnen würde, da der Moderator schließlich immer nur Ziegentüren öffnet und immer ein Ferrari im Spiel ist. Wenn Ihnen das immer noch nicht einleuchtet, ist das für meinen Punkt umso besser. Es zeigt, dass die größten Probleme oft in der Wahl eines geeigneten Modells liegen und nicht unbedingt durch den Rechenapparat gelöst werden. Außerdem zeigen unsere intuitiven Probleme mit der Lösung, dass wir in unseren tatsächlichen Überlegungen keineswegs Bayesianer zu sein scheinen.

Allerdings kann auch gerade der Mix von Wahrscheinlichkeit und Nutzenwerten zu speziellen Problemen führen. Das zeigt sich im *St. Petersburg Paradox*. Dabei wird jemand gefragt, welchen Einsatz er für das folgende Spiel zahlen würde, um daran teilnehmen zu dürfen: Es wird jeweils eine faire Münze geworfen. Im ersten Schritt erhält man 2 Euro, wenn Kopf fällt, und das Spiel ist dann zu Ende. Fällt dagegen Zahl, geht es weiter zum zweiten Schritt. Im zweiten Schritt erhält man 4 Euro, wenn Kopf fällt und es stoppt dann. Fällt dagegen Zahl, gehen wir zum dritten Schritt, bei dem der Einsatz wieder verdoppelt wird (also auf 8 Euro) usf. Hält man längere Zeit durch (fällt also häufiger Zahl) kann man höhere Gewinne erzielen. Doch wie wahrscheinlich sind solche Gewinne? Normale Menschen sind kaum jemand bereit, mehr als 20 Euro für die Teilnahme an diesem Spiel zu bezahlen. Ist das rational oder sind wir etwa alle dumm? Der Bayesianer scheint uns das nahezulegen, denn er rechnet wie folgt:

$$\text{Erwartungswert(Spiel)} = 1/2{\cdot}2 + 1/4{\cdot}4 + 1/8{\cdot}8 + \ldots = 1 + 1 + 1 + \ldots = \infty$$

Danach dürfen wir unendlich hohe Gewinne in diesem Spiel erwarten und sollten also auf jeden Fall unser ganzes Vermögen dafür hergeben, daran teilnehmen zu dürfen. Doch trotz dieser Rechnung wirkt das kaum überzeugend. Die hohen Gewinne treten nur bei langen Serien von Zahl auf. Warum sollte ich die so stark berücksichtigen? Der Trick ist natürlich, dass die Auszahlungen gerade so stark ansteigen, dass sie das Abfallen der Wahrscheinlichkeit genau ausgleichen. Auch hierfür wurden viele Lösungen vorgeschlagen, die aber auch nicht ganz unproblematisch sind.

Ein erster naheliegender Schritt der Bernoulli-Brüder war es, eine spezielle Nutzenfunktion einzuführen. Wir dürfen nicht einfach annehmen, dass zwei Milliarden Euro tatsächlich für uns doppelt so viel Wert besitzen wie eine Milliarde Euro. Der Nutzengewinn nimmt mit höheren Beträgen langsam ab. Die Ökonomen sprechen vom *abnehmenden Grenznutzen*. Wir benötigen daher eine Nutzenfunktion U, die einem Geldbetrag einen Nutzen zuordnet. Ein Vorschlag ist etwa, den Logarithmus zur Basis 2 zu wählen. Dann ergibt sich für unseren Erwartungswert:

$$\text{Erwartungswert(Spiel)} = \sum_n 1/2^n \cdot \log_2(2^n) = \sum_n n/2^n$$

Das ist zumindest ein endlicher Wert und wir können neue Hoffnung schöpfen, dass wir durch derartige Nutzenfunktionen auch solche Spiele wieder besser repräsentieren können. Allerdings bleibt das Problem bestehen, dass jede Nutzenfunktion U eine feste Funktion der Geldbeträge für eine bestimmte Person darstellt, die für sie ebenso in anderen Spielen gilt. Normalerweise wird darüber hinaus angenommen, dass jeder Zuwachs an Geld zumindest noch einen bestimmten wenn auch kleinen Nutzenzuwachs bietet und die Nutzenfunktion daher monoton wächst. Dann lässt sich wieder ein ähnliches Spiel mit etwas höheren Auszahlungen finden, das uns genauso wenig unendlich wertvoll erscheint, für das wir aber wieder unendliche Erwartungsnutzenwerte erhalten. Zu einer monoton wachsenden Nutzenfunktion U müssen wir nur die Umkehrfunktion U^{-1} heranziehen und die Auszahlungen auf $U^{-1}(2^n)$ erhöhen. Dann erhalten wir:

$$\text{Erwartungswert(Spiel)} = \sum_n 1/2^n \cdot U(U^{-1}(2^n)) = \sum_n 2^n/2^n = \infty$$

In unserem Beispiel müssten wir auf jeder Stufe nicht die Auszahlung 2^n vornehmen, sondern $2^{(2\hat{}n)}$. Trotzdem würde niemand (mit einem gewissen Vermögen) sein ganzes Geld dafür hergeben, an diesem Spiel teilzunehmen.

Es bleibt dann nur noch der (ad-hoc?) Ausweg, Zuwächse an Geld ab einem bestimmten Niveau überhaupt nicht mehr als Nutzenzuwachs zu betrachten (vgl. Büchter & Henn 2007). Dadurch bleibt natürlich der

Erwartungswert im endlichen Bereich. Ob sich dadurch sinnvolle Nutzenwerte ergeben, ist allerdings eine andere Frage, die wir der weiteren Debatte in der Entscheidungstheorie überlassen müssen. Jedenfalls zeigt Martin Peterson (2009, 84f.), dass sich dagegen wiederum neue Varianten des Paradoxes entwickeln lassen.

Das Beispiel sollte nur noch einmal belegen, dass die bayesianischen Kalkulationen im Rahmen der Entscheidungstheorie keineswegs nur erfolgreiche und intuitive Anwendungen des Bayesianismus darstellen müssen. Hier lauern viele Fallstricke, für deren Lösung wir an unterschiedlichen Stellen unserer Modellierung eingreifen und Anpassungen vornehmen müssen (hier zumindest an der Nutzenfunktion), um ein brauchbares Resultat zu erhalten. Trotzdem ist klar, dass wir unsere epistemischen Resultate letztlich u.a. in den Dienst praktischer Entscheidungen stellen müssen und dazu hat der Bayesianismus zumindest ein konkretes Angebot vorgestellt. Das können wir als Pluspunkt für ihn betrachten, selbst wenn es eine ganze Reihe von Wermutstropfen gibt. Ob er dabei aber tatsächlich besser abschneidet als die klassische Erkenntnistheorie, möchte ich dagegen offen lassen.

5.9.2 Das Duhem-Quine-Problem

Das Duhem-Quine-Problem kennen wir bereits aus dem hypothetisch-deduktiven Ansatz. Wenn Quine mit seinem nicht weiter qualifizierten erkenntnistheoretischen Holismus Recht hätte, dann könnten wir nur das *ganze Netz unserer Überzeugungen* mit unseren Beobachtungen vergleichen und hätten bei Konflikten zwischen unseren Überzeugungen und unseren Beobachtungen die Wahl, irgendwo im Netz (sogar bei den logischen Regeln) eine Änderung vorzunehmen, um diesen Konflikt zu beseitigen. Wir würden dabei aus praktischen Gründen vor allem in der Peripherie nach Änderungsmöglichkeiten suchen, da das vermutlich den kleinsten Umbau unseres Überzeugungssystems mit sich bringen würde, aber erkenntnistheoretische Gründe sind nicht so leicht für diese Vorgehensweise zu finden. Somit können wir nicht aus objektiven Gründen auf spezielle Hypothesen als Übeltäter zeigen und diese dann als falsifiziert betrachten. Ebenso wenig können wir bestimmten Hypothesen den Verdienst zuschreiben, wenn ein beobachtetes Datum

(probabilistisch) aus unseren Hypothesen abgeleitet werden kann. Das Verdienst kommt immer nur dem ganzen Überzeugungssystem zu und kann eben nicht auf die einzelnen Aussagen daraus verteilt werden.

Doch die Praxis der Wissenschaften sieht ganz anders aus. Mir sind jedenfalls keine Fälle aus der Wissenschaft bekannt, in denen man zur Lösung eines Konflikts zwischen Theorien und Daten die Regeln der Logik verändert hätte. Natürlich kann man durch Aufgabe des Modus Tollens und das Akzeptieren von Inkonsistenzen solche Probleme formal aus der Welt schaffen, aber die Frage ist, ob das auch ein methodologisch sinnvolles Vorgehen für die Wissenschaft wäre. Es gibt zwar schon Fälle, in denen wir scheinbar inkonsistente Theorien akzeptieren. Lakatos (1974) hatte ein solches Beispiel im bohrschen Atommodell gesehen, doch hier liegen eigentlich andere Interpretationen nahe, nach denen wir schlicht den Anwendungsbereich der klassischen Elektrodynamik einschränken und gebundene Elektronen davon ausnehmen (vgl. Bartelborth 1989).

Quine (1979) erwähnt noch den Einsatz einer dreiwertigen Logik in der Quantenmechanik. Dabei ging es jedoch nicht darum, die Theorien mit den Daten zu versöhnen, sondern vielmehr darum, ein verständliches Interpretationsmodell für die Quantenmechanik zu entwickeln. Doch selbst dazu konnte die dreiwertige Logik nicht sehr viel beisteuern, weil der dritte Wahrheitswert in diesem Anwendungsfall nicht wirklich verständlicher ist, als das zugrunde liegende Problem der Quantenmechnik, wonach bestimmte Größen nicht zugleich eindeutig bestimmte Werte aufweisen können. Also scheint Quine die Konsequenzen für die Wissenschaft deutlich zu übertreiben.

Es stimmt jedoch, dass wir normalerweise nicht allein aus einer entwickelten wissenschaftlichen Theorie T bereits bestimmte Daten ableiten können, sondern, dass wir weitere Hilfsannahmen A dazu benötigen. Im einfachen Fall sagt etwa T in einer bestimmten Anwendung einen Wert r für eine theoretische Größe g vorher, und wir akzeptieren die Hilfsannahme A, dass ein bestimmtes Messgerät oder Verfahren diese Größe korrekt misst. Oder A beinhaltet, dass bestimmte Anfangswerte oder entsprechende Annahmen gegeben sind. Nehmen wir also z.B. an, dass eine Aussage E* (etwa g = r in unserer Anwendung) oder einfach ¬E aus T und A deduktiv folgt, wir aber tatsächlich E beobachten, das

mit E* logisch unverträglich ist. Unsere Frage ist dann typischerweise: Wen trifft hier die Schuld, T oder A? Klar ist, dass T&A falsifiziert ist, aber welche der beiden Teilaussagen sollten wir in so einem Fall als falsifiziert betrachten? Dazu müssen wir nachsehen, inwiefern T und A unterschiedlich behandelt werden können. Doch wo können in der bayesianischen Betrachtung die Asymmetrien zu finden sein, die uns hier weiterhelfen?

Sie könnten z.B. in unterschiedlichen Likelihoods bestehen. In unserer Situation gilt jedenfalls: P(E|T&A) = 0. Nun betrachten wir das Updaten von T und A mit E jeweils für sich:

$$(1)\ P^{+}(T) = P(T|E) = P(T\&\neg A|E)+P(T\&A|E) = P(E|T\&\neg A)\cdot P(T\&\neg A)\ /\ P(E)$$

Das erste Problem dürfte sein, dass P(E|T&¬A) nicht objektiv bestimmt ist. Aus T&A folgt zwar E*, aber für T&¬A wissen wir nicht viel. Wenn unsere Theorie zwar stimmt, aber unser Messverfahren nicht mehr, wie groß soll dann unser Messwert für g sein? Das können wir wohl zunächst nur als subjektive Wahrscheinlichkeit konzipieren. Die Likelihoodanbindung versagt hier, da die entsprechende objektive Likelihood im Normalfall nicht bestimmt ist.

Das ist auch ein Problem der klassischen Statistik, die ganz auf objektive Likelihoods vertraut. Um etwa einen Signifikanztest anwenden zu können (der auch eine Form von statistischer Falsifikation darstellt), müssen daher die Hypothesen typischerweise ganz einfach sein, so dass die Messung der betreffenden Größen keine zusätzlichen Risiken mit sich bringt und wir einfach davon ausgehen dürfen, dass die erforderlichen Hilfsannahmen wie A in jedem Falle wahr sind. Dann trifft die Falsifikation eindeutig die betrachteten Hypothesen und die Likelihoods lassen sich eher bestimmen.

In unserem Beispiel können wir (ähnlich wie das Howson & Urbach 1993 vorgeben) einfach davon ausgehen, dass g endlich viele, unterschiedliche diskrete Werte annehmen kann und wir auf denen eine Gleichverteilung vornehmen dürfen. Dann wird P(E|T&¬A) relativ klein werden. Sollten wir darauf allerdings eine Lösung des Duhem-Quine-Problems aufbauen wollen, wäre die ganz auf unsere subjektive Vorgabe von Vorher-Wahrscheinlichkeiten angewiesen und damit eigentlich keine

Lösung, die durch den bayesianischen Apparat nahegelegt würde. Wir erhalten ganz symmetrisch die Gleichung für A:

(2) $P^+(A) = P(E|\neg T\&A)\cdot P(\neg T\&A) \;/\; P(E)$

Für die Likelihood $P(E|\neg T\&A)$ gilt aber dasselbe, was wir oben für $P(E|T\&\neg A)$ gesagt haben und somit kämen wir vermutlich am ehesten zu denselben subjektiven Werten. Dann können wir die Likelihoods später rauskürzen. Wir kommen nun ein Stück weiter, indem wir wieder den Quotienten der beiden Wahrscheinlichkeiten betrachten und zusätzlich noch annehmen, dass T und A statistisch unabhängig sind, so dass zumindest $P(T\&A) = P(T)\cdot P(A)$ ist. Diese Annahme ist oft nicht ganz unplausibel, denn die Wahrheit der Theorie und die korrekte Arbeit des Messinstruments hängen nur dann zusammen, wenn das Messinstrument sich bereits auf die Theorie stützen muss. Das kann allerdings auch passieren, aber wird zunächst einmal ausgeklammert, um überhaupt zu weiteren Einsichten für diesen Fall zu gelangen. Damit ergibt sich:

(3) $P^+(T) \;/\; P^+(A) = [P(E|T\&\neg A)\cdot P(\neg A)\cdot P(T)] \;/\; [P(E|\neg T\&A)\cdot P(\neg T)\cdot P(A)]$

Darin können wir den Faktor

(4) $Q(T/A) = [P(E|T\&\neg A)\cdot P(\neg A)] \;/\; [P(E|\neg T\&A)\cdot P(\neg T)]$

als Quotientenupdatefaktor betrachten, der darüber entscheidet, wie sich das Verhältnis der Wahrscheinlichkeiten von T und A durch das Updaten ändert. Dazu können wir nun zwei Fälle betrachten. Erstens nehmen wir an, dass die beiden Likelihoods, die wir nun schätzen müssen, schlicht gleich sind, da wir eigentlich nicht viel über sie wissen (s.o.), also:

$P(E|T\&\neg A) = P(E|\neg T\&A)$

Dann entscheiden nur noch die Vorher-Wahrscheinlichkeiten von T und A über den Quotientenupdatefaktor.

(5) $Q(T/A) = [1–P(A)] \;/\; [1–P(T)]$

Das bedeutet, dass bei gleichen Wahrscheinlichkeiten für T und A, also P(T) = P(A), der Faktor 1 ist. Dann sind also auch die Nachher-Wahrscheinlichkeiten von A und T gleich. Oder wir müssen wieder auf die Gleichung (4) zurückgehen und doch noch Gründe für einen Unterschied in den Likelihoods finden. Die könnten dann den Fall entscheiden.

Ist aber P(T) > P(A), so wird der Updatefaktor größer als 1 und damit verändert sich zumindest das Verhältnis der Wahrscheinlichkeiten von A und T weiter zugunsten von T.

Beispiel dazu: P(T) = 0,9 und P(A) = 0,7, dann ist $P(T)/P(A) \approx 1{,}3$ und Q(T/A) = 3, d.h., dann ist $P^+(T)/P^+(A) \approx 3{,}9$

Das Verhältnis der Wahrscheinlichkeiten für T relativ zu A ändert sich also von 1,3 auf 3,9. Das heißt, dass die weniger plausible Aussage hier weiter abgewertet wird gegenüber der plausibleren Aussage. Da in unserem Beispiel die Theorie T ursprünglich plausibler ist, trifft die Falsifikation von T&A vor allem A. Als Konsequenz könnte man sagen, dass A durch E falsifiziert wird und T nicht so betroffen ist.

Mit Hilfe einiger Zusatzannahmen ist es uns gelungen, einen relativ allgemeinen Zusammenhang zu finden, der die bayesianische Variante der *Regel des schwächsten Gliedes* darstellt. Diejenige Aussage, die unwahrscheinlicher ist, sollte aufgegeben werden, wenn eine Konjunktion falsifiziert wurde.

Das wird von Bayesianern als großer Erfolg des bayesianischen Programms gesehen (vgl. Earman 1992, Kap. 3.7). Doch genau genommen beruht der ganze Erfolg vor allem auf der Einschätzung der Startwahrscheinlichkeiten und für die liefert der subjektive Bayesianismus keine Hilfestellung. Auch die Quantifizierung der jeweiligen Abwertungen hängt wesentlich an bloß subjektiven Größen (da uns die Likelihoodanbindung hier normalerweise nicht weiterhilft) und stellt daher keine besondere Leistung des Bayesianismus dar. Er sagt uns das, was auch als sinnvolle Regel für den hypothetisch-deduktiven Ansatz gilt: Wenn T&A zu einem Widerspruch mit den Daten führt, dann gib die Aussage von beiden auf, die uns schon vorher als unwahrscheinlicher erschien bzw., die schon vorher schlechter begründet war.

Dazu gibt es aber auch ganz andere Antworten, die etwa auf die innere Struktur von Theorien und ihre Erklärungskraft Bezug nehmen. Innerhalb der sogenannten strukturalistischen Theorienauffassung (vgl. Bartelborth 1996) werden entwickelte Theorien als komplexe Theoriennetze dargestellt, die verschiedene Theorieelemente T_i mit unterschiedlichem Spezialisierungsgrad und verschiedenen Komponenten aufweisen. Kommt es nun zu Konflikten mit der Erfahrung, wird normalerweise zunächst versucht, Änderungen in den entfernten Spezialisierungen und dort in den weniger zentralen Komponenten vorzunehmen und damit die Konflikte aufzulösen. Erst wenn das nicht mehr gelingt, geht man zu Abänderungen an zentraleren Bestandteilen über (vgl. Gähde & Stegmüller 1988 für eine entsprechende Fallstudie). Das hat u.a. etwas mit der Erklärungsstärke der abgeänderten Theorie zu tun. Die Erklärungsstärke nimmt am wenigsten ab, wenn wir nur in der Peripherie der Theorie Abschwächungen vornehmen und die meisten Erklärungen der Theorie weiter aufrechterhalten werden können, denn diese Vereinheitlichungsleistung ist ein wichtiger Aspekt der Erklärungsstärke.

Derartige Überlegungen können tatsächlich Einblick in unsere Methodologie der Theorienänderung geben und könnten dann auch – wenn sie erfolgreich sind – eine gehaltvolle Antwort auf das Duhem-Quine-Problem bieten, doch die bayesianische Rekonstruktion der Regel vom schwächsten Glied kann keine derartigen Einsichten vermitteln, so dass ich hier keine spezielle bayesianische Auflösung des Problems erkennen kann. Der Bayesianismus geht somit nicht wesentlich über das hypothetisch-deduktive Schließen kombiniert mit dieser Regel hinaus, zumal er auf weitere Zusatzannahmen über die Likelihoods und statistische Unabhängigkeit angewiesen ist.

Wir können den Effekt auch an einem explizit und vollständig bestimmten probabilistischen Überzeugungssystem vorführen. Daran können wir auch relativ einfach den Effekt bestimmter Wahlen subjektiver Likelihoods nachzeichnen. Gehen wir also davon aus, dass wir es mit den drei Aussagen T, A und E zu tun haben und dass T und A unabhängig sind: $P(T\&A) = P(T)\cdot P(A)$. Dann haben wir es mit einem kleinen bayesschen Netz zu tun: $A \Rightarrow E \Leftarrow T$, für das wir nur bestimmte Wahrscheinlichkeiten bestimmen müssen, um damit alle Wahrscheinlichkeiten unseres Überzeugungssystems festzulegen.

Nehmen wir dazu etwa die folgenden und stellen die Negationen wieder durch kleine Buchstaben dar:

P(T) = 0,9	P(t) = 0,1
P(A) = 0,6	P(a) = 0,4
P(E|TA) = 0	P(e|TA) = 1
P(E|Ta) = x = 0,1	P(e|Ta) = 1–x = 0,9
P(E|tA) = y = 0,2	P(e|tA) = 1–y = 0,8
P(E|ta) = z = 0,3	P(e|ta) = 1–z = 0,7

Wir müssen also insgesamt 6 Wahrscheinlichkeiten festlegen, während sich alle anderen daraus ergeben. Dabei bin ich davon ausgegangen, dass die Wahrscheinlichkeit E klein bleibt, wenn etwa nur unsere Hilfstheorie A falsch sein sollte und erst dann größer wird, wenn auch T falsch ist, aber hier könnte man natürlich ebenso gut x, y und z variieren und nachsehen, was dann jeweils passiert. Entsprechend der Formel für bayessche Netze können wir dann leicht die weiteren gemeinsamen Verteilungen angeben, so gilt z.B. P(EtA) = P(t)·P(A)·P(E|tA) = 0,0012. Entsprechend erhalten wir:

P(ETA) = 0	P(eTA) = 0,54
P(ETa) = 0,036	P(eTa) = 0,324
P(EtA) = 0,012	P(etA) = 0,048
P(Eta) =0,012	P(eta) = 0,028

Jetzt können wir mit E updaten (wodurch T&A falsifiziert wird) und sehen, was sich für die Annahmen T und A jeweils einzeln dabei ergibt:

P(E) = 0,036 + 0,012 + 0,012 = 0,06
P(ET) = 0,036
P(EA) = 0,012
$P^+(T) = P(ET)/P(E) = 0{,}036/0{,}06 = 0{,}6$
$P^+(A) = P(EA)/P(E) = 0{,}012/0{,}06 = 0{,}2$

Man sieht sehr deutlich, dass beide Aussagen A und T abgewertet wurden, aber dass die Abwertung von A viel deutlicher ist als die von T. Wir können nun schnell auch noch den Fall betrachten, in dem wir mit

¬E updaten und damit unsere Theorie T und die Hilfsannahmen A ein Stück weit bestätigen. Man könnte sagen, dass es sich dabei um das umgekehrte Duhem-Quine-Problem handelt. Wir möchten nun nicht wissen, wer der Hauptschuldige für die Falsifikation ist, sondern vielmehr, wer am meisten von einer Bestätigung profitiert:

P(e) = 0,54 + 0,324 + 0,048 + 0,028 = 0,94
P(eT) = 0,54 + 0,324 = 0,864
P(eA) = 0,54 + 0,048 = 0,588
P*(T) = P(eT)/P(e) = 0,864/0,94 = 0,92
P*(A) = P(eA)/P(e) = 0,588/0,94 = 0,626

Auch hier kann der Bayesianer genauer beschreiben, wie sich die Bestätigungswirkung durch ¬E auf die beiden Annahmen verteilt. Auffällig ist hier etwa, dass selbst T noch sehr deutlich von der Bestätigung profitiert, obwohl T bereits eine hohe Wahrscheinlichkeit aufwies.

Deutlicher lässt sich das noch für andere Ausgangswerte erkennen. Wählen wir etwa für unsere Theorie nun noch geringere Ausgangswerte und schauen, was dann passiert:

P(T) = 0,6 P(t) = 0,4
P(A) = 0,3 P(a) = 0,7
P(E|TA) = 0 P(e|TA) = 1
P(E|Ta) = x = 0,1 P(e|Ta) = 1–x = 0,9
P(E|tA) = y = 0,2 P(e|tA) = 1–y = 0,8
P(E|ta) = z = 0,3 P(e|ta) = 1–z = 0,7

Nun müssen wir die grundlegende Verteilung neu bestimmen:

P(ETA) = 0 P(eTA) = 0,18
P(ETa) = 0,042 P(eTa) = 0,378
P(EtA) = 0,024 P(etA) = 0,096
P(Eta) =0,084 P(eta) = 0,196

Dann können wir wieder mit ¬E updaten und erhalten diesmal:

P(e) = 0,85
P(eT) = 0,18+0,378 = 0,558
P(eA) = 0,378 + 0,096 = 0,474
P*(T) = P(eT)/P(e) = 0,558/0,85 = 0,656
P*(A) = P(eA)/P(e) = 0,474/0,85 = 0,558

Hier wird deutlich, dass vor allem die schwächere Hypothese durch die bestätigenden Daten bestätigt wird. Das zeigt ein allgemeines Phänomen, das aber natürlich auch von den Werten von x, y und z abhängt. Der Leser möge nun selbst mit entsprechenden Werten weiter experimentieren.

5.9.3 Ad-hoc-Hypothesen

Durch eine Abänderung von Hilfsannahmen können wir unsere Theorien praktisch immer vor einer Falsifikation retten. Wenn also T&A zu einem Widerspruch mit der Erfahrung führt, suchen wir nach einer Hypothese A*, so dass T&A* wieder mit unseren Beobachtungen verträglich ist. Eine Hypothese vom Typ A*, die nur dazu entwickelt wurde, um eine Theorie zu retten, die aber selbst keine oder nur geringe Stützung durch andere Aussagen erfährt, nennen wir eine *Ad-hoc-Hypothese*. Da sie verhindert, dass wir unsere Theorie T falsifizieren können, hat insbesondere Popper Ad-hoc-Hypothesen als epistemisch minderwertig betrachtet und wollte ihren Einsatz ganz verbieten.

In der Wissenschaft war man aber nicht daran interessiert, bei Problemen mit den Daten sogleich seine besten Theorien aufzugeben und suchte daher immer wieder nach passenden Ad-hoc-Hypothesen. Als man im Rahmen der Phlogistontheorie, wonach eine Verbrennung das Entweichen des Feuerstoffs Phlogiston darstellt, die bei einer Verbrennung übrig bleibenden Stoffe gewogen hat, und diese schwerer waren als vor der Verbrennung, hat man kurzerhand vermutet, dass Phlogiston negatives Gewicht haben müsste. Oder als der Planet Merkur sich nicht so verhielt, wie es die newtonsche Gravitationstheorie vorhersagte, nahm man einfach an, es müssen noch einen sonnennäheren noch nicht entdeckten Planeten Vulkan geben, der für die Abweichungen verantwortlich sei. Oder als Michelson und Morley den Ätherwind beim Durchgang der Erde durch den Äther mit Hilfe eines Interferometers messen wollten, aber dabei immer nur Nullergebnisse bekamen, hat man

nicht gleich die Äthertheorie aufgegeben, sondern Lorenz hat behauptet, dass sich das Messgerät beim Durchfliegen des Äthers genau so weit verkürzt, dass dadurch der Effekt nicht mehr messbar wird. Sind das dann keine Fälle von sauberer Wissenschaft mehr?

Der Bayesianer könnte sagen, dass Ad-hoc-Hypothesen solche sind, die eine geringe Vorher-Wahrscheinlichkeit aufweisen, da sie ja nicht durch andere Erkenntnisse gestützt werden. Das spräche ebenfalls dafür, sie eher zu verbieten. Doch schon der Kohärenztheoretiker sieht das anders. Sie sind dadurch gerechtfertigt, dass sie zur Gesamtkohärenz eine Menge beisteuern können, weil sie schließlich eine im Übrigen erklärungsstarke Theorie retten helfen. Die kann aber eindeutig kohärenzstiftend sein. Ob die Ad-hoc-Hypothesen sich schließlich als Fortschritt erweisen werden oder doch nur als Sackgasse kann man nicht von vornherein wissen. Hinterher sind wir natürlich immer schlauer und können vielleicht über bestimmte Irrwege der Phlogistontheoretiker lachen, doch viele, die zu ihrer Zeit verlacht wurden, hatten fortschrittliche Ideen zu bieten. Die Bahnstörungen des Uranus ließen ebenfalls im 19. Jh. vermuten, dass es noch einen weiteren Planeten (den Neptun) geben müsste, und der wurde dann tatsächlich an der vermuteten Stelle entdeckt. Die Annahme des Uranus war also erfolgreich, die des Vulkans bekanntlich nicht. Beide Annahmen erfolgten durch den Astronomen Leverrier.

Keiner kann also im Voraus sagen, welche Ad-hoc-Hypothesen sich als progressiv und welche sich als Sackgasse entpuppen werden. Ad-hoc-Hypothesen, die darüber hinaus praktisch empirisch kaum testbar sind, stellen sicherlich ein Problem dar, aber das muss keineswegs mit dem Ad-hoc-Charakter der Hypothesen zu tun haben. Wir müssten also zwischen guten und schlechten Ad-hoc-Hypothesen unterscheiden können, wenn wir Popper überhaupt folgen möchten. Da bisher keineswegs klar ist, wie das geschehen sollte, sollten wir lieber davon Abstand nehmen. Dann hat aber auch der Bayesianismus nicht viel dazu zu sagen. Der Kohärenztheoretiker wird hingegen eine klare Antwort zu Ad-hoc-Hypothesen geben: Akzeptiere die Ad-hoc-Hypothesen, solange sie weniger Inkohärenzen aufweisen, als sie vermeiden helfen.

Beim negativen Gewicht des Phlogistons kann man schon sagen, dass wir eine geradezu analytische Inkonsistenz heraufbeschwören, wenn wir von einem Gewicht annehmen, dass es negativ sein könnte.

Dem mussten also schon gute Gründe gegenüberstehen, um das noch wenigstens zu einem bestimmten Zeitpunkt zu akzeptieren. Ob das historisch der Fall war, scheint mir eher zweifelhaft, aber das können wir hier nicht verfolgen. Jedenfalls ist die Wissenschaft voll von solchen Beispielen von Ad-hoc-Hypothesen und der Bayesianismus gibt uns nicht unbedingt hilfreiche Einblicke in die erkenntnistheoretischen Probleme dieser Annahmen. Das scheint vor allem daran zu liegen, dass hier bestimmte holistische Effekte im Spiel sind. Die Ad-hoc-Hypothesen sind nicht für sich genommen epistemisch wertvoll und werden insbesondere nicht direkt durch die Daten gestützt, aber sie sind epistemisch wertvoll über den Umweg der Theorie, die sie retten. Das kann eine Kohärenzkonzeption der Rechtfertigung besser erfassen als die anderen Ansätze zum induktiven Schließen.

5.9.4 Die Rabenparadoxie

Wir haben bereits als Problem für das hypothetisch-deduktive Schließen die Rabenparadoxie kennengelernt. Da der Bayesianismus deduktive Beziehungen reproduziert, ist er ebenso mit diesem Paradox belastet. Allerdings schrieben sich Bayesianer auf die Fahnen, dass sie eine Lösung anzubieten haben. Zur Erinnerung: Das Paradox besteht darin, dass die Hypothese $H \equiv \forall x(Rx \rightarrow Sx)$ (Alle Raben sind schwarz) intuitiv durch eine Instanz der Art $A \equiv Ra\&Sa$ bestätigt wird und dann die zu H äquivalente Kontraposition $H_k \equiv \forall x(\neg Sx \rightarrow \neg Rx)$ dementsprechend durch die »Gegeninstanz« $E \equiv \neg Sa\&\neg Ra$ bestätigt würde, was natürlich auch zugleich H mitbestätigen sollte. Doch wie kann ein Gegenstand, der weder schwarz noch ein Rabe ist, unsere ursprüngliche Hypothese H bestätigen? Das kommt uns recht unplausibel vor. Wie sieht das im Bayesianismus aus? Die Lösungsidee der Bayesianer ist simpel. Sie akzeptieren, dass A unsere Hypothese H normalerweise bestätigt und ebenso, dass die Gegeninstanz H bestätigt, aber sie versuchen zu zeigen, dass die Instanzen die Hypothese zumindest deutlich besser bestätigen als die Gegeninstanzen. Gegeninstanzen sollten bei geeigneten Hintergrundannahmen die Hypothese überhaupt nur minimal bestätigen.

Um diese Ergebnisse zu erhalten, werden normalerweise bestimmte Annahmen akzeptiert (vgl. Vranas 2004, Fitelson & Hawthorne 2010 mit ausführlichen Literaturhinweisen zur Rabenparadoxie).

Annahmen zum Rabenparadox
(1) P(¬Sa) > P(Ra)
(2) P(Ra|H) = P(Ra)
(3) P(¬Sa|H) = P(¬Sa)

Dabei steht S für schwarz und R für Rabe und die Gleichungen sind jeweils immer auf ein bestimmtes normales Hintergrundwissen K zu relativieren, das ich aber hier und im Folgenden der Übersichtlichkeit halber nicht explizit anführe. Das Beispiel von Good (s. Kap. 3.9) hat schon bewiesen, dass wir die Bestätigung durch eine Instanz nicht für jedes Hintergrundwissen erwarten dürfen, weshalb wir hier von einem gewöhnlichen Hintergrundwissen ausgehen.

Dann besagt die Annahme (1), dass es viel mehr nicht-schwarze Dinge als Raben gibt. Das ist in der Regel eine unbestrittene Annahme, die der entscheidende Hintergrund dafür ist, dass die Gegeninstanzen weniger bestätigend sind als die Instanzen, weil es davon eben viel mehr gibt als von den Instanzen. Wir müssen daher viel mehr Dinge untersuchen, um sicherzustellen, dass die nicht-schwarzen Dinge nicht-Raben sind, als wir sie untersuchen müssen, wenn wir direkt die Raben anschauen. Kontroverser sind da schon die Annahmen (2) und (3), nach denen die Beobachtungen Ra und ¬Sa jeweils statistisch unabhängig von H sind. Vranas (2004) bemängelt, dass es dafür keine Rechtfertigungsversuche gibt. Man geht schlicht von diesen Annahmen aus, um das gewünschte Resultat zu erzielen. In Fitelson & Hawthorne (2010, Kap. 7) leiten die Autoren eine der Lösungen ohne ganz so starke Annahmen ab.

Zunächst einmal untersuchen sie unterschiedliche Bestätigungsmaße c und zeigen, dass das erwünschte Ergebnis äquivalent zu einer einfachen Ungleichung in den Wahrscheinlichkeiten ist. Es gilt:

(4') c(H,A) > c(H,E) ist für die Maße 1-3 äquivalent zu
(4) P(H|A) > P(H|E)

Nur das zweite Differenzmaß fällt hier aus dem Rahmen. Da es ebenfalls aus anderen Gründen nicht sehr plausibel wirkt, ist das aber kein

Beinbruch. Es genügt demnach, (4) zu zeigen. Dazu nehmen die Autoren bestimmte »Nichttrivialitätsannahmen« (NT) an:

$0 < P(H|Ra\&Sa) < 1$ und $0 < P(H|\neg Ra\&\neg Sa) < 1$ und
$P(\neg Sa\&Ra) > 0$ (immer relativ zu einem Hintergrundwissen K)

Dann leiten sie u.a. (durch das Betrachten bestimmter Likelihoodquotienten) mit schwächeren Annahmen als (1) bis (3) ein entsprechendes Theorem ab, wonach (gleichgültig, ob die Instanzen überhaupt H stützen) A unsere Hypothese H mehr stützt als E.

Außerdem erhalten wir entsprechende Abschätzungen für die Größenordnung des Unterschieds in der Bestätigung durch die Instanzen und die Gegeninstanzen. Dazu wird allerdings doch wieder eine approximative Variante von Annahme (3) benötigt. Außerdem erweisen sich die Unterschiede nicht gerade als beeindruckend.

Damit ist jedenfalls klar, wie die Bayesianer die Paradoxie zu lösen gedenken. Aber ist das schon eine überzeugende Lösung? Die Annahmen bei Hawthorne und Fitelson (2010) scheinen tatsächlich relativ akzeptabel, doch die Frage bleibt, ob wir Gegeninstanzen überhaupt als bestätigende Instanzen für die Hypothese H zu betrachten möchten. Einen Ausweg dazu bot schon der Ansatz, nach dem unsere Konditionale eher nomische Konditionale sind und sie daher keine Äquivalenz zu ihren Kontrapositionen aufweisen.

Auch Mark Siebel (2004) verneint das mit guten Gründen. Denken wir an unseren Biologen aus Kapitel 3.9. Er hofft auf ein Forschungsprojekt, in dem er die Hypothese H bestätigen möchte, indem er lauter Gegeninstanzen sammelt, wobei er die schwächere Bestätigung durch eine größere Zahl von Objekten wettmachen möchte. Selbst wenn wir es wissenschaftlich für wertvoll hielten, H zu begründen, wird uns das kaum überzeugen können. Der Biologe sammelt also weiße Schuhe, rote Sportwagen, grüne Blätter, braune Äste, gelbe Bücher, etc. Er könnte auf diese Weise sicher eine bedeutende Sammlung von Objekten aufstellen, die alle nicht-schwarz und Nicht-Raben sind. Sagt uns das schon irgendetwas Interessantes über die Farbe von Raben? Angesichts der Mengen von Objekten, die er hier sammeln könnte und angesichts dessen, dass er sich nie den Raben zuwendet, scheint uns das Projekt ziemlich sinnlos zu sein.

Wir sollten hier nicht davon sprechen, dass die Gegeninstanzen unsere Hypothese überhaupt stützen – noch nicht einmal schwach. Damit haben wir womöglich wieder einen ähnlichen Punkt erreicht wie in den Kritiken von Achinstein (s. Kap. 5.8.8), in dem wir auch festgestellt haben, dass nicht jede Wahrscheinlichkeitserhöhung schon als Bestätigung zu deuten ist. Man muss also wohl resümieren, dass die Überlegungen der Bayesianer das Paradox zwar etwas abschwächen können, aber doch nicht wirklich lösen.

5.9.5 Irrelevante Konjunktionen

Aus dem hypothetisch-deduktiven Ansatz haben wir außerdem noch das Problem der irrelevanten Konjunktionen bzw. das Tacking-Paradox geerbt. Plagt es tatsächlich ebenso den Bayesianismus? Das scheint zumindest für den Fall zu gelten, in dem ein Datum E deduktiv aus einer Hypothese H und unserem Hintergrundwissen K folgt. Dann scheint das Problem der hypothetisch-deduktiven Bestätigung B_{dh} sich zu wiederholen, dass wir eine irrelevante Hypothese X zu unserer Theorie H hinzufügen können, so dass sie gleich mitbestätigt wird:

(IK) $B_{dh}(H|E;K) \Rightarrow B_{dh}(H\&X|E;K)$

Das ist zumindest sehr unschön, da X überhaupt keine Verbindung zu H und E haben muss und trotzdem mitbestätigt wird. Außerdem ist es auch für andere Probleme wünschenswert, wenn wir genauer bestimmen könnten, welche Teile einer Theorie T durch E bestätigt werden und gegenüber welchen Teilen von T E neutral ist oder welche es sogar schwächt.

Die Beziehung (IK) finden wir ebenso im bayesianischen Fall wieder, gegeben, dass E noch nicht logisch aus unserem Hintergrundwissen K folgt (und tatsächlich $P(E|K)<1$ ist), aber aus H deduzierbar ist (vgl. Fitelson 2002 und Hawthorne & Fitelson 2004). Die erste Frage ist schon, was wir überhaupt von einer Lösung in diesem Fall erwarten. Selbstverständlich gilt:

(1) $P(H\&X|E\&K) \leq P(H|E\&K)$

Dabei gilt die Gleichheit normalerweise nur in den Fällen, in denen X keine substantiellen Behauptungen zu H&E&K hinzufügt. Aber das ist nur eine triviale Weisheit über Wahrscheinlichkeiten. Wir haben nicht den Eindruck, dass damit schon unser Problem gelöst sein dürfte. Stattdessen müssen wir den Fall wieder mit Hilfe der geeigneten Bestätigungsmaße untersuchen. Und tatsächlich beweist Fitelson (2002), dass unter den Annahmen, dass X durch E bestätigt wird und X bestätigungsirrelevant für H, E und H&E ist, gilt:

(2) c(H,E|K) > c(H&X,E|K)

im Falle, dass c das Differenzmaß d oder das Likelihood-Ratio-Maß l darstellt (vgl. Kap. 5.6.1). Also gilt auch für entsprechende Bestätigungsmaße, dass die Hypothese H allein stärker durch E bestätigt wird als die Kombination H&X. Dabei bedeutet Bestätigungsirrelevanz für Fitelson bereits, dass probabilistische Irrelevanz vorliegt. Eigentlich geht man damit aber nur mit Hilfe der Unabhängigkeitsannahme über (1) hinaus. Erst Hawthorne/Fitelson (2004) verstärken das Resultat etwas, in dem sie die ursprüngliche Irrelevanzannahme von X etwas anders deuten, und erhalten die Aussage:

Theorem (Hawthorne/Fitelson): Wenn E H bestätigt und
P(E|X&H&K) = P(E|H&K) und es gilt P(X|H&K) < 1, dann gilt:
(2*) c(H,E|K) > c(H&X,E|K) (für die Maße c=d und c=l)

Dabei wird also nur noch verlangt, dass X keinen Beitrag mehr zur Ableitung von E leistet, der nicht schon in H und K enthalten wäre. Klar, dass die Bedingung erfüllt ist, sollte E ableitbar sein aus H&K. Das entspricht besser der ursprünglichen Idee der Irrelevanz im deduktiven Fall und liefert ein recht allgemeines Ergebnis, das wir als Lösung für das Irrelevanzproblem ansehen können. Viel mehr können wir wohl nicht erwarten. Hinter der Einsicht steckt allerdings vor allem der simple Zusammenhang aus (1). Aber immerhin wird also durch die Quantifizierung der Bestätigung damit deutlich, dass die Kombination H&X schlechter abschneidet (weniger gut bestätigt wird) als H alleine, so dass die Mitbestätigung zumindest die Gesamtbestätigung abschwächt. Das ließ sich im Falle der hypothetisch-deduktiven Theorienbestätigung

natürlich noch nicht ableiten, da sie keine unterschiedlichen Grade der Bestätigung kennt.

5.9.6 Variation der Daten

Nehmen wir an, jemand möchte die Hypothese H bestätigen, dass Wasser bei 100°C kocht. Dazu betrachten wir zwei Vorgehensweisen:

V1. Jemand kocht 1000-mal immer wieder im selben Topf mit demselben Verfahren, an derselben Stelle das Wasser und erhält immer das gewünschte Ergebnis.

V2. Jemand kocht ebenfalls 1000-mal Wasser, variiert dabei aber alle möglicherweise einflussreichen Umstände, wie die Art des Gefäßes (insbesondere seine Geometrie), die Schnelligkeit des Erhitzens, den Ort des Experiments u.v.m. Auch er erhalte immer das gewünschte Ergebnis.

Dann werden wir die Bestätigung durch V2 im Normalfall für deutlich besser und damit stärker erachten als die durch V1. Es scheint also so zu sein, dass eine Variation der Umstände die Daten gewichtiger macht. Doch warum ist das so und wie können wir das im Bayesianismus verdeutlichen?

Wir erwarten beim Verfahren V2 viel eher, dass die Hypothese falsifiziert wird, sollte sie (zumindest in dieser Allgemeinheit) falsch sein. Daher bietet V2 einen viel *strengeren Test* für H, als es für V1 der Fall ist. Sollte die Theorie diesen strengeren Test bestehen, spricht das stärker für sie als ein bloß schwacher Test. Man könnte hier im Sinne von Deborah Mayo (1996) sagen, dass ein Test T umso strenger für eine Hypothese H ist, umso größer die Wahrscheinlichkeit dafür ist, dass der Test negativ für die Hypothese ausgeht, wenn sie falsch ist. Wir werden sehen, dass das auch die Grundidee der statistischen Signifikanztests ist. In unserem Beispiel sollte es klar sein, dass es bei V2 sehr unwahrscheinlich war, dass alle Daten unsere Hypothese bestätigen und es schon ein ziemlicher Zufall war, dass wir gerade die Variationen des Ortes nicht erwischt haben, die H aufgrund unterschiedlichen Luftdrucks widerlegt hätten.

Wie kann aber ein Bayesianer diese Idee umsetzen? Dazu betrachtet etwa Earman (1992, Kap. 3.5) die Korrelationen der Daten untereinander

und nimmt an, dass die größer sind, wenn es sich um gleichartige Daten handelt, während sie kleiner sind, wenn die Umstände der Datenerhebung stärker variieren. Er definiert die Variation gerade über die Korrelation der Daten untereinander. Haben wir etwa zwei Daten $E \equiv E_1\&E_2$, so betrachten wir $P(E|K) = P(E_2|E_1\&K)\cdot P(E_1|K)$, was mitentscheidet über den Update-Faktor beim Updaten mit E. Je stärker E1 und E2 korreliert sind, umso größer ist $P(E_2|E_1\&K)$ und damit wird der Update-Faktor $P(E|H\&K)/P(E|K)$ umso kleiner und die Bestätigung durch E dann kleiner. Das wird besonders deutlich, wenn wir davon ausgehen, dass E aus H&K logisch ableitbar ist. Dann gilt:

$$P(H|E\&K) = P(H|K) \;/\; P(E|K) = P(H|K) \;/\; [P(E_2|E_1\&K)\cdot P(E_1|K)]$$

Allgemeiner gilt für

$E \equiv E_1\&\ldots\&E_n$ mit $H\&K \Rightarrow E$:

$$P(H|E\&K) = P(H|K) \;/\; [P(E_n|E_{n-1}\&\ldots\&E_1\&K)\cdot\ldots\cdot P(E_2|E_1\&K)\cdot P(E_1|K)]$$

Dann definiert Earman (1992, 78f) die Variation der Daten durch die Rate, mit der die Werte $P(E_n|E_{n-1}\&\ldots\&E_1\&K)$ für wachendes n anwachsen. Je schneller sie anwachsen, umso schneller nimmt die Bestätigung unserer Hypothese durch neue Daten E_i ab.

Dieser Korrelationsansatz sieht zunächst plausibel aus, denn intuitiv ist es wahrscheinlicher, dass unser Wasser wieder bei 100° C kochen wird, wenn wir den Versuch genauso wie beim letzten Mal gestalten, statt die Umstände stärker zu variieren. Das Problem bei dieser Lösung ist aber, dass die Korrelationen zwischen den Daten nur durch subjektive Wahrscheinlichkeiten bestimmt werden, statt durch eine objektive Ähnlichkeitsbeziehung zwischen den Daten. Das heißt, die Ähnlichkeiten müssen durch das epistemische Subjekt geschätzt werden. Die Lösung beruht also mehr auf den Einschätzungsfähigkeiten des Subjekts statt auf dem bayesianischen Apparat. Der bietet allerdings immerhin die Möglichkeit, diese Zusammenhänge entsprechend zu repräsentieren, was im hypothetisch-deduktiven Ansatz so noch nicht gegeben war.

Wie reagieren übrigens Vertreter der Abduktion auf das Problem der Diversität der Daten? Fehlt diese Diversität, bieten sich eingeschränkte Hypothesen als möglicherweise beste Erklärungen der Daten an. So

könnten die Daten aus V1 etwa dadurch erklärt werden, dass unsere Hypothese H nur für Töpfe eines bestimmten Typs und das Erhitzen auf eine ganz bestimmte Weise gilt. Verfügen wir nur über die Daten aus V1, bietet diese alternative lokale Hypothese dieselbe Vereinheitlichungsleistung für die tatsächlich vorliegenden Daten wie unsere ursprüngliche Hypothese. Erst die variierten Daten aus V2 werden weniger gut erklärt, wenn wir eine Reihe solch lokaler Hypothesen formulieren, statt einer globalen und damit invarianteren Hypothese, die für alle Fälle postuliert, dass Wasser bei 100°C kocht. Jedenfalls scheint diese globale Hypothese die besseren Erklärungen zu liefern, solange wir noch nicht festgestellt haben, dass sie eigentlich falsch ist. Dann sind wir natürlich doch gezwungen auf lokalere Varianten umzusteigen oder nach ganz neuen Ansätzen zu suchen, die wieder eine umfassendere Vereinheitlichung etwa über Siedepunkte unterschiedlicher Stoffe bieten.

5.9.7 Ein bayesianischer Gottesbeweis

Richard Swinburne hat in (1987) versucht, einen ausführlicheren Gottesbeweis zu führen, der nicht mehr auf metaphysischen Prinzipien beruht wie etwa einem entsprechenden Kausalprinzip, nach dem jedes kontingente Ereignis eine Ursache haben muss (und dann bleibt als letzte Ursache eben nur Gott übrig), sondern der sich stattdessen auf Einfachheitsüberlegungen stützt, die zu bestimmten Wahrscheinlichkeitseinschätzungen führen. Winfried Löffler (2002 und 2006) hat Swinburnes Argument in komprimierter Form rekonstruiert und mit konkreten Zahlen versehen, die ein überraschendes Ergebnis brachten. In gewisser Weise zeigen sich an diesem Beispiel die Stärken und die Schwächen der bayesianischen Konzeption. Außerdem ermöglicht das Argument von Swinburne auch einen Vergleich mit dem abduktiven Schließen, zumal er selbst viel von Erklärungen und der Einfachheit spricht, die für den Schluss von Bedeutung seien.

Die informellen Schritte des Arguments sind die Folgenden: Zunächst akzeptieren wir ein erkenntnistheoretisches Prinzip, nach dem wir bestimmten Erfahrungen und Berichten darüber solange Vertrauen schenken, wie sie nicht aus konkreten Gründen als wahrscheinlich irreführend einzustufen sind. Ein ähnliches Prinzip konservativen Vorgehens

wird auch in Bartelborth (1996) diskutiert. Dann kommt das Prinzip zum Einsatz, indem es auf religiöse Erfahrungen und Berichte darüber angewandt wird, die die Existenz Gottes zur Folge haben, sollten wir ihnen vertrauen. Des weiteren wird dann geklärt, dass die Existenz Gottes nicht aus anderen Gründen unwahrscheinlich ist. Sie ist nicht widersprüchlich und selbst die Existenz des Übels in der Welt macht sie nicht unwahrscheinlich. Dafür argumentiert Swinburne ausführlicher, was wir hier aber nicht verfolgen können. Des weiteren gibt es 6 weitere Belege für die Existenz Gottes (Rekonstruktion nach Löffler):

(1) die Existenz eines komplexen physikalischen Universums
(2) die erkennbare Ordnung im Universum
(3) die Existenz bewusstseinsbegabter Wesen
(4) die Übereinstimmung zwischen menschlichen und tierischen Bedürfnissen einerseits und Umweltgegebenheiten andererseits
(5) das möglicherweise Vorkommen von Wundern
(6) die Feinabstimmung grundlegender Naturkonstanten (erst in der zweiten Auflage enthalten)

Diese insgesamt 7 Belege (6 + die religiösen Erfahrungen) für die Existenz Gottes versucht Swinburne dann in Form bayesianischer Überlegungen auszuwerten.

Bevor ich darauf eingehe, möchte ich schon in etwas ketzerischer Weise darauf hinweisen, dass der Bayesianismus eben typischerweise Dämonentheorien unterstützt. Nehmen wir an, wir hätten eine Dämonentheorie d, die besagt, ein Dämon existiert, der e möchte und e herbeiführen kann und es dann auch tut. Dabei ist e ein beliebiges Ereignis, das wir beobachten können. Es könnte sich z.B. um einen Unfall handeln, den wir erleiden. Nehmen wir an, der Unfall geschah gemäß unserem Hintergrundwissen nicht zwangsläufig: $P(e) < 1$. Außerdem ist gemäß unserer Annahme der Dämon sehr effektiv: $d \Rightarrow e$. Dann bestätigt der Unfall ein Stück weit unsere Dämonenhypothese d, weil der Update-Faktor $P(e|d)/P(e) = 1/P(e)$ größer als eins wird.

Man sieht hier wieder, dass es sich um eine einfache Rekonstruktion eines hypothetisch-deduktiven Bestätigungszusammenhangs handelt. Außerdem genügen im Prinzip komparative Wahrscheinlichkeitsüberlegungen, um diesen Begründungszusammenhang zu erkennen. Auf

solche komparativen Überlegungen stützt sich auch Swinburne. Außerdem scheint es auch nichts an unseren Prämissen zu kritisieren zu geben. Unsere Dämonenhypothese dürfen wir ja zunächst einmal so gestalten, wie es uns sinnvoll erscheint. Das Einzige, was wir so noch nicht erreicht haben, ist, dass die Dämonenhypothese auch eine Wahrscheinlichkeit >0,5 erzielen muss. Dazu wären wir auf weitere Überlegungen in dieser Richtung angewiesen. Weitere Unfälle und Abschätzungen der Startwahrscheinlichkeiten für Dämonenhypothesen könnten uns dabei behilflich sein. Doch das will ich nicht weiter treiben. Es bleibt zumindest die Tatsache bestehen, dass es im Bayesianismus sehr leicht ist, Hypothesen so zu konstruieren, dass sie durch beliebige Ereignisse bestätigt werden. Man könnte das die Anfälligkeit des Bayesianismus für Dämonenhypothesen nennen.

Lässt sich das so ohne Weiteres auf das abduktive Schließen übertragen? Das glaube ich nicht, aber es hängt davon ab, ob wir der Dämonenhypothese auch genuine Erklärungskraft für das Auftreten bestimmter Unfälle zubilligen und ob wir keine anderen Erklärungshypothesen (wie meine Unachtsamkeit bzw. mein Schlafdefizit etc.) finden können, die bessere Erklärungen dieser Unfälle darstellen würden. Das können wir gleich noch einmal in Swinburnes Beispiel erörtern.

Löffler (2002) vergibt nun die entsprechenden Wahrscheinlichkeiten für die Variante des Bayeschen Theorems, auf die sich Swinburne stützt (das Hintergrundwissen k habe ich aus Gründen der Vereinfachung wieder weggelassen), h ist die Hypothese, dass ein allmächtiger, allgütiger Gott existiert und e ist die Konjunktion unserer 7 Belege:

$$P(h|e) = [P(e|h)\cdot P(h)] \,/\, [(P(e|h)\cdot P(h))+P(e\&\neg h)] =$$
$$Y \,/\, [Y+P(e\&\neg h)], \text{ mit } Y = P(e|h)\cdot P(h)$$

Hierbei kommt dem Term $P(e\&\neg h)$ eine besondere Bedeutung zu und Swinburne argumentiert dafür, dass er sehr klein wird, denn es sei unwahrscheinlich, dass all die obengenannten Dinge auftreten, wenn es keinen entsprechenden Gott gäbe. Geht der Term $P(e\&\neg h)$ aber gegen Null, so wächst die Nachher-Wahrscheinlichkeit unserer Gotteshypothese schnell gegen 1. Das ist auch das Überraschende, das uns Löffler an vielen Zahlenbeispielen verdeutlich: Swinburne zeigt sogar

mehr als er selbst behauptet. Swinburne gibt sich damit zufrieden, dass die Gotteshypothese eine Wahrscheinlichkeit größer als 0,5 aufweist, aber wenn wir ihm folgen, hat er eigentlich schon einen richtigen Gottes*beweis* vorgetragen. Wenn wir etwa P(e) = P(e|h) = 0,01 wählen, aber das Auftreten von e&¬h noch für unwahrscheinlicher halten (etwa P(e&¬h) = 0,00000001), dann ist die Wahrscheinlichkeit für h bereits bei 99,9%.

Doch das wirkt selbst auf einen theistisch gesinnten Philosophen wie Löffler als eine zu leichte Art und Weise, um Gott zu beweisen. Löfflers Kritik wendet sich gegen die Wahrscheinlichkeitsabschätzungen, die Swinburne uns vorgibt, die u.a. Abschätzungen dafür sind, wie wahrscheinlich es ist, dass dieses Universum entstanden ist, statt eines anderen (mit und ohne Gott). Er ist auch der Überzeugung, dass die dabei resultierenden Wahrscheinlichkeiten keineswegs objektiv sind (wie Swinburne es gerne betrachtet), sondern bereits durch bestimmte theistische Einstellungen und unsere metaphysischen Annahmen (wieder etwa eines Kausalprinzips) gefärbt sind. Sie sind nicht wirklich frei von Voreinstellungen und allein anhand von Einfachheitsüberlegungen zu gewinnen. Löffler nimmt daher an, dass wir eben weiterhin in unseren Argumenten für die Existenz Gottes auf diese metaphysischen Annahmen angewiesen sind. Das wird durch die bayesianische Argumentation nur verschleiert.

Das ist auch mein Punkt zum Bayesianismus. Er erlaubt es, selbst Überlegungen zur Existenz Gottes in eine einfache Berechnung zu gießen, aber er bietet dafür nur ein Buchhaltungssystem an, dessen genaue Wahrscheinlichkeiten wir anhand anderer Überlegungen vorgeben müssen. Ein subjektiver Bayesianer hätte damit schnell einen Gottesbeweis im Stile von Swinburne beisammen. Allerdings gäbe es genauso schnell viele Indizien für Dämonen, wenn er unsere Dämonentheorie einmal erwägen würde. Man sieht dabei, dass die ganze epistemische Arbeit eigentlich in der Wahl der entsprechenden Ausgangswahrscheinlichkeiten steckt und dahinter letztlich metaphysische Annahmen stehen, die wir m.E. nur anhand sorgfältiger Kohärenzerwägungen vornehmen können. Gehen wir dagegen schnell von unserem Hintergrundwissen ohne eine genaue Begründung anhand weiterer Argumente über die inneren Zusammenhänge unserer Annahmen zu subjektiven Wahrscheinlichkeiten über,

lassen sich schnell die Vorurteile reproduzieren, die wir eben gerne pflegen möchten. Der Bayesianismus hindert uns nicht daran und liefert nicht viel Hilfestellung, wie wir in solchen Fällen vorgehen sollten. Die Gotteshypothese oder auch Dämonenhypothesen werden so schnell und damit zu leicht bestätigt.

Kann das abduktive Schließen hier tatsächlich helfen? Das glaube ich schon, aber wir müssen ein wenig tiefer in die Argumentationen eindringen. Swinburne spricht viel von Erklärungen durch *einfache* Hypothesen und seiner Meinung nach ist das Gotteskonzept eines solchen unendlichen (allmächtig, allgütig etc.) Gottes ein besonders einfaches Konzept. Doch selbst wenn das so ist, ist damit nicht gesagt, dass die entsprechenden Erklärungen eine große Vereinheitlichung bieten oder überhaupt wissenschaftliche Erklärungen darstellen.

Nehmen wir etwa den Fall der Wunder. Wir gehen von n Wundern $w_1,\ldots,w_n$ aus und untersuchen die entsprechenden Gotteserklärungen dazu. Die sehen kurz gesagt etwa so aus:

> Ein allmächtiger Gott existiert. Gott möchte w_1 in Situation s_1. Also geschieht w_1 in s_1.
> Ein allmächtiger Gott existiert. Gott möchte w_2 in Situation s_2. Also geschieht w_2 in s_2.
> ...
> Ein allmächtiger Gott existiert. Gott möchte w_n in Situation s_n. Also geschieht w_n in s_n.

Doch schon dabei fällt auf, dass wir auf eine ganze Reihe *wesentlicher Zusatzannahmen* angewiesen sind, deren epistemischer Status selbst recht fragwürdig ist. Woher wissen wir, was Gott sich für bestimmte Situationen jeweils wünscht? Selbst wenn wir den Gottesbegriff also für besonders einfach halten, werden die dadurch generierten Erklärungen das noch nicht unbedingt.

Die einfache Gotteshypothese selbst stellt außerdem *kein nomisches Muster* dar, das von sich aus schon eine Generalisierung liefert, was Gott sich jeweils in bestimmten Situationen wünscht. Eine solche Art von invarianter Generalisierung müsste zumindest vorliegen, damit die Gotteshypothese so ausgestaltet ist, dass sie auch tatsächlich Erklärungskraft

und Vereinheitlichungskraft besitzt. Das Problem wird deutlich, wenn wir nach der Vorhersageleistung des obigen Schemas fragen. Die ist nämlich praktisch gleich Null, das sollte auch der Theist zugeben, denn so einfach lässt sich Gott nicht in die Karten sehen und ebenso wenig können wir entsprechende Wunder durch ganz bestimmte Interventionen gezielt herbeiführen.

Dazu kommt, dass wir nach *alternativen Erklärungen* Ausschau halten müssen. Wir können uns nie sicher sein, dass es sich tatsächlich um Wunder handelt, die also ein Eingreifen Gottes darstellen, sondern müssen immer wieder fragen, ob es nicht doch andere natürliche Erklärungen dafür gibt. Dieser komparative Aspekt und wie er genau einzubringen ist, wird ebenfalls besonders deutlich im abduktiven Ansatz, wird aber in der bayesianischen Rekonstruktion zunächst nicht deutlich.

Ganz ähnlich wie für Wunder können wir genauso für die oben angeführten sechs Phänomene und ihre vielen Einzelinstanzen detailliert nachfragen, ob die Gotteshypothese hier tatsächlich eine genuine und vereinheitlichende Erklärung anbieten kann, die zumindest deutlich besser ist als alternative Ansätze. Zudem müssen wir für das abduktive Schließen holistisch vorgehen und uns immer wieder fragen, ob die sich so ergebenden Überzeugungssysteme insgesamt kohärenter sind als ähnliche Überzeugungssysteme ohne Gottesannahme. Was dabei herauskommt, können wir an dieser Stelle sicher nicht endgültig klären, dafür ist die Debatte zu komplex, aber der Schluss auf die beste Erklärung führt uns zumindest vor Augen, wie umfassend die dabei auftretenden Abwägungen zu sein haben. Im Bayesianismus können wir dagegen zu leicht zu einem »Gottesbeweis« gelangen. Die wirklichen Problemfelder bleiben dabei links liegen und werden in der Wahl entsprechender Ausgangswahrscheinlichkeiten versteckt.

5.9.8 Zeugenaussagen und unabhängige Daten

Die Beurteilung der Aussagen anderer Menschen (im englischen kurz unter dem Stichwort »testimony« zusammengefasst) ist ein kompliziertes Gebiet der Erkenntnistheorie. Dem werden wir hier nicht in umfassender Weise gerecht werden können, aber wir können doch einige Aspekte

betrachten. Es ist schon deshalb so wichtig, weil wir wohl das meiste, das wir über die Welt wissen, durch Berichte anderer Menschen erfahren haben. Woher wissen wir z.B. etwas über die Daten und Experimente, die unseren wissenschaftlichen Theorien zugrunde liegen? Woher wissen wir etwas über die Politik und alle möglichen anderen Ereignisse in der Welt wie Erdbeben etc.? In den wenigsten Fällen waren wir selbst dabei und haben uns davon direkt überzeugen können, dass etwa Angela Merkel zur Kanzlerin gewählt wurde. Wir verlassen uns stattdessen auf Zeitschriftenberichte, Bücher, Fernsehreportagen etc. Das kennen wir schon aus unserer Schulzeit. Aber andererseits wissen wir ebenfalls, dass diese Berichte keineswegs immer zuverlässig sind. Wir sollten sie nicht gleich für bare Münze nehmen, sondern uns eigentlich erst einmal fragen, ob wir ihnen vertrauen dürfen. In den meisten Fällen *neigen* wir allerdings dazu, solchen Berichten blind zu vertrauen, aber viele Experimente über die Verlässlichkeit von Zeugenaussagen zeigen, dass wir das in wichtigen Fällen lieber nicht tun sollten.

In den erkenntnistheoretischen Debatten geht es noch um die Frage, ob diese Erkenntnisform einen besonderen *eigenständigen* Charakter hat oder eher auf die normalen Formen reduzierbar ist (vgl. Adler 2006). Das soll hier nicht weiter diskutiert werden, d.h., wir nehmen schlicht den Reduktionismus als gegeben an. Es soll uns also nur darum gehen, wie der Bayesianismus und das abduktive Schließen auf Zeugenaussagen anwendbar sind. Nehmen wir etwa an, in einer Zeugenaussage E wird behauptet, ein Ereignis H (etwa, dass X bei einer Tat beobachtet wurde) wäre beobachtet worden. Dann wird das normale Updaten bei festem Hintergrundwissen K betrachtet, das wir allerdings in unseren Formeln weglassen:

$$P(H|E) = P(H)\cdot P(E|H)] \;/\; [P(E|H)\cdot P(H)+P(E|\neg H)\cdot P(\neg H)]$$

Dabei muss der Bayesianer wieder einige Vorher-Wahrscheinlichkeiten einschätzen. Nehmen wir z.B. an, dass P(H) = 0,4 ist. Bisher war X also nur mäßig verdächtig. Dann müssen wir weiterhin wissen, wie hoch die Wahrscheinlichkeit P(E|H) dafür ist, dass unser Zeuge H behauptet, wenn H wahr ist. Vielleicht ist es dafür einfacher, P(¬E|H) zu bestimmen. Im Prinzip sind es zwei Möglichkeiten (plus ihre Überlappung), die hier zu beachten sind. Die erste besteht darin, dass der Zeuge sich irrt, und die

zweite darin, dass der Zeuge lügt. Man kann vereinfachend annehmen, dass sich diese beiden Möglichkeiten insgesamt addieren. Setzen wir also beispielsweise an, dass beides zusammen in 10% der Fälle vorliegt: $P(\neg E|H) = 0{,}1$. Somit wird der Zeuge in ca. 10 von 100 Situationen, in denen das Ereignis tatsächlich vorliegt, dennoch behaupten, dass es nicht stattgefunden hat. Diese Wahrscheinlichkeit dürfte sehr stark von den genauen Einzelheiten des jeweiligen Falles abhängen. Wie waren die Lichtverhältnisse, wie aufmerksam war der Zeuge bzw. wie abgelenkt bei der Beobachtung, wie lang hat er zugeschaut u.v.m. Hatte der Zeuge Gründe für eine Lüge? Wenn ja, wie hoch ist dann die Wahrscheinlichkeit, dass er das auch tut? Mit allgemeinen Annahmen über die Zuverlässigkeit von Zeugen, oder eines bestimmten Zeugen, werden wir hier kaum weiterkommen, und die Einschätzung wird in der Praxis sicher subjektiv bleiben. Ob uns eine bayesianische Berechnung dabei tatsächlich helfen kann, diese Überlegungen zusammenzunehmen, wird weiter zu diskutieren sein.

Spannend ist auch die andere Wahrscheinlichkeit, die wir nicht vergessen dürfen, dass ein Zeuge H behauptet, obwohl es gar nicht stattgefunden hat: $P(E|\neg H)$. Nehmen wir dafür einmal 15% an, dann können wir unsere Berechnung durchführen. (Wiederum *addieren* sich hier in etwa die 2 Möglichkeiten des Irrtums und der Lüge zu den 15%, und wir könnten das entsprechend weiter aufschlüsseln.)

Beispiel: Vorher: $P(H) = 0{,}4$, dann ergibt sich:
$P(H|E) = (0{,}9 \cdot 0{,}4) \,/\, (0{,}9 \cdot 0{,}4 + 0{,}15 \cdot 0{,}6) = 0{,}8$

Das heißt, die Wahrscheinlichkeit dafür, dass H vorliegt, hat sich durch die Zeugenaussage verdoppelt, obwohl wir diverse Irrtumsmöglichkeiten des Zeugen berücksichtigt haben. Ob diese Einschätzungen einigermaßen realistisch sind, ließe sich empirisch weiter testen, aber es bleiben immer die Unsicherheiten des Einzelfalls bestehen, die wir kaum sicher abschätzen können.

Ein naheliegender Vorschlag aus der Praxis ist es, nicht nur eine Aussage bzw. einen Bericht herzunehmen, sondern lieber auf mehrere Berichte zu setzen, die möglichst unabhängig voneinander entstanden sind. Nehmen wir an, in der FAZ findet sich ein Artikel (E_1), der meldet,

dass die Kanzlerin ihren baldigen Rücktritt angekündigt hat (H). Das kommt uns trotz des Berichtes noch recht unwahrscheinlich vor und wir suchen daher nach weiteren Bestätigungen. Wittgenstein schildert einen Mann, der dann einfach ein weiteres Exemplar der FAZ kauft und anhand des entsprechenden Artikels das noch einmal überprüft. Das würden wir im Normalfall kaum als weitere Bestätigung betrachten. Allerdings gibt es selbst dafür Ausnahmen. Wenn etwa ein Spion glaubt, man habe ihm nur eine getürkte Ausgabe der FAZ untergeschoben, damit er falsche Nachrichten weitergibt und sich damit unglaubwürdig macht, kann es durchaus sinnvoll sein, noch ein weiteres Exemplar der FAZ zu kaufen. Doch bleiben wir beim Normalfall. Der Mann, der hier zu einer zweiten FAZ greift, käme uns dann ziemlich dumm vor. Wir sollten lieber eine andere Zeitung kaufen (etwa die Zeit). Finden wir dort dieselbe Nachricht H wieder, hat sie sich ein Stück weit bestätigt. Allerdings haben vielleicht beide Zeitungen nur eine dpa-Meldung abgedruckt. Dann haben wir immer noch keine wirklich unabhängige Bestätigung erlangt und stehen eigentlich genauso da wie Wittgensteins Mann mit den zwei Exemplaren der FAZ.

Damit der zweite Bericht (E_2) eine weitergehende Bestätigung des ersten sein kann, muss er auf einem unabhängigen kausalen Weg vor dem ersten Bericht entstanden sein. Es sollten also zwei Reporter vor Ort gewesen sein, auf deren Augenzeugenberichte nun die beiden Artikel zurückzuführen sind. Das sollten wir allerdings nicht so beschreiben, dass E_1 und E_2 statistisch unabhängig sind (wie das etwa Howson und Urbach in 2003 annehmen), sondern sie sollten statistisch unabhängig sein gegeben H und gegeben ¬H. Selbst diese statistische Unabhängigkeit kann nicht immer die geforderte kausale Unabhängigkeit genau darstellen (s. Kap. 7). Doch sie gibt uns noch das beste probabilistische Indiz dafür, dass die angestrebte Unabhängigkeit vorliegt. Um das abschätzen zu können, nehmen wir etwa die folgende Umrechnung des bayesschen Theorems:

(*) $P(H|E_1\&E_2) = P(H) \,/\, [P(H) + P(\neg H)\ q]$, wobei q der folgende Quotient ist: $q = P(E_1\&E_2|\neg H) \,/\, P(E_1\&E_2|H)$

Nun hängt alles an q und wir können uns überlegen, was passiert, wenn die Reporter ihre Meldungen z.B. von dpa bzw. einer gemeinsamen

ähnlichen Quelle beziehen oder die Zeugen sich absprechen. Betrachten wir dafür die folgenden beiden Fälle:

1. Der zweite Zeuge übernimmt einfach das Urteil des ersten Zeugen: $P(E_1\&E_2|H) = P(E_1|H)$ und $P(E_1\&E_2|\neg H) = P(E_1|\neg H)$.
Das ist praktisch genauso, als ob wir nur einen Zeugen hätten.
2. Beide Zeugen urteilen unabhängig voneinander, so dass gilt: $P(E_1\&E_2|\neg H) = P(E_1|\neg H)\cdot P(E_2|\neg H)$ und
$P(E_1\&E_2|H) = P(E_1|H)\cdot P(E_2|H)$

Dann wird schnell klar, was aus unserem Quotienten q in den beiden Fällen wird. Im ersten Fall ergibt sich $q_1 = P(E_1|\neg H)/P(E_1|H)$ und im zweiten Fall wird das noch mit Faktor f multipliziert: $f = P(E_2|\neg H)/P(E_2|H)$, also $q_2 = q_1\cdot f$. Wenn wir nun wieder (*) anschauen, wird deutlich, dass die Aussage des zweiten Zeugen genau dann eine bessere Bestätigung von H bietet, wenn q durch ihn kleiner wird, also seine Urteilskraft der folgenden Bedingung genügt:

(**) Bedingung für zweiten Zeugen: $P(E_2|\neg H) < P(E_2|H)$

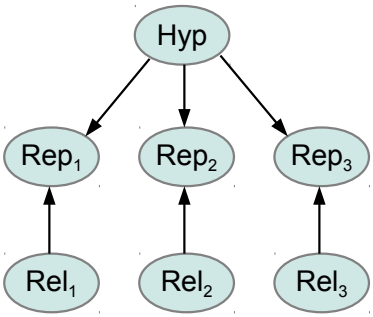
Der Bayesianismus liefert hier das gewünschte Ergebnis: Ein zweiter unabhängiger Zeuge bringt nur dann eine bessere Bestätigung für unsere Annahme H, wenn die Wahrscheinlichkeit dafür, dass er H aussagt, in jedem Fall kleiner ist, wenn H nicht der Fall ist, als wenn H der Fall ist. Gilt hier die Gleichheit ist er irrelevant und gilt sogar die umgekehrte Ungleichung ist seine Aussage eher schädlich.

Man könnte sagen, solange Bedingung (**) erfüllt ist, besitzt der 2. Zeuge zumindest eine gewisse positive *Zuverlässigkeit* in seinen Aussagen über H. Ist selbst diese Minimalbedingung nicht mehr gegeben, wäre es besser, er schließt sich einfach der Meinung des 1. Zeugen an. Ist der ebenfalls unzuverlässig, stehen wir zwar schlecht da, aber es wird nur noch schlechter, wenn wir dann noch den 2. Zeugen hören. Das sieht mir nach einem kleinen Erfolg für den bayesianischen Apparat aus, zumal sich die genannten Wahrscheinlichkeitszusammenhänge auf die meisten der Bestätigungsmaße übertragen (s.o.). Allerdings müssen wir weiterhin die subjektive Einschätzung vornehmen, dass die Bedingung

(**) gilt, und erraten, ob eher der 1. Fall oder der 2. Fall vorliegt, und dafür bietet uns der Bayesianismus keine besondere Hilfestellung.

Man kann nun noch versuchen, die Zuverlässigkeit der Zeugen anhand ihrer Übereinstimmung bayesianisch zu beurteilen (vgl. Bovens & Hartmann 2006). Stimmen viele Zeugen in ihren Urteilen überein, spricht das eher dafür, dass sie alle richtig liegen, als dass sie alle falsch liegen. Das bringen die Autoren auch noch mit probabilistischen Kohärenzüberlegungen in Verbindung (vgl. Kap. 4.4). Allerdings ist dabei Vorsicht geboten, denn eine zu große Übereinstimmung von Zeugenaussagen kann auch wieder Hinweise darauf geben, dass sie durch Absprachen oder andere gegenseitige Beeinflussungen entstanden ist, wodurch ihr epistemischer Wert wieder geschmälert würde. Eine ganz so einfache Regel für die Bewertung von Zeugenaussagen werden wir wohl nicht erhalten.

Bovens und Hartmann modellieren jedenfalls unterschiedliche Situationen anhand bayesscher Netze, in die unsere Unabhängigkeitsannahmen eingespeist werden. Dabei kann die jeweilige Zuverlässigkeit (Rel) der Berichte (Rep), dass ein bestimmter Sachverhalt vorliegt (Hyp), für jeden Zeugen als z.B. unabhängiger Faktor (Rel) angenommen werden, so dass wir z.B. für drei Zeugen das folgende Bild erhalten und natürlich entsprechende Graphen für n Zeugen:



Bayessches Netz 4: Drei Zeugenaussagen

In diesem Netz können wir entsprechende Berechnungen anderer Wahrscheinlichkeiten vornehmen, wenn uns bestimmte Dinge bekannt sind, wie etwa, dass alle Zeugen übereinstimmende Berichte abgeliefert haben. Die Zuverlässigkeit kann gewissermaßen von außen vorgegeben

werden, oder wir können auch versuchen, sie innerhalb des Netzes (endogen) zu bestimmen.

Bovens und Hartmann (2006, Kap. 3) nutzen das so beschriebene Netz in der folgenden Weise. Sie wählen als Faktoren zweiwertige Aussagenvariablen und führen damit gewisse intuitive Größen ein:

(1) $P(Rel_i) = \varrho$ sei der Zuverlässigkeitsparameter bzw. die Anfangswahrscheinlichkeit dafür, dass der betreffende Zeuge zuverlässig ist, wobei zur Vereinfachung alle Zeugen hier gleich behandelt werden.
(2) $P(Hyp) = h$ sei die Startwahrscheinlichkeit der Behauptung, dass ein bestimmter Sachverhalt vorliegt.
(3) $P(Rep_i|Hyp \,\&\, Rel_i) = 1$ und $P(Rep_i|\neg Hyp \,\&\, Rel_i) = 0$ für zuverlässige Zeugen und
(4) $P(Rep_i|Hyp \,\&\, \neg Rel_i) = P(Rep_i|\neg Hyp \,\&\, \neg Rel_i) = a$ für unzuverlässige Zeugen mit einem Randomisierungsparameter a.

Dabei sollen alle Parameter echt zwischen 0 und 1 liegen: $0 < \varrho; h, a < 1$. Zuverlässige Zeugen berichten jeweils korrekt darüber, ob Hyp vorliegt oder nicht (3), während unzuverlässige Zeugen eher zufällig mit einer gewissen Wahrscheinlichkeit a über das Vorliegen von Hyp berichten, die aber unabhängig davon ist, ob die in Frage stehende Behauptung (oder Hypothese) wahr ist oder nicht (4). Weiterhin können wir einige Unabhängigkeitsannahmen aus dem Graphen ablesen. Die Zuverlässigkeit Rel_i der Zeugen und die Hypothese Hyp werden als Wurzelfaktoren unseres Netzes alle als voneinander statistisch unabhängig betrachtet. Das modelliert unsere Annahme, dass es sich um unabhängige Zeugenberichte von unzuverlässigen Zeugen handelt und die Zuverlässigkeit auch nicht von dem speziellen Sachverhalt abhängt, um den es geht:

$Rel_i \perp Rel_j$, Hyp für alle $i \neq j$

Außerdem folgt in dieser Modellierung, dass die Berichte voneinander unabhängig sind gegeben Hyp:

$(Rep_i \perp Rep_j \mid Hyp)$ für alle $i \neq j$

Wenn wir nun die Information erhalten, dass n Zeugen übereinstimmend berichten, dass Hyp vorliegt, so können wir uns fragen und das dann

ausrechnen, wie sich das auf die Wahrscheinlichkeiten für die anderen Faktoren auswirkt. Definieren wir etwa die entsprechende n-fache Updatefunktion $P^{+(n)}$ für eine Variable X so:

$$P^{+(n)}(X) := P(X|Rep_1 \,\&\ldots\&\, Rep_n),$$

dann liefern uns Bovens & Hartmann die folgenden Resultate:

$$(1)\ P^{+(n)}(Rel_i) = [h\cdot(1-x)] \,/\, [h+(1-h)\cdot x^n]$$

mit dem Likelihoodquotienten (für $0 < x < 1$) für einen einzelnen Bericht:

$$x = P(Rep_i|\neg Hyp) \,/\, P(Rep_i|Hyp) = a\cdot(1-\varrho) \,/\, [\varrho+a\cdot(1-\varrho)]$$

Im Grenzfall, in dem die Anzahl der Zeugen immer weiter wächst, erhalten wir so das Resultat:

$$(1^*)\ \lim_{n\to\infty} P^{+(n)}(Rel_i) = 1-x = \varrho \,/\, [\varrho+a\cdot(1-\varrho)]$$

Unsere Annahmen über die Zuverlässigkeit der Zeugen konvergieren also gegen eine obere Schranke, die echt kleiner als 1 ist und vor allem vom Randomisierungsparameter a abhängt. Dieses Ergebnis wird von Bovens und Hartmann noch ausführlicher diskutiert. Aber das wichtigere Resultat ist wohl, dass zumindest die Wahrscheinlichkeit für die Hypothese Hyp im Grenzfall gegen 1 konvergiert:

$$(2)\ P^{+(n)}(Hyp) = h \,/\, [h+(1-h)\cdot x^n]$$

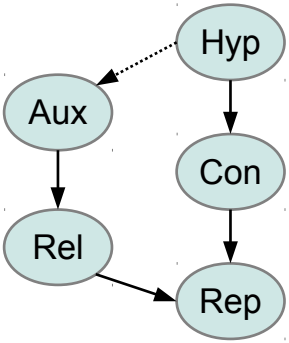
Damit ist klar, dass gilt:

$$(2^*)\ \lim_{n\to\infty} P^{+(n)}(Hyp) = 1$$

Das illustriert wiederum, wie wir den Bayesianismus durch geeignete Modellierungen, bei denen unsere kausalen Annahmen über die betreffende Situation eingehen, einsetzen können. Der Graph unseres bayesschen Netzes beschreibt dazu in sehr anschaulicher Weise die grundlegenden kausalen Zusammenhänge, die wir voraussetzen. Bovens und Hartmann (2006, Kap. 4) untersuchen in ähnlicher Weise, was passiert, wenn wir

mehrere unzuverlässige Messungen erhalten, die alle für eine Theorie sprechen.

Insbesondere modellieren sie aber noch in ähnlicher Weise eine Situation, in der wir eine Konsequenz (Con) aus einer Hypothese ziehen und Rep die bestätigende Messung bezeichnet, während mit Aux eine Hilfshypothese beschrieben wird, die gerade besagt, dass unser Messgerät zuverlässig arbeitet. Anhand von dieser Modellierung untersuchen sie dann auch das Duhem-Quine-Problem.



Bayessches Netz 5: Messung mit Hilfshypothesen

Dieses bayessche Netz soll zugleich zwei Situationen darstellen, die sich darin unterscheiden, ob die Hypothese Hyp und die Hilfshypothese Aux unabhängig sind oder nicht. Das wird in der Grafik durch den gestrichelten Pfeil wiedergegeben. Das Besondere ist dabei vor allem, dass überhaupt der Fall betrachtet wird, dass Hyp und Aux probabilistisch abhängig sind. Denken wir etwa an das Beispiel, dass wir eine Stromstärke durch ein Amperemeter messen, dessen Funktionieren selbst z.T. durch die Maxwellsche Elektrodynamik beschrieben wird, die wir dabei aber letztlich erst noch testen wollen.

Als Ergebnis werden genauere Bereiche für die Startwahrscheinlichkeiten benannt, für die eine Messung im unabhängigen Fall eine bessere Bestätigung der Hypothese liefert und wann sie im abhängigen Fall eine bessere Stützung liefert. Doch die möchte ich hier nicht weiter verfolgen, sondern noch ein wenig auf die Problematik der Modellierung solcher Situationen eingehen. Die obigen Modellierungen zeichnen sich durch relativ viele Faktoren aus, die hier zunächst unterschieden werden,

doch die werden im Prinzip danach wieder gleich reduziert, indem etwa angenommen wird, dass das Messgerät genau dann zuverlässig ist, wenn die Hilfshypothese stimmt und die Konsequenz genau dann wahr ist, wenn die Hypothese wahr ist:

$P(Rel|Aux) = 1 \quad P(Con|Hyp) = 1$
$P(Rel|\neg Aux) = 0 \quad P(Con|\neg Hyp) = 0$

Dadurch wird die Anzahl der Faktoren implizit wieder reduziert und wir sehen, wie schwierig es oft ist, die angenommenen Zusammenhänge überhaupt in ein bayesianisches Modell zu übersetzen. Das verlangt zumindest eine gewisse Expertise für das betreffende Gebiet, und man kann keineswegs behaupten, dass uns der Bayesianismus hier automatisch zu bestimmten Modellierungen und Berechnungen hin führt. Dazu kommt noch die Schätzung der Wahrscheinlichkeiten, die ebenfalls alles andere als unproblematisch ist. In den konkreten Anwendungen offenbaren sich so die Probleme, aber natürlich auch etliche Stärken des Bayesianismus.

Wie würde der Abduktionsvertreter an die Frage herangehen, wann wir bestimmten Zeugenaussagen Glauben schenken sollten? Er fragt danach, wie die Berichte wohl entstanden sind, was also als beste Erklärung dieser Aussagen dienen kann. Ist die beste Erklärung für die zwei Zeugenaussagen, dass sie beide H beobachtet haben, oder ist eine bessere Erklärung, dass der erste lügt und der andere sich hat vom ersten Zeugen beeinflussen lassen oder ist eine andere Erklärung die beste? Dazu kann er gezielt nach weiteren Fakten suchen, die zwischen den beiden Erklärungsansätzen zu entscheiden helfen. Dabei erhält der Abduktionsvertreter allerdings noch keine quantitative Abschätzung der Bestätigung von H. Dazu müsste er versuchen, die jeweiligen Erklärungsstärken miteinander zu vergleichen, wozu man u.a. wieder die betreffenden Likelihoods mit heranziehen würde. Der Faktor q bestimmt dabei gerade die Erklärungsstärke, die die Annahme ¬H im Vergleich zur Erklärung durch H für das Auftreten der beiden Zeugenaussagen E_1 und E_2 bietet. Auf diesem Wege reproduziert der Abduktionsvertreter auf etwas andere Weise mit etwas anderen Werten hier bestimmte Ergebnisse der Bayesianer. Wenn q größer wird, spricht das für ¬H, und wenn q kleiner wird, spricht das für H. Hierbei werden die Likelihoods allerdings möglichst objektiv gedeutet, was aber nicht verhindert, dass auch der

Abduktionsvertreter sie schätzen muss und sie nicht einfach messen kann.

5.9.9 Der Trugschluss des Anklägers

Im Bereich der Verwertung von Indizien bei Gericht sind viele Anwendungen für das induktive Schließen zu finden und gerade bei Wahrscheinlichkeitsschlüssen kann man dort erstaunliche Fehlschlüsse entdecken. Einige haben sogar einen Namen erhalten und sollen kurz vorgestellt werden. Beginnen möchte ich mit einer kurzen Schilderung eines Beispiels aus dem Buch von Gigerenzer (2002, 209ff.), das viele weitere Beispiele für die Anwendung bayesianischer Überlegungen in der Praxis bietet und insbesondere darauf verweist, dass wir viele Fehler in unserem intuitiven Umgang mit Wahrscheinlichkeiten vermeiden könnten, wenn wir zu entsprechenden Häufigkeitsüberlegungen übergehen, was ich des Öfteren in diesem Buch zu beherzigen versucht habe.

In Los Angeles kam es 1964 zu einem Raubüberfall und Zeugen sahen die Täterin enteilen. Sie beschrieben sie als Blondine mit Pferdeschwanz. Sie stieg in ein gelbes Auto, an dessen Steuer ein Schwarzer mit Backenbart und Schnurrbart saß. Die Polizei verhaftete daraufhin das Ehepaar Collins, auf das die genannten Merkmale alle zutrafen. Da eine sichere Identifizierung nicht gelang, musste die Staatsanwaltschaft die Anklage ganz auf diese Merkmale stützen. Daraufhin trat ein Mathematiker als Sachverständiger auf, der das Ehepaar Collins »überführte«. Er gab Wahrscheinlichkeiten für die einzelnen Merkmale an und multiplizierte sie dann:

P(Mädchen mit blondem Haar) = 1/3
P(Mädchen mit Pferdeschwanz) = 1/10
P(teilweise gelbes Auto) = 1/10
P(Mann mit Schnurrbart) = 1/4
P(Schwarzer mit Bart) = 1/10
P(Pärchen beiderlei Hautfarbe) = 1/1000
Also: P(Paar mit allen Merkmalen) = 1/12 000 000

Er schloss also, dass es nur eine Chance von 1 zu 12 Millionen gäbe, dass ein Paar alle diese Merkmale aufwiese. Und das wäre noch vorsichtig

gerechnet. Daher wäre die Wahrscheinlichkeit, dass jemand Anderes die Tat begangen hätte (statt des Ehepaar Collins), also nur 1/12 000 000. Klar, dass die Geschworenen das Ehepaar Collins bei so viel Sachverstand verurteilten.

Erst der oberste Gerichtshof von Kalifornien hob das Urteil auf und verwies darauf, dass hier wohl alle Fehler gemacht wurden, die man in solchen Fällen überhaupt nur machen kann. 1. Die Wahrscheinlichkeiten waren nur Schätzungen und entbehrten einer empirischen Grundlage. Gerichte geben sich in der Regel nicht mit rein subjektiven Wahrscheinlichkeiten zufrieden, sondern verlangen eine Likelihoodanbindung oder eine Herleitung der Startwahrscheinlichkeiten mit Hilfe des statistischen Syllogismus. Die Schätzungen wirken dagegen eher willkürlich. 2. Die Wahrscheinlichkeiten wären nur dann zu multiplizieren, wenn für diese Merkmale eine statistische Unabhängigkeit vorliegen würde. Das ist keineswegs plausibel, denn z.B. einen Bart zu haben ist sicher damit korreliert, auch einen Schnurrbart zu haben. Das müsste letztlich empirisch genauer untersucht werden. 3. Die Unsicherheiten, die in den Aussagen enthalten waren und darin, dass die Merkmale veränderlich sind, waren in den Abschätzungen nicht enthalten. Es könnte sich auch um eine relativ hellhäutige Farbige mit blond gefärbten Haaren handeln. Das alles zeigt schon, mit welcher Vorsicht die Werte zu behandeln sind, aber der schwerwiegendste Fehler ist der sogenannte *Trugschluss des Anklägers*: 4. Es findet eine Verwechslung statt zwischen der Wahrscheinlichkeit dafür, dass alle gesuchten Merkmale in einer Person übereinstimmen P(Übereinstimmung) und der Wahrscheinlichkeit, dass eine solche Person unschuldig ist: P(unschuldig|Übereinstimmung).

Trugschluss des Anklägers: Verwechslung von
P(Merkmale treffen auf eine bestimmte Person zu) und
P(Person ist unschuldig|Merkmale treffen auf diese Person zu)

Betrachten wir das obige Beispiel, um die Verwechslung deutlich zu machen. Nehmen wir an, dass eine Korrektur der genannten Wahrscheinlichkeiten zu dem Ergebnis komme, dass die Wahrscheinlichkeit für eine Übereinstimmung nur 1/1000 wäre. Ist dann die Wahrscheinlichkeit für die Unschuld gegeben eine Übereinstimmung auch nur noch 1/1000 und damit die Gegenwahrscheinlichkeit, also die Wahrscheinlichkeit für

die Schuld 999/1000? Das würde schließlich immer noch massiv gegen die Collins sprechen – oder nicht? Nein, das ist keineswegs so einfach. Nehmen wir an, es gab ca. 2,5 Millionen Einwohner in Los Angeles in den 60er Jahren. Nehmen wir weiter an, dass wir keine weiteren Einschränkungen vornehmen können, als die bereits genannten und jeder Bürger von LA als potentieller Verdächtiger in Frage kommt, dann würden auf jeden 1000-ten Bürger von LA alle Merkmale zutreffen. Damit hätten wir in LA ca. 2500 Bürger (lassen wir die Komplikation einfach einmal weg, dass es sich um zwei Personen dabei handelt), auf die die Merkmale alle zutreffen. Nur eine davon wäre die Täterin, d.h., wir hätten eine Auswahl von 1 aus 2500 und damit wäre P(schuldig|Übereinstimmung) = 1/2500 und damit P(unschuldig|Übereinstimmung) = 2499/2500. Das würde jedenfalls nicht ausreichen, um von der Schuld der Collins überzeugt zu sein. Man erkennt daran sehr gut, dass es sich bei dem Schluss des Sachverständigen und des Anklägers um einen Trugschluss handelt. Wir müssen typischerweise genau hinschauen, von welchen Wahrscheinlichkeiten jeweils die Rede ist. Leider kommen solche Trugschlüsse häufig bei Laien und sogar bei Sachverständigen vor (vgl. Schweizer 2006).

Ähnliche probabilistische Trugschlüsse versuchen natürlich auch die Verteidiger für sich zu nutzen (ob bewusst oder unbewusst sei dahingestellt). Im Prozess gegen O.J. Simpson wurde gegen ihn u.a. vorgebracht, dass er seine Frau geschlagen hatte. Doch spricht das für ihn als Täter? Intuitiv und psychologisch gesehen scheint das klar gegen ihn zu sprechen, selbst wenn es noch kein zwingender Hinweis auf seine Täterschaft sein kann. Alan Dershowitz (einer der Verteidiger von O.J.) versuchte statistisch aber sogar positives Kapital aus den Prügelattacken zu ziehen. Er wies darauf hin, dass nur etwa einer von 1000 der Männer, die ihre Frauen schlagen, diese auch umbringen. Spricht das nicht für die Unschuld seines Mandanten? Leider hat die Anklage diesen Punkt nicht entkräftet. Man hätte auf folgenden Unterschied hinweisen können: P(Ehemann ist schuldig|hat Frau geschlagen) ist eine ganz andere als P(Ehemann ist schuldig|hat Frau geschlagen & Frau wurde ermordet) (vgl. dazu Beck-Bornholdt & Dubben 2003, 95 ff). Es geht uns doch um die Frage, wie viele von den ermordeten Frauen wurden von ihrem Ehemann ermordet. Gibt in diesen Fällen nicht doch das Prügelverhalten einen Hinweis auf die Täterschaft. Tatsächlich ergaben

Abschätzungen, dass von den ermordeten Frauen mindestens jede zweite von ihrem prügelnden Ehemann ermordet wurde (vgl. Schweizer 2006) oder sogar 4 von 5 (vgl. dazu Beck-Bornholdt & Dubben 2003, 95 ff). Bei Schweizer (2006) finden sich dazu weitere Literaturhinweise auf die kontrovers geführte Diskussion, ob solche bayesianischen Überlegungen überhaupt in Gerichtsverfahren eingesetzt werden sollten. Doch für einige Zusammenhänge kommt man wohl um bestimmte probabilistische Überlegungen nicht herum. Das zeigt sich z.B. dort, wo wir uns fragen, wie sehr wir typischen Indizien wie Fingerabdrücken oder DNA-Spuren vertrauen dürfen. Das ist keineswegs so einfach zu beantworten, wie es leider des Öfteren vor Gericht angenommen wird.

5.9.10 Die Bayessche Analyse von DNA-Beweisen

DNA-Beweise im kriminologischen Bereich (auch als genetischer Fingerabdruck bezeichnet) lassen sich mit dem üblichen bayesianischen Schema bewerten. Allerdings benötigen wir dazu eine ganze Reihe von Daten und daran krankt es. Bei der DNA-Analyse handelt es sich um ein komplexes Geschäft mit schwierigen Interpretationsproblemen. Auf die inhaltlichen Details werde ich hier nicht eingehen, sondern nur speziell auf einige Ergebnisse zur Fehlerhäufigkeit solcher Analysen. Gigerenzer (2002, 227 ff.) weist auf einige Irrtumsquellen in entsprechenden Gerichtsverfahren hin und führt dazu bestimmte Begriffe ein. Nehmen wir der Einfachheit halber an, dass es praktisch auszuschließen ist (sehr kleine Wahrscheinlichkeit), dass zwei Personen dasselbe DNA-Profil haben. Das stimmt zwar insbesondere für enge Verwandte nicht (und schon gar nicht für eineiige Zwillinge), aber mit diesen Spezialfällen werden wir uns nicht weiter beschäftigen. Dann müssen wir jedoch trotzdem die Laborfehler oder Möglichkeiten der Verunreinigung berücksichtigen, um die *Urheberwahrscheinlichkeit* zu bestimmen, d.h. die Wahrscheinlichkeit, mit der eine bestimmte Person tatsächlich der Urheber der DNA-Spuren ist. Des weiteren muss auch eine Person, deren DNA sich am Tatort findet, keineswegs gleich der Täter sein. Hier gibt es viele weitere Fehlermöglichkeiten, die allerdings stark vom Einzelfall abhängen dürften.

Mir geht es hier vor allem um die genannte Urheberwahrscheinlichkeit und die Wahrscheinlichkeit falsch positiver Laborergebnisse: $P(L^+|\neg H)$ (Wahrscheinlichkeit dafür, dass Laborergebnis positiv ausfällt (L^+), obwohl der Verdächtige nicht der Täter ist ($\neg H$)). Diese Wahrscheinlichkeit sollte doch zumindest sehr niedrig sein. Doch welche Hinweise haben wir darauf? Leider versuchen staatliche Stellen wie das FBI alle Daten darüber geheim zu halten (vgl. Gigerenzer 2002, 229ff.), was uns natürlich stutzig machen muss. Trotzdem leugnen viele Sachverständige bereits die Möglichkeit einer fehlerhaften Ermittlung der Urheberschaft überhaupt. Im Prozess gegen O.J. Simpson musste das Labor dagegen eine Falsch-positiv-Rate von 1 zu 200 einräumen. Im Rahmen einer größeren Untersuchung kam Jonathan Koehler (1995) sogar auf eine Falsch-positiv-Rate von 1 zu 100. Das können wir für unsere Likelihoodanbindung nutzen.

Was bedeutet das aber in konkreten Fällen? Solche DNA-Tests werden gerne bei Massenuntersuchungen eingesetzt, mit vielen Verdächtigen und kaum weiteren Hinweisen auf den Täter. Gigerenzer (2002, 221) schildert einen Fall aus Oldenburg aus dem Jahre 1998, in dem 15000 Männer zwischen 18 und 30 getestet wurden. Nehmen wir also an, wir wüssten schon, dass einer der Männer der Täter ist. Fritz ist einer von ihnen und hat daher eine Ausgangswahrscheinlichkeit von 1/15000 dafür der Täter zu sein. Nehmen wir einfach an, die Wahrscheinlichkeit für ein positives Laborergebnis bei tatsächlicher Urheberschaft sei sogar $P(L^+|H) = 1$ und die Falsch-positiv-Rate (zugunsten des Täters) $P(L^+|\neg H) = 1/100$. Dann erhalten wir beim Updaten:

$$P(H|L^+) \approx \frac{1}{151} \approx 0{,}0067 \text{ oder } \frac{2}{3}\%$$

Damit bliebe die Wahrscheinlichkeit für die Urheberschaft von Fritz selbst bei einem positiven Ergebnis deutlich unter einem Prozent. Hätte der Verdächtige nicht gestanden, wäre ihm eigentlich noch nicht viel nachgewiesen worden. Doch hätten die Ermittler oder die Geschworenen das auch erkannt? Leider steht zu befürchten, dass wir hier wieder dem Basisratenfehlschluss begegnet wären. Natürlich hätte man den Test wiederholen können, und es ist nicht leicht zu sagen, zu welchen neuen Wahrscheinlichkeiten das geführt hätte. Das hängt vor allem davon ab, wie es genau zu dem ersten Fehler gekommen sein könnte.

Mit einem Geständnis könnten wir natürlich wiederum fragen, wie hoch die Wahrscheinlichkeit für ein falsches Geständnis ist, wenn die Verdächtigen massiv unter Druck gesetzt werden, wie das in diesem Fall sicherlich gegeben war. Um das weiter zu verfolgen, müssten wir jedoch zunächst möglichst genaue empirische Daten darüber haben, wie hier die Falsch-positiv-Rate unter verschiedenen Bedingungen aussieht. Dann könnten wir wieder normal updaten. Das dürfte aber kaum zu ermitteln sein und welchen Wert sollte man dafür schätzen? Die bayesianische Analyse droht dann selbst subjektiv zu werden.

Einen ganz anderen Aspekt finden wir noch in der Unschuldsvermutung in unserem Strafrecht. Soll die etwa bedeuten, dass wir nur dann von der Täterschaft ausgehen dürfen, wenn die Täterwahrscheinlichkeit über 90% oder sogar über 99% liegt? Dann gäbe es vermutlich nur noch sehr wenige Verurteilungen. Es wäre aber zumindest interessant, Juristen dazu zu befragen, wo ihrer Meinung nach diese Schwelle zu suchen sein sollte. Doch das ist kein überwiegend erkenntnistheoretisches Problem mehr und soll daher hier nicht weiter verfolgt werden.

5.10 Fazit 1: Möglichst objektiver Bayesianismus für die Theorienwahl

Die Glosse über Prof. Wichtig (Kap. 5.5.11) sollte belegen, wie wichtig für den Bayesianismus *vernünftige* Wahrscheinlichkeitseinschätzungen sind. Wir können das auch so formulieren, dass wir die Kritik am Bayesianismus, er sei zu subjektiv, sehr ernst nehmen müssen. Wir sollten also alle Gelegenheiten für Objektivierungen beim Schopf ergreifen. Überhaupt hat sich an einigen Beispielen gezeigt, wie schwierig es ist, ein guter Bayesianer zu sein, weil es noch nicht einmal leicht ist, überhaupt konsistente Wahrscheinlichkeiten für mehrere Aussagen zu vergeben. Das alles spricht dafür, dass wir dort, wo es irgendwie möglich erscheint, unsere subjektiven Wahrscheinlichkeiten an objektiven Wahrscheinlichkeiten bzw. an relativen Häufigkeiten orientieren sollten – und das gilt vor allem für die Startwahrscheinlichkeiten.

Dazu können in erster Linie die *Wahrscheinlichkeitskoordinierungsprinzipien* dienen wie der statistische Syllogismus, das Hauptprinzip und

die Likelihoodanbindung. Daneben sollten wir Überlegungen aus der induktiven Logik ernst nehmen und mit Hilfe von Indifferenzprinzipien nach vernünftigen Startwahrscheinlichkeiten zu suchen, um nicht dem Fehler von Prof. Wichtig anheimzufallen, dass ein Vorurteil all unsere Beobachtungen aus dem Felde schlägt. Wenn es nicht gelingt, mit Hilfe solcher Prinzipien zu eindeutigen Wahrscheinlichkeiten zu gelangen stehen uns die Wege offen, die Hawthorne und Maher nennen, dass wir mit mehreren erlaubten Startwahrscheinlichkeiten arbeiten können oder uns im Zweifelsfall auch einmal einer Wahrscheinlichkeitsvergabe ganz enthalten. Für die weitere Ausgestaltung unseres Überzeugungssystems sind wir dann auf weiteres Hintergrundwissen angewiesen, wie z.B. unser Wissen um grundlegende kausale Zusammenhänge. Die können im Bayesianismus etwa in Form bayesscher Netze eingebracht werden und führen erst zu einigermaßen handhabbaren probabilistischen Überzeugungssystemen.

Überhaupt sollten wir uns im Normalfall auf kleinere Mengen von Aussagen bzw. lokale Überzeugungssysteme beschränken, die für uns noch einigermaßen überschaubar bleiben. So können wir uns etwa der Aufgabe widmen aus einer vorgegebenen Liste von Hypothesen anhand bestimmter Daten einen Sieger zu küren. Weiteres Hintergrundwissen kann dann etwa über die Startwahrscheinlichkeiten einfließen. Das Verfahren der Theorienwahl sollte dann etwa wie folgt aussehen:

Vorschlag: Bayesianisches Verfahren der Theorienwahl

1. Schritt: Man stelle eine Liste $L = \{H_1,\ldots,H_n\}$ von Hypothesen auf, die einander ausschließen, die man aber auch für erschöpfend hält. Dazu lasse man nur Hypothesen zu, die alle bereits über eine erhebliche Erklärungskraft verfügen, um dem zweiten epistemischen Ziel zu genügen.

2. Schritt: Man vergebe Startwahrscheinlichkeiten für die Hypothesen und unsere Daten $E_1,\ldots,E_m$ nach möglichst objektiven Gesichtspunkten. Für die Daten sollten wir uns vor allem auf die Likelihoodanbindung stützen (soweit objektive Likelihoods vorhanden sind) für die Hypothesen auf den statistischen Syllogismus (wenn entsprechende Daten vorliegen) und ansonsten auf ein Indifferenzprinzip (IP), das allen Hypothesen dieselbe Wahrscheinlichkeit einräumt. Zusätzlich

kann weiteres Hintergrundwissen (HW) einfließen, das die Startwahrscheinlichkeit der Hypothesen in moderater Weise verändern kann: $P(H_i) = P_{IP+HW}(H_i)$. Auch spezielles kausales Hintergrundwissen kann durch die Aufstellung bayesscher Netze eingebracht werden, die es uns erleichtern, die Startwahrscheinlichkeiten zu vergeben.

3. Schritt: Updaten der Hypothesen H_i mit den auftretenden Daten $E_1,\ldots,E_m$ (bzw. $E^m \equiv E_1\&\ldots\&E_m$) möglichst anhand einer stabilen Likelihoodanbindung (d.h. $P(E^m|H_i) = P_{Hi}(E^m)$) ganz im Sinne reiner Bestätigungsfunktionen. Aufgrund der Likelihoodanbindung sind die Daten bedingt unabhängig und es gilt dann:

$P(E^m|H_i) = P(E_1|H_i)\cdot\ldots\cdot P(E_m|H_i) = P_{Hi}(E^m) = P_{Hi}(E_1)\cdot\ldots\cdot P_{Hi}(E_m)$

damit können wir nun auch $P(E^m)$ auf objektive Wahrscheinlichkeiten zurückführen:

$P(E^m) = \sum_j\ P(E^m|H_j)\cdot P(H_j) = \sum_j\ P_{Hj}(E^m)\cdot P_{IP+HW}(H_j)$

und insgesamt ergibt sich:

(*) $P(H_i|E_1\&\ldots\&E_m) = P_{IP+HW}(H_i)\ (P(E^m|H_i)\ /\ P(E^m))$

Wir sind im günstigen Fall so nur auf objektiv bestimmbare Wahrscheinlichkeiten angewiesen und das Updaten erhält damit den Status einer objektiven Bestätigungsfunktion.

4. Schritt: Auswahl der besten Theorie relativ zu den anderen. Dabei wird man vor allem darauf setzen, ob es einer Theorie gelingt, sich durch das Updaten (*) von den anderen abzusetzen und einen deutlichen Wahrscheinlichkeitsvorsprung gegenüber allen anderen zu erzielen. Die sollten wir dann als unseren besten Tipp betrachten und entsprechend (mit den nötigen Vorbehalten für induktives Schließen) als akzeptiert auswählen. Hier können wir eventuell noch bestimmte untere Grenzen für ihre Wahrscheinlichkeit fordern (etwa > 0,5), doch das könnte in der Praxis durchaus zu Problemen führen, wenn die Startwahrscheinlichkeiten der Hypothesen schon recht klein sind. Daher würde ich darauf verzichten.

Das so beschriebene Verfahren entspricht mehr unserem tatsächlichen Vorgehen auch in Beispielfällen. Es ist sicher realistischer, als anzunehmen, wir verfügten über ein gesamtes probabilistisches Überzeugungs-

system mit seiner riesigen Anzahl an erforderlichen Startwahrscheinlichkeiten. Das weitere Hintergrundwissen sollten wir uns vermutlich eher als zweiwertig (bzw. dreiwertig) im klassischen Sinne einsortiert vorstellen. So wird der Bayesianismus handhabbar und entspricht dann in der konkreten Anwendung weitgehend den Bestätigungsfunktionen im Sinne von Hawthorne.

Außerdem sollten wir natürlich immer über den Tellerrand schauen und Vergleiche mit anderen Ansätzen der Datenauswertung ernst nehmen. Die Likelihoodisten und die klassischen Statistiker setzen ganz auf die objektiven Likelihoods und versuchen so, gerade die möglicherweise subjektiven Anteile am Bayesianismus zu vermeiden. Allerdings übersehen sie damit möglicherweise wichtiges Hintergrundwissen, das wir bereits besitzen etwa über Basisraten oder kausale Zusammenhänge. Außerdem steht der Bayesianismus ganz in der Tradition des Empirismus und übersieht daher in seiner klassischen Konzeption ein wesentliches epistemisches Ziel unserer Theoriensuche, nämlich die Suche nach erklärungsstarken Theorien. Dem möchte ich hier aber noch einmal besondere Aufmerksamkeit und somit einen eigenständigen Abschnitt widmen.

5.11 Fazit 2: Bayesianismus und abduktives Schließen

Sind der Schluss auf die beste Erklärung und das bayesianische Schließen Konkurrenten um das beste Induktionsverfahren? Davon muss man z.T. wohl ausgehen, denn beide stellen unterschiedliche Aspekte in den Vordergrund und sollten daher in geeigneten Fällen zu unterschiedlichen Ratschlägen führen. Trotzdem gibt es natürlich auch eine ganze Reihe von Gemeinsamkeiten, von denen einige bereits zur Sprache gekommen sind und die in dem oben beschriebenen Verfahren der bayesianischen Theorienwahl deutlich werden. Die Schwerpunkte und Zielsetzungen der beiden Ansätze sind aber zunächst andere. Der Bayesianismus ist nicht in erster Linie ein Verfahren zur Theorienwahl, denn es werden jeder Theorie und ihren Konkurrenten jeweils immer von Null verschiedene subjektive Wahrscheinlichkeiten bzw. Plausibilitäts- oder Glaubensgrade zugewiesen. Dabei mag eine Theorie vorne liegen, aber es ist im

klassischen Bayesianismus nichts darüber gesagt, ob wir sie und unter welchen genauen Umständen wir sie akzeptieren sollten. In Abschnitt 5.3.2 hatte ich dafür argumentiert, dass wir durchaus Kriterien für das *Akzeptieren von Theorien* benötigen und dazu am ehesten eine Schwellenwertkonzeption passt (die man noch im Sinne von Hawthornes Glaubenslogik modifizieren kann s.o.), wonach wir A akzeptieren sollten, sobald etwa P(A)>0,5 gilt. Im letzen Abschnitt habe ich aber für eine liberalere Regelung plädiert. Mit diesen beiden Vorschlägen möchte ich hier der Einfachheit halber arbeiten, obwohl Bayesianer sich oft weigern, über das Akzeptieren von Theorien zu sprechen. Sie sollten jedenfalls zumindest zustimmen, dass eine höhere Wahrscheinlichkeit für eine Theorie eine epistemisch positive Auszeichnung der Theorie darstellt, denn sonst verlieren sie jede Anbindung an die klassische Erkenntnistheorie.

Der Schluss auf die beste Erklärung ist dagegen typischerweise ein Verfahren der Theorienwahl, das aber natürlich nicht in jedem Fall zu einer Auswahl einer Theorie führen muss. Eventuell gibt es nahezu gleich gute Erklärungskandidaten oder die beste Erklärung stellt immer noch eine sehr schlechte Erklärung dar und erfüllt noch nicht einmal minimale Anforderungen, dann wird auch der Abduktivist keinen Sieger im Sinne der Theorienwahl küren. Doch gehen wir zunächst einmal von einfachen und klaren Fällen aus. Wählen dann beide Verfahren jeweils dieselben Theorien?

Das ist nicht zu erwarten. Während der Bayesianismus eher auf der Linie der logischen Empiristen liegt und vor allem darauf setzt, dass die ausgewählte Theorie als sehr wahrscheinlich im Lichte unsere Daten dasteht, setzt der Abduktivismus vor allem auf die Erklärungskraft der Theorie und natürlich darauf, dass die vorliegenden Daten gut durch die Theorie erklärt werden. Beide Ansätze verlangen also, dass die Daten in bestimmter Weise aus der zu wählenden Theorie folgen. Für den Bayesianer sollten sie im Lichte der Theorie eine möglichst hohe Wahrscheinlichkeit aufweisen und der Idealfall ist der, in dem sie deduktiv aus unserer Theorie herleitbar sind. Das ist für eine Abduktion nicht nötig. Sie verlangt vielmehr vor allem, dass die Theorie nomische Muster enthält, die aufzeigen, wie die zu erklärenden Daten sich kausal aus früheren Zuständen entwickelt haben. Auch hier bleibt die deduktive

Herleitung der Daten ein Grenzfall, doch die deduktive Ableitung ist für sich genommen noch alles andere als ideal, denn oft stellt sie keine Erklärung der Daten dar. Den einfachsten Fall finden wir in der reinen »Beobachtungstheorie« $t \equiv d_1 \& \ldots \& d_n$, die nur aus einer Konjunktion der Beobachtungsdaten d_1 bis d_n besteht. Sie deckt zugleich wichtige Unterschiede zwischen den Ansätzen auf.

Die Theorie t ist für einen Bayesianer eigentlich eine ideale Theorie, denn sie folgt deduktiv aus den Daten und muss daher die Wahrscheinlichkeit 1 erhalten. Das scheint noch nicht problematisch zu sein, denn natürlich sollten wir t vertrauen, wenn wir den Daten vertrauen. Das Problem tritt allerdings auf, wenn wir t mit irgendeiner anderen vollwertigen Theorie T vergleichen, die die Daten erklärt, indem sie sie auf zugrundeliegende Gesetze zurückführt. Eine richtige Theorie T sollte dabei einen Überschussgehalt gegenüber den Daten haben und wird daher im Normalfall auch nach dem Updaten nicht die Wahrscheinlichkeit 1 aufweisen. Dann müsste ein Bayesianer aber t immer gegenüber T vorziehen, doch das entspricht überhaupt nicht dem Verfahren der Theorienwahl in der Wissenschaft und noch nicht einmal der im Alltag. Die Theorie t gestattet es zwar, unsere Daten abzuleiten, aber sie besitzt überhaupt keine Erklärungskraft; sie sollte daher kein Konkurrent für T sein. Hier kommt leider die Theoriefeindlichkeit des Bayesianismus zum Vorschein. Auch der Abduktivist wird natürlich die deduktiven Zusammenhänge respektieren (im Rahmen der Kohärenztheorie), aber für ihn verlangt eine Theorienwahl vor allem nach Erklärungen, die t nicht liefern kann. Denken wir an das Beispiel der Choleratheorie von Snow. Diese Theorie schätzen wir für ihre große Erklärungskraft und Vereinheitlichungsleistung, zählt dagegen aber nur ihre Wahrscheinlichkeit, hätte sie keine Chance gegen eine Theorie vom Typ t.

Wir sind dazu in unserem Vorschlag für die bayesianische Theorienwahl dafür eingetreten, dass man im Hinblick auf die Wahrscheinlichkeit nur solche Theorien miteinander vergleichen sollte, die dieselbe Erklärungskraft aufweisen. Das ist sicher ein wesentlicher Schritt in die richtige Richtung. Allerdings führt er plötzlich ganz neue Gesichtspunkte in die bayesianische Erkenntnistheorie ein. Darüber sollten wir uns im Klaren sein. Wir müssen nun doch die *Erklärungsstärke* der

beteiligten Theorien kennen. Dafür hat der Bayesianer bisher keine eigenen Ressourcen. Es ist auch nicht zu erwarten, dass sich die auf die Wahrscheinlichkeiten zurückführen lässt, die dem Bayesianer vertraut sind. Die Erklärungsdebatte hat gezeigt, dass es keine einfache Explikation von Erklärung durch Wahrscheinlichkeitszusammenhänge gibt und dass insbesondere die *epistemischen* Wahrscheinlichkeiten dazu ungeeignet sind; zumindest würden wir physikalische (bzw. ontische) Wahrscheinlichkeiten benötigen, um mit ihrer Hilfe eine Erklärung eines Ereignisses abgeben zu können. Das ist ebenso das Problem, wenn Bayesianer versuchen, mit rein probabilistischen Mitteln Kohärenz zu explizieren. Dann können sie jedenfalls nicht die *Erklärungs*kohärenz damit erfassen, und meinen somit etwas anderes als das, was sich in der Rekonstruktion von historischen Fallstudien als hilfreicher Leitfaden der Theorienbeurteilung durch den Kohärenztheoretiker herausgestellt hat (vgl. etwa Thagard 2000).

Außerdem ergeben sich Probleme, wenn die Theorien nicht alle über die exakt gleiche Erklärungsstärke verfügen. Dann werden Verrechnungen zwischen Wahrscheinlichkeit und Erklärungsstärke bzw. dem Gehalt der Theorien erforderlich, wie wir sie auch im Rahmen unserer Konzeption von Erklärungskohärenz benötigen. Man könnte also sagen, dass wir hier wieder auf Poppers Idee zurückgreifen, dass wir bei Theorienwahl nach Theorien suchen, die einen hohen Gehalt haben. Für Popper bleibt das das einzige Kriterium, um unter den nichtfalsifizierten Theorien nach der besten Ausschau zu halten. Das sieht der Abduktivist anders. Er sucht nach einer Theorie, die möglichst viele der *vorliegenden Daten* auch tatsächlich erklären kann, also diese Daten etwa ableiten kann, dabei aber zugleich einen besonders hohen Gehalt aufweist. Unter hohem Gehalt ist wiederum die Erklärungsstärke gemeint, die auch Popper manchmal im Sinn zu haben scheint, wenn er etwa davon spricht, dass wir nach tiefen Theorien suchen. Der Vorschlag zur bayesianischen Theorienwahl aus dem letzten Abschnitt versucht bereits, diesen Gesichtspunkt zumindest mit einzubeziehen.

Während die logischen Empiristen also einen Pol der Debatte besetzen, an dem man nur auf die hohe Wahrscheinlichkeit der Theorien setzt, und Popper einen anderen Pol, an dem man nur auf den Gehalt der (nichtfalsifizierten) Theorien vertraut (was eher kleinen Wahrscheinlichkeiten

entspricht), sucht der Abduktivist und Kohärenzvertreter nach einem mittleren Weg. Einerseits zählt, dass eine Theorie möglichst viele Daten erklärt und dabei natürlich epistemisch durch diese Daten gestützt und bzw. begründet wird, andererseits zählt dabei die Erklärungsstärke und damit der Gehalt, den die Theorie zu bieten hat, der sich u.a. in möglichst invarianten nomischen Mustern manifestiert. Damit zählen etwa auch mögliche Daten, denn es geht darum, zugleich für neue mögliche Daten bereits eine Theorie zur Verfügung zu stellen, deren nomische Muster darauf anwendbar sind.

Es gibt einige wenige Bayesianer, die zumindest sehen, dass der (empirische) Gehalt ein wichtiges Kriterium der Theorienwahl darstellen sollte. Doch die Explikation des gesuchten Gehalts von Theorien bleibt dabei ein Problem, denn der ist m.E. nicht schlicht über die Wahrscheinlichkeit bzw. Plausibilitätsgrade einer Theorie explizierbar. Diese ändern sich, ohne dass sich deswegen schon die Erklärungskraft der Theorie verändert. Hier gibt es also Bedarf für eine wichtige Ergänzung des Bayesianismus, wobei noch unklar ist, wie die genau aussehen könnte. Gerade wenn der Bayesianismus etwas zur Theorienwahl in der Wissenschaft beisteuern möchte, muss er den Aspekt des Gehalts von Theorien im Blick behalten. Intuitiv werden wir oft vergleichbar starke Theorien gegeneinander antreten lassen, aber dieser Aspekt muss dann im bayesianischen Ansatz explizit genannt und möglichst weiter präzisiert werden.

Wir finden diesen Aspekt bereits in der allgemeinen Erkenntnistheorie. Es wird vermutlich unter dem Einfluss der logischen Empiristen und durch Betonung der Auseinandersetzung mit dem Skeptiker gerne übersehen, dass wir genau genommen zwei Ziele in der Erkenntnistheorie verfolgen. Vielleicht wird das zweite Ziel auch immer implizit vorausgesetzt, so dass wir nur vergessen, es explizit zu formulieren:

Die zwei grundlegenden Ziele der Erkenntnistheorie
1. Vermeide es, *falsche* Aussagen zu akzeptieren.
2. Akzeptiere möglichst viele wahre und *informative* Aussagen.

Im zweiten Ziel werden mit den »informativen« Aussagen gerade solche gesucht, die es uns ermöglichen zu verstehen, was in der Welt um uns

herum passiert, die also dafür Erklärungen liefern, und die es möglichst gestatten, in die Welt in unserem Sinne einzugreifen. Dabei handelt es sich also primär um erklärungsrelevantes Wissen, das gerade die Konzeption der Erklärungskohärenz in den Mittelpunkt stellt.

Besonders die Empiristen haben immer das erste Ziel in den Vordergrund gestellt und große Teile der Debatte in der Erkenntnistheorie, wie die Bekämpfung des Skeptikers, sind diesem Ziel gewidmet. Auch in der normalen Wissensdebatte geht es nur darum, Forderungen in puncto Sicherheit unserer Erkenntnis und Abwesenheit von relevanten Unterminierern aufzustellen und man fragt sich eigentlich nie, wofür man diesen Aufwand treibt. Ginge es wirklich nur um das erste Ziel, könnten wir uns schlicht den Skeptikern anschließen, die uns raten, dass wir uns einfach aller Mutmaßungen über unsere Welt enthalten sollen oder nur auf analytische Aussagen zurückzugreifen oder es bei den einfachen »Beobachtungstheorien« t zu belassen.

Das erste Ziel ist also für sich genommen auf recht triviale Weise zu erfüllen. Ebenso natürlich das zweite. Dazu müssten wir nur alle Aussagen unserer Sprache akzeptieren. Erst im Zusammenspiel der beiden Forderungen liegt die Brisanz unserer Erkenntnis. (In Niiniluoto 1999 werden daher beide Aspekte auch in seiner Konzeption von Wahrheitsähnlichkeit berücksichtigt.)

In der Wissenschaft wird das besonders deutlich, aber ebenso im Alltag: Um zu erklärungsrelevantem Wissen (also insbesondere zu Hypothesen mit Erklärungskraft) zu gelangen, sind wir gezwungen, ein gewisses epistemisches Risiko einzugehen (vgl. Bartelborth 2005). Sonst könnten wir bei mathematischem Wissen stehenbleiben oder noch die reinen Beobachtungsüberzeugungen hinzunehmen. Schon für die Auswahl der Beobachtungsüberzeugungen, die auch tatsächlich relativ sicher über die Welt Auskunft geben, sind wir allerdings auf allgemeines Wissen darüber, wie die Welt funktioniert angewiesen. Wir müssen etwa typische Irrtumsquellen kennen und dafür plädieren, dass die im konkreten Beobachtungsfall nicht vorliegen.

Angesichts der zwei Ziele wird der Unterschied zwischen den Bayesianern und den Abduktivisten und Kohärenzvertretern deutlicher. Letztere versuchen beide Ziele explizit zu berücksichtigen und sehen deshalb auch so gut aus, wenn es darum geht, Episoden der Wissenschaft zu re-

konstruieren. Im Bayesianismus wird das zweite Ziel bestenfalls versteckt vorkommen und höchstens implizit von den Bayesianern in ihre Vorher-Wahrscheinlichkeiten einfließen. Der Bayesianismus sollte an dieser Stelle ergänzt werden. Das gilt zumindest für einige Anwendungen. Im Falle der diagnostischen Schlüsse etwa und manch anderen Situationen mit vorgegeben Konkurrenzhypothesen kommt das Problem nicht so zum Tragen, da die Theorien wie die reine Beobachtungstheorie t nicht zur Debatte stehen. Man kann also auch dafür argumentieren, dass die verschiedenen Verfahren eher bereichsspezifisch mehr oder weniger gut geeignet sind. Für die diagnostischen Schlüsse liefert der Bayesianismus die überzeugendste Verrechnung all unserer Erkenntnisse, für die allgemeine Theorienwahl ist das nicht der Fall. Um dem Rechnung zu tragen, muss dieses Ziel bereits in die Auswahl einer Liste von Hypothesen, die dann bayesianisch miteinander verglichen werden eingebaut werden.

Es gibt aber natürlich zugleich engere Zusammenhänge zwischen dem Bayesianismus und dem Schluss auf die beste Erklärung. Die erkennt man am besten, wenn wir das bayessche Theorem in der schon in Kapitel 5.1 angeführten ausführlicheren Form betrachten und zunächst etwa annehmen, dass es nur zwei einander ausschließende Hypothesen H_1 und H_2 gibt (oben waren das H und ¬H) und dazu Daten $E^k \equiv E_1\&\ldots\&E_k$. Damit erhalten wir:

(1) $P(H_1|E^k) = P(H_1) \,/\, [P(H_1) + P(H_2)\ q_{1,2}(E^k)]$, wobei $q_{1,2}$ der folgende Quotient ist: $q_{1,2}(E^k) = P(E^k|H_2) \,/\, P(E^k|H_1)$

Hieran kann man erkennen, dass H_1 durch die Daten ceteris paribus umso besser bestätigt wird, je kleiner $q_{1,2}(E^k)$ ist, d.h., umso besser H_2 durch die Daten E^k relativ zu H_1 bestätigt wird. Doch der Quotient $q_{1,2}(E^k)$ stellt zugleich eine wichtige Dimension des Vergleichs der Erklärungskraft unserer beiden Hypothesen dar. Neben $q_{1,2}(E^k)$ gehen für den Bayesianer noch die Vorher-Wahrscheinlichkeiten in die Betrachtung ein.

Für eine vollständige Liste einander ausschließender Hypothesen $H_1,\ldots,H_n$ erhalten wir dann einen etwas komplexeren Ausdruck:

(2) $P(H_1|E^k) = P(H_1) \,/\, [P(H_1) + \sum_{i>1} P(H_i)\, q_{1,i}(E^k)]$, wobei $q_{1,i}$ der folgende Quotient ist: $q_{1,i}(E^k) = P(E^k|H_i) \,/\, P(E^k|H_1)$ und wir den Ausdruck $Q = \sum_{i>1} P(H_i)\, q_{1,i}(E^k)$ definieren.

Hierbei ist klar, dass Q klein werden muss, damit H_1 durch die Daten gut bestätigt wird. Das ist ceteris paribus wiederum der Fall, wenn die $q_{1,i}(E^k)$ jeweils alle klein werden (was dafür spricht, dass H_1 bessere Erklärungen liefert als die anderen Hypothesen), aber in Q finden wir auch noch die Vorher-Wahrscheinlichkeiten der anderen Hypothesen, so dass der Abduktionsvertreter diesen Ausdruck nicht ohne weiteres übernehmen kann.

Weitere Gemeinsamkeiten finden sich in den möglichen Unterminierern für beide Ansätze. Wir haben schon die engen Zusammenhänge beider Ansätze zur Grundidee der eliminativen Induktion analysiert. Typische Unterminierer sind hierfür übersehene relevante Hypothesen. Das ist auch für das abduktive Schließen und ebenfalls für den Bayesianismus eine Todsünde. Für das abduktive Schließen ist das wohl offensichtlich, aber auch im Bayesianismus erkennt man das Problem leicht wieder (vgl. a. Hawthorne 1993). Wir bestimmen die Wahrscheinlichkeiten der einzelnen Hypothesen in der Regel als Liste einander ausschließender aber insgesamt vollständiger Hypothesen, und insbesondere müssen wir P(E) anhand dieser Liste bestimmen, um überhaupt einen Wert zu erhalten und um die Likelihoodanbindung einbringen zu können. An dieser Stelle lauert also eine ähnliche Fehlerquelle für beide Ansätze, die uns schnell den Aufstieg zu wissenschaftlichem Wissen verstellen kann. Der Fachwissenschaftler (und auch der Detektiv) muss in möglichst kreativer Weise immer nach neuen Konkurrenzhypothesen zu seiner eigenen Ausschau halten. Doch wir wissen wohl alle, dass diese Suche nicht gerade unsere Stärke ist, wo es uns doch eigentlich darum geht, alle Daten als positive Bestätigungen unserer Hypothesen zu betrachten. Diese Unterminierer sind spezifisch für ganz bestimmte Induktionsverfahren. Im (konservativen) Extrapolieren, der hypothetisch-deduktiven Theorienbestätigung, den Signifikanztests und vielen anderen sind vollständige Listen von Hypothesen nicht vorgesehen und können daher im Rahmen dieser Verfahren scheinbar nicht zu Fehlschlüssen führen.

Doch das zeigt nur, dass diesen Verfahren der wichtige komparative Aspekt der Bestätigung fehlt.

Fehlerhafte Daten oder Messgeräte sind natürlich Unterminierer für praktisch alle Induktionsverfahren, aber allgemeinere Hilfshypothesen sind wieder etwas spezifischere Unterminierer für alle absteigenden Bestätigungskonzeptionen. Hier finden sich klare Gemeinsamkeiten zwischen unseren Ansätzen zum induktiven Schließen.

5.12 Andere Plausibilitätsmaße

Probabilistische Positionen liefern sicher die prominentesten und am weitesten ausgearbeiteten Plausibilitätsmaße, aber wir sollten zur Kenntnis nehmen, dass sie keineswegs die einzigen Kandidaten dafür darstellen. Es gibt gerade unter Informatikern einige Versuche, den wahrscheinlichkeitstheoretischen Ansatz zu verallgemeinern etwa im Rahmen des Dempster-Shafer-Ansatzes (Shafer 1976, Huber 2011) und eine allgemeinere Theorie von Plausibilitätsmaßen zu konstruieren, die bereits einige Anwendungsbereiche befruchten kann wie etwa das Default-Schließen, bei dem wir z.B. schließen, dass aus p *typischerweise* oder *normalerweise* q folgt. Wenn Tweety ein Vogel ist, so kann er normalerweise fliegen, aber das gilt nicht in jedem Fall, so z.B. nicht für Pinguine. Ähnlich werden manchmal auch Ceteris-paribus-Gesetze verstanden. Sie behaupten demnach, dass Rauchen im Normalfall Lungenkrebs hervorruft, aber gestatten eben auch Ausnahmefälle wie Helmut Schmidt, ohne dass dadurch das CP-Gesetz falsifiziert würde.

Auf ein erstes Problem stoßen wir in der Beschreibung dieser Ansätze, wenn wir die Objekte des Glaubens bzw. für bestimmte Plausibilitätsmaße genauer bestimmen möchten. Diese Frage haben wir hier bisher einfach ausgeklammert bzw. ich habe schlicht von *Aussagen* als den Objekten des Glaubens bzw. den Objekten unserer epistemischen Bewertungen gesprochen. Wir können das auch als Frage nach der geeigneten *Individuierung* von Glaubensinhalten bzw. Aussagen verstehen. Einen Zugang dazu bietet die mögliche Welten Semantik bzw. eine entsprechende Semantik von Möglichkeiten, die etwa durch eine Menge W dargestellt werden. Man sagt dann, dass eine Proposition

p mit einer Teilmenge A von W identifiziert wird; also etwa mit der Menge A der möglichen Welten, in denen p wahr ist. Das hat allerdings gewisse problematische Konsequenzen wie z.B. die folgende: Die beiden Aussagen A1 ≡ »2+2 = 4« und A2 ≡ »Junggesellen sind unverheiratet« sind beide in allen Welten wahr und bestimmen damit dieselbe Proposition, nämlich die Menge aller Welten W. Intuitiv scheinen sie uns aber unterschiedliche Aussagen zu machen. Wir können die Aussagen aber auch nicht anhand der Sätze, mit denen wir sie aussagen, festlegen, denn die beiden Aussagen »Aprikosen sind schmackhaft« und »Marillen schmecken gut« stellen zwar recht unterschiedliche Sätze dar, während sie im Wesentlichen dieselbe Aussage machen. Während Propositionen im Sinne der möglichen Welten unsere Aussagen also wohl zu grob individuieren, stellen Sätze offensichtlich eine zu feine Individuierung dar. Wie wir dann eine geeignete Semantik und Individuierung von Aussagen beschreiben können, soll jedoch hier nicht weiter verfolgt werden (vgl. etwa Huber 2011 und King 2011).

Die im Folgenden zu betrachtenden Ansätze verwenden zumindest meist die mögliche Welten Semantik und beziehen sich daher auf die entsprechenden Propositionen, was ich hier übernehme. Man kann sich die grundlegenden Möglichkeiten w∈W auch durch die Vollkonjunktionen von gewissen Basisaussagen gegeben denken und erhält damit einen Zusammenhang zu unseren früheren Darstellungen. Oft wird jedenfalls eine Mengenalgebra A⊆Pot(W) als Menge der Propositionen angenommen, die als stabil unter Vereinigungen, Durchschnitts- und Komplementbildung konstruiert wird. Außerdem benötigen wir eine zumindest partiell geordnete Menge D für die Plausibilitätsmaße: Pl:A→D (vgl. Halpern 2001/a und 2003). Ich vereinfache das etwas und wähle für D das reelle Intervall D = [0;1]. Für das Plausibilitätsmaß Pl werden zunächst drei einfache Axiome für Pl angenommen:

1) Pl(W) = 1
2) Pl(∅) = 0
3) Für U ⊆ V gilt: Pl(U) ≤ Pl(V)

Als Nächstes zeigt Halpern, dass viele Plausibilitätsmaße diese recht schwachen Bedingungen erfüllen – zumindest, wenn wir eine recht allgemeine Menge D annehmen. Der Clou liegt dann aber vor allem

darin, dass nun konditionale Plausibilitätsmaße betrachtet werden, mit denen wir analog zum bayesianischen Updaten neue Informationen einbeziehen und unser Plausibilitätsmaß *updaten* können. Für die konditionalen Plausibilitätsmaße wird etwa für beliebige U, U* sowie V und V* aus A verlangt:

Pl:A×A→D ist ein konditionales Plausibilitätsmaß, wenn gilt:
K1) Pl(W|V) = 1
K2) Pl(∅|V) = 0
K3) Für U ⊆ U* gilt: Pl(U|V) ≤ Pl(U*|V)
K4) Pl(U|V) = Pl(U∩V|V)

Die ersten drei Axiome entsprechen denen der unbedingten Plausibilitätsmaße, während das vierte sicherstellen soll, dass die bedingte Plausibilität von U tatsächlich nur von dem Bereich der Übereinstimmung von U und V abhängt, also tatsächlich eine Relativierung auf V vorliegt. Halpern schlägt auch noch ein mögliches stärkeres Ersatzaxiom für (K4) vor:

K5) Pl(U|V∩V*) ≤ Pl(U*|V∩V*) gdw. Pl(U∩V|V*) ≤ Pl(U*∩V|V*)

Aus (K5) ist (K4) ableitbar, und (K5) sieht zwar ebenfalls einleuchtend aus, wird von einigen Ansätzen jedoch schon nicht mehr erfüllt.

Das sieht alles nach einigermaßen einfachen Verallgemeinerungen des probabilistischen Ansatzes aus. Um unsere Phantasie etwas weiter anzuregen, möchte ich zumindest einen Ansatz kurz vorstellen, der sich bei Philosophen einer gewissen Beliebtheit erfreut, aber doch ganz anders aussieht als die Wahrscheinlichkeitsansätze, nämlich die sogenannte Rangtheorie von Wolfgang Spohn (vgl. etwa Spohn 2009, 2012 und Huber 2011). Es werden dabei Propositionen in Stufen eingeteilt, wobei die plausibleren den Wert 0 zugewiesen bekommen, während die unplausibleren Propositionen natürliche Zahlen bis einschließlich unendlich erhalten. Spohn spricht von negativen Rangfunktionen, da sie mit höheren Werten höhere Grade von Unplausibilität ausdrücken. Dazu kommen wieder die typischen Anforderungen an Plausibilitätsmaße. Zugrunde liegt der ganzen Konstruktion eigentlich schon eine punktweise Funktion κ:W→ℕ∪{∞}, die einer Möglichkeit oder möglichen Welt

w∈W (die wie erwähnt durch die Vollkonjunktionen von Basisaussagen gegeben sein könnten) einen Grad von Unplausibilität κ(w) zuweist. Dabei wird gefordert, dass wir nicht alle Möglichkeiten als unplausibel ablehnen, d.h. für mindestens ein w aus W sollte gelten: κ(w) = 0. Für Mengen A∈**A**, sei dann κ(A) = min{κ(w); w∈A} definiert. Das heißt, eine Proposition ist nur so unplausibel, wie die plausibelste Welt, in der sie gilt. Man kann dazu auch direkt Rangfunktionen axiomatisch einführen:

κ:A→ℕ∪{∞} ist eine *negative Rangfunktion*, wenn für alle A, B ∈ A gilt:
(1) κ(∅) = ∞
(2) κ(W) = 0
(3) κ(A∪B) = min{κ(A), κ(B)}

Dabei erkennt man wieder die Anforderungen an Plausibilitätsmaße, und es lässt sich u.a. ableiten, dass κ(A) = 0 oder κ(¬A) = 0 gilt, wobei wir mit ¬A nun das Komplement von A in W meinen, denn die Rangfunktion liefert für die Vereinigung beider Mengen die 0, also muss eine der beiden Mengen auf 0 abgebildet werden. Außerdem ergibt sich für A ⊆ B, dass κ(B) ≤ κ(A) ist (man denke daran, dass es sich um eine *negative* Rangfunktion handelt), denn B = A∪B\A und daher κ(B) = min{κ(A), κ(B\A)} ≤ κ(A).

Wenn man die Rangfunktion im Sinne von Glaubensgraden deuten möchte, dann besteht die Menge $\kappa^{-1}(0)$, die auf die 0 abgebildet wird, aus den Aussagen, die wir akzeptieren und denen, denen gegenüber wir neutral sind. Unsere Überzeugungsmenge G ist dann dadurch gekennzeichnet, dass κ(¬A) > 0 ist: G = {A∈A; κ(¬A)>0}, während für die neutralen Überzeugungen N gilt: κ(N) = 0 und κ(¬N) = 0.

Hier lässt sich eine konditionale Rangfunktion auf besonders einfache Weise definieren:

Für eine negative Rangfunktion κ und κ(A) < ∞ lässt sich eine *konditionale negative Rangfunktion* durch κ(B|A) := κ(A∩B) − κ(A) definieren, die im ersten Argument wiederum eine negative Rangfunktion darstellt.

Die konditionalen Rangfunktionen gestatten es dann wiederum, ein Updaten zu beschreiben, wobei sich sogar ein Analogon zum Update-Verfahren von Jeffreys angeben lässt. Außerdem lassen sich einige interessante Konsequenzen ableiten, und die Rangfunktionen können

sogar dazu dienen, bestimmte Konzeptionen kontrafaktischer Kausalität zu bestimmen. Allerdings bleibt das Problem virulent, wie sich die Rangwerte zunächst in sinnvoller Weise bestimmen lassen. Es existiert leider keine Anbindung an relative Häufigkeiten, die wir wie im Falle der Wahrscheinlichkeiten nutzen könnten.

Andere Ansätze, wie etwa die Fuzzy-Logik, bieten gegenüber dem Wahrscheinlichkeitsansatz den Vorteil, dass sich einfache Kombinationsregeln festlegen lassen – also Regeln dafür, welche Plausibilität A&B zukommt, wenn wir die Plausibilität von A und die von B bereits kennen. Für Wahrscheinlichkeiten können wir nur sagen, dass $P(A\&B) \in [0; \max\{P(A),P(B)\}]$ gilt. Daher benötigten wir extrem viele Einzelwahrscheinlichkeiten. In der Fuzzy-Logik werden für die Fuzzy-Wahrheitswerte üblicherweise sogenannte t-Normen als Kombinationsregeln untersucht (vgl. Gottwald 1993, 2010).

> Eine Funktion $t:[0;1]^2 \rightarrow [0;1]$ heißt *t-Norm*, wenn sie die folgenden Bedingungen für alle a, b, c $\in [0;1]$ erfüllt:
> (T1) $t(a,1) = a$ (neutrales Element)
> (T2) $a \leq b \Rightarrow t(a,c) \leq t(b,c)$ (Monotonie)
> (T3) t ist kommutativ und assoziativ

Um die Möglichkeiten für solche t-Normen weiter einzuschränken und überschaubarer zu gestalten, wird oft zusätzlich verlangt, dass t auch noch stetig ist. Einen typischen Vertreter finden wir z.B. in $t(a,b) := \min\{a,b\}$, einen anderen in $t(a,b) := a \cdot b$. Ist also erst einmal eine Entscheidung für eine ganz bestimmte t-Norm gefallen, lassen sich die Fuzzy-Wahrheitswerte für eine Konjunktion aus denen der einzelnen Konjunkte ableiten. Leider ist zunächst nicht so klar, welche t-Norm die richtige/beste ist und außerdem muss das Ergebnis nicht immer zu unseren relativen Häufigkeiten passen.

Diese Ansätze bieten also sicher alle interessante theoretische Konstruktionen, aber in der Wissenschaftspraxis sind sie nicht sehr gut verankert, wenn es um das induktive Schließen geht, und es bleibt immer die Frage, wie wir zu konkreten Rängen für Propositionen oder zu Fuzzy-Wahrheitswerten oder zu anderen Plausibilitätsmaßen kommen bzw., wie wir sie mit unseren Daten in Verbindung bringen. Für die

Wahrscheinlichkeiten haben wir zu diesem Zweck einige Wahrscheinlichkeitskoordinierungsprinzipien wie den statistischen Syllogismus oder die Likelihoodanbindung etc. kennengelernt, die uns hier weiterhelfen. Solange diese Fragen für die anderen Plausibilitätsansätze nicht beantwortet sind, werden diese Ansätze vermutlich weiter eine Nebenrolle für das induktive Schließen spielen, wie das auch in diesem Buch der Fall ist.

6 Grundlegende Schlüsse der klassischen Statistik

Die klassische Statistik ist das Instrument, das überwiegend in der Wissenschaft zum Einsatz kommt, wenn man sich dort explizit auf bestimmte Methoden des induktiven Schließens stützt. Im Mittelpunkt des Kapitels wird der klassische Signifikanztest als Nullhypothesentest stehen, da er auch das in der Praxis der Sozialwissenschaften und der Medizin am meisten eingesetzte Werkzeug sein dürfte. Der klassische Statistiker lehnt den Einsatz *epistemischer Wahrscheinlichkeiten* ab und möchte sich nur auf *objektive Wahrscheinlichkeiten* stützen, die sich als *relative Häufigkeiten* interpretieren lassen. Dabei stehen vor allem die objektiven Likelihoods im Mittelpunkt, bei denen unsere Theorien direkte Wahrscheinlichkeiten für bestimmte Ereignisse vergeben.

Die hatten wir im letzten Kapitel im Rahmen der Likelihoodanbindung kennengelernt. Abgelehnt werden insbesondere Wahrscheinlichkeiten für Hypothesen als zu subjektive Einschätzungen. Die untersuchten Theorien können daher nicht als zu 70% bestätigt oder auf ähnliche Weise eingestuft werden. Obwohl die klassische Statistik also mit Wahrscheinlichkeiten arbeitet, lehnt sie den Probabilismus strikt ab. Man könnte etwas boshaft sagen, dass die klassische Statistik entsteht, wenn wir im Bayesianismus bis auf die Likelihoodanbindung alles andere weglassen. Doch die klassische Statistik hat natürlich eigenständige Methoden entwickelt, die in zahllosen Büchern und Artikeln vorgestellt und häufig auch sehr kritisch diskutiert werden.

Klassische Statistiker möchten sich also ganz auf die direkten Likelihoods beschränken, in denen uns die Hypothese H (ganz im Sinne der Likelihoodanbindung) direkt eine bestimmte Wahrscheinlichkeit für A vergibt. Man könnte mit Kent Staley (2014, Kap. 9) vielleicht besser P(A;H) schreiben, um diese Größe dadurch von den bedingten Wahrscheinlichkeiten der Bayesianer zu unterscheiden. Insbesondere dürfen wir natürlich nicht behaupten P(A;H) würde sich hier als P(A&H)/P(H)

definieren lassen, da der klassische Statistiker diese Größen nicht zulässt. Zahlenmäßig stimmen P(A|H) und P(A;H) nur dann überein, wenn der Bayesianer sich strikt an die Likelihoodanbindung hält. Ich bleibe hier bei der Notation P(A|H), aber wir sollten dabei im Hinterkopf behalten, dass es sich in Kapitel 6 immer nur um objektive direkte Likelihoods handelt.

Genaugenommen geht es auch meistens nicht nur um eine spezielle Hypothese H, sondern gleich um ein umfassenderes statistisches Modell einer bestimmten Situation, in dem wir z.B. noch annehmen, dass innerhalb einer bestimmten Versuchsreihe in jedem Versuch dieselbe Wahrscheinlichkeit p aufweist und die Versuche untereinander probabilistisch unabhängig sind. Wenn unsere Hypothese H z.B. besagt, dass eine bestimmte Münze fair ist (d.h. P(Kopf)=0,5), so liefert das nur dann eine Wahrscheinlichkeit für bestimmte Ergebnisse wie 55-mal Kopf in 100 Würfen, wenn wir davon ausgehen, dass diese Wahrscheinlichkeit von 0,5 nicht durch besondere Umstände des Werfens verändert wird, dass die Würfe alle dieselbe Wahrscheinlichkeit aufweisen und dass sie untereinander probabilistisch unabhängig sind. Diese weiteren Modellannahmen werden wir in der Regel aber nur stillschweigend voraussetzen.

Aus der Ablehnung des Probabilismus resultieren allerdings wiederum Interpretationsprobleme. Zunächst stellt sich die Frage, wie sich die auftretenden Likelihoods frequentistisch(?) interpretieren lassen. Was sagt uns etwa eine Aussage der Form »P(A|H) = r« über die Welt, wobei A ein spezielles Ergebnis eines konkreten Experiments beschreibt und H eine Hypothese? Zunächst liegt es nahe, das so zu übersetzen: Im Falle, dass H wahr wäre (in den möglichen H-Welten), ist bei diesem Experiment (nach unserem Dafürhalten) das Ergebnis A mit der Stärke r zu *erwarten*. Das ist jedenfalls die Behauptung, die in H für direkte Likelihoods enthalten ist. Doch damit würden wir schon wieder in die Redeweise der epistemischen Wahrscheinlichkeiten verfallen. Als Übersetzung in *relative Häufigkeiten* könnten wir vielleicht eher sagen: In den H-Welten tritt das Ergebnis A (bei einem bestimmten Experiment) im Durchschnitt mit einer *relativen Häufigkeit* r auf. Übersetzungen in relative Häufigkeiten sind aber erstens nicht ohne Verluste möglich – wie

das letzte Kapitel gezeigt hat – und sind außerdem speziell für einzelne Experimente nicht sinnvoll.

Also sollte man stattdessen wohl besser sagen: In den H-Welten hat ein bestimmtes Experiment eine *Tendenz* (Propensität) der Stärke r dazu, dass A auftritt. Wir müssen damit die enthaltenen Modalitäten schon physikalischer deuten, als in dem bloß epistemischen oberen Szenario. Ob sich die oft empiristisch gesinnten klassischen Statistiker damit wirklich anfreunden können, ist zumindest fraglich. Außerdem werden wir sehen, dass wir wenigstens an ganz bestimmten Stellen doch wieder gezwungen sind, aus relativen Häufigkeiten oder Propensitäten epistemische Schlussfolgerungen zu ziehen und dann an bestimmten Glaubensgraden nicht vorbeikommen.

Mir geht es hier – wie auch schon im Falle des Bayesianismus – nicht darum, diese Methoden in allen Einzelheiten vorzustellen und zu veranschaulichen, sondern vielmehr darum, die zugrundeliegende Logik des induktiven Schließens bzw. Rechtfertigens in der klassischen Statistik offenzulegen, und auf dieser Grundlage bestimmte Problem in ihrer Anwendung aufzuzeigen. Die führen u.a. zu der Einsicht, dass die klassischen Hypothesentests keineswegs immer das leisten können, was wir uns von ihnen in der Praxis der Wissenschaften oft erwarten. Außerdem können wir erkennen, dass es sich wieder um Spezialfälle des Schlusses auf die beste Erklärung handelt. Tatsächlich sind ganz ähnliche Abwägungsprozesse für ihre Ergebnisse in der Anwendung erforderlich.

Da es so viele gute Einführungen in die Statistik gibt und etliche davon sogar im Internet zu finden sind, verzichte ich darauf, alle Begriffe der Statistik, die ich verwende, ausführlich zu erläutern. Das würde den Rahmen dieses Buches überschreiten, und ich muss daher den Leser bitten, der damit nicht vertraut ist, dafür ein Standardwerk der Statistik zu Rate zu ziehen. Für das Verstehen der Methoden der klassischen Statistik wird sich immer wieder der Vergleich mit dem bayesianischen Ansatz als hilfreich erweisen. Insbesondere werden sich so, einige neue Sichtweisen auf die klassische Statistik ergeben und damit Stoff für eine weitere Debatte liefern. Für eine ausführliche Darstellung der klassischen Statistik gibt es also hier keinen Bedarf aufgrund der vielen guten Lehrbücher in diesem Bereich, ein direkt zugängliches Buch möchte ich trotzdem noch erwähnen: Unter anderem finden wir in Diez et al.

2012 auf »http://www.openintro.org/« ein kostenloses Textbuch zum Herunterladen mit vielen Anwendungen und weiteren Materialien. Jedes Buch in diesem Bereich hat aber Stärken und Schwächen und ich kenne leider keines, das ich nun als mein Referenzwerk nennen möchte. Zur Sprache kommen sollen in diesem Kapitel vor allem Signifikanztests im Sinne von Fisher (etwa Fisher 1935) und Schätzverfahren sowie die Regressionsrechnung, die aber erst im Kapitel 7 mit behandelt wird.

6.1 Hypothesentests im Rahmen der klassischen Statistik

In der klassischen Statistik wird den untersuchten Hypothesen also niemals eine Wahrscheinlichkeit zugewiesen, denn die hier eingesetzten Wahrscheinlichkeiten werden immer objektiv verstanden und meist als frequentistisch interpretierbar gedacht. Trotz der Probleme solcher frequentistischen Interpretationen von Wahrscheinlichkeit werden wir an dieser Stelle dem klassischen Statistiker entgegenkommen und einfach davon ausgehen, dass wir die auftretenden Wahrscheinlichkeiten in Aussagen über relative Häufigkeiten übersetzen können. Jedenfalls können Hypothesen nicht wie im Bayesianismus anhand ihrer epistemischen Wahrscheinlichkeit oder Plausibilität bewertet werden, denn die haben sie in der klassischen Statistik nicht. Sie sind entweder wahr oder falsch (bzw. gut bestätigt und akzeptiert oder nicht gut bestätigt und nicht akzeptiert) und weitere epistemische graduelle Auszeichnungen, wie gut sie genau bestätigt sind oder wie plausibel sie uns erscheinen, kommen in der klassischen Statistik streng genommen nicht vor.

Letztlich geht es wie in der klassischen Erkenntnistheorie um eine Entscheidung, wann wir bestimmte Hypothesen zumindest vorläufig akzeptieren sollten. Das muss man sich immer wieder vor Augen führen, denn manche Redeweisen der klassischen Statistiker sind in dieser Hinsicht irreführend, und es erscheint uns natürlich sehr verlockend, von mehr oder weniger gut gestützten Hypothesen zu sprechen, als ob man das genauer quantifizieren könnte. Doch die klassische Statistik ist ganz im Rahmen der klassischen Erkenntnistheorie angesiedelt, nach

der wir Hypothesen nur akzeptieren oder ablehnen können oder uns ansonsten eines Urteils einfach enthalten dürfen.

Leider sind die Regeln, wann wir bestimmte Theorien akzeptieren sollten, nicht ganz klar zu explizieren. Genügt etwa ein signifikantes Testergebnis, um eine Hypothese als akzeptierbar auszuzeichnen? Oder benötigen wir dazu mehrere Testergebnisse und wie viele sollten es denn sein? Das wird z.T. offenbleiben, wie es auch im Rahmen der klassischen Statistik nicht wirklich geklärt ist.

Im Rahmen der Nullhypothesen-Signifikanztests (NHST) wird sogar manchmal explizit gesagt, dass es nicht um eine Entscheidung zugunsten einer Hypothese geht, anders als etwa im Neyman-Pearson-Ansatz. Doch das ist m.E. irreführend, denn dann ist die Frage nicht mehr zu beantworten, was ein positives Testergebnis für unsere Überzeugungssysteme bedeutet. Man bräuchte dann zumindest eine Art von Buchhaltungssystem für die Erfolge und Misserfolge von Theorien, in dem auch eine Verrechnung vorgenommen werden kann, die uns letztlich erklärt, ab wann wir uns nun in unseren Entscheidungen (etwa Therapien) auf eine bestimmte Theorie stützen dürfen. (Genügen etwa drei bestandene Signifikanztests zum Niveau 5%?) Da es ein solches Buchhaltungssystem nicht gibt, scheint mir die einzig sinnvolle Variante des NHST (Nullhypothesen-Signifikanztestens) zu sein, dass wir die signifikant getesteten Hypothesen (signifikant entsprechend dem jeweils erforderlichen Signifikanzniveau) zunächst einmal in den Bereich der akzeptierten Hypothesen aufnehmen, solange keine weiteren widersprüchlichen Informationen zu ihnen vorliegen. Wie man solche Informationen dann damit zu verrechnen hat, ist allerdings ein weiteres offenes Problem des Ansatzes. Doch solange die nicht vorliegen, sollten positiv bestandene Signifikanztests irgendwie verbucht werden und da bleibt eben im klassischen Ansatz nur der Weg über die Aufnahme in den Akzeptanzbereich.

Diese Anbindung der Signifikanztests an bestimmte Entscheidungen ist für meine Überlegungen im weiteren nicht von großer Bedeutung. Sie erleichtert es nur, über das Signifikanztesten zu sprechen. Auch die Redeweise von den zwei Fehlertypen (erster und zweiter Art s.u.), die wir begehen, wenn wir fälschlicherweise die Nullhypothese ablehnen oder sie fälschlicherweise nicht ablehnen, passt zu meiner Konzeption einer

vorläufigen Entscheidung. In der Praxis wird auch so verfahren, auch wenn Fisher selbst das nicht propagiert hat. Er war in seinen späteren Phasen aber auch nicht mehr der Ansicht, dass der NHST überhaupt ein gutes Testverfahren für die wissenschaftliche Forschung wäre (vgl. Gigerenzer 2004).

Tatsächlich möchten wir uns zunächst auf die Theorien *stützen*, die signifikant bestätigt wurden. Ein bestandener Signifikanztest stellt eine positive epistemische Auszeichnung der Theorie dar. Hat sich ein neues Medikament als signifikant besser als ein altes erwiesen, werden wir ceteris paribus nun das neue Medikament dem alten vorziehen. Dabei geht es im klassischen Rahmen immer wieder um die Entscheidung, eine Hypothese vorläufig als wahr zu akzeptieren. Das ist die einzige Vorgehensweise, die im Rahmen der klassischen Erkenntnistheorie sinnvoll erscheint. Daher ist in den allermeisten Statistiklehrbüchern auch von einer Ablehnung der Nullhypothese (bzw. einem Beibehalten der Nullhypothese) die Rede, woraus in der Regel logisch folgt, dass die alternative Hypothese zu akzeptieren ist, da sie eine Gegenhypothese zur Nullhypothese darstellt (s.u.). Die alternative Hypothese nenne ich auch oft die »Forschungshypothese« (oder »Zielhypothese«), denn es ist die Hypothese, die wir meistens begründen möchten, wenn wir einen Signifikanztest durchführen.

Es geht mir aber vor allem darum, genauer zu ermitteln, wie *stark* die Stützung einer Hypothese durch einen positiven Signifikanztest ist, und das Thema ist natürlich von entscheidender Bedeutung für jede Art von Einsatz der Signifikanztests. Auch wenn wir nicht vom Akzeptieren einer Hypothese sprechen möchten, müssen wir doch herausfinden, wie plausibel eine Hypothese nach bestandenem Test nun ist. Schließlich geht es in vielen Fällen darum, relevantes Wissen für praktische Entscheidungen zu generieren und dazu sind wir darauf angewiesen, zumindest die Plausibilität der ratgebenden Hypothesen zu kennen. Gigerenzer (2004) hat dazu die These formuliert, dass ein Grund für die Beliebtheit der NHST der ist, dass ihre Ergebnisse sowohl von Studierenden wie auch von erfahrenen Forschern häufig falsch interpretiert werden und zwar so, als ob sie uns diese Information über die Plausibilität der Hypothesen geben könnten.

Dazu gab es bereits eine Vielzahl von empirischen Untersuchungen, die z.T. von Gigerenzer (2004) referiert werden, die gezeigt haben, dass in unserem Verständnis von Signifikanztests an irgendeiner Stelle falsche epistemische Beurteilungen oder Wahrscheinlichkeiten mit den Tests verknüpft werden. Die wurden anhand mehrerer Aussagen abgefragt, denen die Interviewten zustimmen oder nicht zustimmen konnten. Manchmal waren dabei alle vorgegeben Antworten falsch und trotzdem erhielten viele davon hohe Zustimmungsraten von über 50%. Es zeigte sich dabei die Ansicht, dass uns ein bestandener Test letztlich eine hohe Plausibilität der getesteten Hypothesen beweist und viele Wissenschaftler glauben, dass die sich mehr oder weniger direkt aus dem Signifikanzniveau oder dem p-Wert (den wir noch kennenlernen werden) ergibt. Gigerenzer konnte auch darauf verweisen, dass diese falsche Annahme eine lange Tradition in den entsprechenden Statistikeinführungen besitzt. Ohne diese falschen Annahmen müssten wir zugeben, dass der NHST uns nicht das liefert, was wir eigentlich benötigen und wäre damit weit weniger attraktiv. Meine Frage wird dazu sein, welche Stützung gibt uns ein solcher Test denn tatsächlich und wie können wir das bemessen (in einer Weise, die auch den klassischen Statistiker überzeugen sollte)?

Statt also die Wahrscheinlichkeiten der Hypothesen »upzudaten« kommen Hypothesentests in der klassischen Statistik dadurch zustande, dass die Likelihoods der Hypothesen H im Lichte der Daten E betrachtet werden, oder intuitiver ausgedrückt: Wir ziehen nur die Wahrscheinlichkeit bzw. Likelihood P(D|H) der Daten D gegeben eine Hypothese H heran, wobei nur solche Likelihoods zugelassen werden, die wir als objektiv ansehen können, die sich also aus der oben so bezeichneten Likelihoodanbindung ergeben. (Die Redeweise von den Likelihoods ist leider nicht einheitlich und stellt manchmal die Hypothesen als Träger der Likelihoods dar, was ich aber nicht übernehmen möchte.) Das heißt zusammengefasst: Zugelassen sind nur objektive (physikalische) Wahrscheinlichkeiten und objektive Likelihoods, in denen uns eine Theorie sagt, welche Wahrscheinlichkeiten bestimmte Ereignisse hätten, wenn die Theorie wahr *wäre*. Damit kommen allerdings bereits gewisse modale Elemente in die klassische Statistik, weshalb sie genau genommen nicht ganz so empiristisch akzeptabel ist, wie ihre Väter sich das gewünscht haben.

Die Grundidee ist wiederum simpel und passt zunächst auch zum bayesianischen Vorgehen oder dem hypothetisch-deduktiven bzw. falsifikationistischen Ansatz:

> **Erste Idee**: Sollte P(D|H) relativ groß sein und D dann auftreten, so stützt das die Hypothese H, sollte P(D|H) dagegen relativ klein sein und D tritt auf, dann spricht das gegen unsere Hypothese.

Die Idee erscheint zunächst plausibel (Man denke z.B. an die Hypothese einer fairen Münze und unterschiedliche Münzwurfergebnisse) und klingt noch ein wenig nach Bayesianismus oder zumindest Likelihoodismus, aber dabei stellen sich sehr schnell Probleme ein, die nur mit größerem Aufwand zu überwinden sind.

6.1.1 Der Fehlschluss der probabilistischen Falsifikation

Ein Grundproblem, das auch beim Signifikanztesten wieder auftreten wird, ist das Folgende: Wir müssen für Hypothesentests die genannte erste Idee für eine *probabilistische Falsifikation* nutzen. Das liegt u.a. daran, dass wir im klassischen Rahmen nur sehr wenig über eine direkte positive Stützung einer Hypothese durch die Daten sagen können. Ein Datum wie 53-mal Kopf bei 100 Würfen ist mit relativ vielen konkurrierenden Hypothesen vereinbar, die etwa behaupten, dass die wahr Wahrscheinlichkeit für Kopf gerade 0,5 ist oder gerade 0,53 ist oder 0,56 ist usw. Dann bestätigt es aber jede dieser Hypothesen höchstens sehr schwach. Deshalb setzt die klassische Statistik ganz auf probabilistische Falsifikationen, denn zumindest werden durch unser Datum einige Hypothese probabilistisch ausgeschlossen. Sollte nämlich die wahre Wahrscheinlichkeit 0,1 sein, wäre so ein Ergebnis praktisch ausgeschlossen und deshalb betrachten wir die Hypothese P(Kopf)=0,1 als probabilistisch falsifiziert.

Wenn also P(D|H) sehr klein ist und D trotzdem auftritt, möchten wir gerne daraus schließen, dass H falsch ist. Doch die probabilistische Falsifikation kann nicht an die Erfolge einer strikten Falsifikation anschließen. Das haben Wissenschaftler und Wissenschaftstheoretiker

immer wieder an Beispielen erläutert. Wir stellen leider fest: Auch wenn die Wahrscheinlichkeit von D im Lichte von H klein ist, spricht das Auftreten von D keineswegs gleich gegen H. Es könnte z.B. der Fall sein, dass D überhaupt nur eine kleine Wahrscheinlichkeit hat, aufzutreten.

Dass mir mein Vermieter ein Auto schenkt (D), wenn er mich mag (Hypothese H), hat nur eine sehr kleine Wahrscheinlichkeit. Wenn er mir dann tatsächlich ein Auto schenkt, spricht das aber eher dafür, dass er mich mag, als dass er mich nicht mag. Das Datum D falsifiziert die Hypothese H also in diesem Beispiel nicht, sondern stützt sie vielmehr. Das können wir mit fiktiven Wahrscheinlichkeiten verdeutlichen: So gilt womöglich $P(D|H) = 10^{-8}$, was schon sehr klein ist, aber es gilt womöglich: $P(D|\neg H) = 10^{-9}$, was noch einmal deutlich kleiner ist. Daher würde ein Autogeschenk meines Vermieters an mich, intuitiv und von den Wahrscheinlichkeiten her dafür sprechen, dass mich mein Vermieter mag. Gemäß der klassischen probabilistischen Falsifikation wäre H aber als falsifiziert zu betrachten. Die Idee ist also zu verbessern.

Einige witzige Beispiele für die Problematik des Hypothesentestens mit Hilfe einer probabilistischen Falsifikation finden wir bei Beck-Bornholdt & Dubben (1998, 210 ff.). Das erste hatte schon zu einer (etwas verwirrten) Kontroverse in *Nature* geführt (Beck-Bornholdt & Dubben 1996). Die Beispiele zeigen deutliche Unterschiede zwischen einer echten Falsifikation und einer bloß probabilistischen. Nehmen wir als Nullhypothese etwa an, dass Johannes Paul II ein Mensch ist. Weiterhin wissen wir, dass es für einen Menschen sehr unwahrscheinlich ist, der Papst zu sein, also P(Papst sein|Mensch sein) ist etwa 1 zu 6 Milliarden. Als Datum haben wir aber, dass Johannes Paul II der Papst ist. (Das Beispiel ist schon etwas älter.) Damit wäre jedoch unsere Nullhypothese widerlegt, dass es sich dabei um einen Menschen handelt, denn aufgrund der Nullhypothese hat unser Datum nur eine sehr kleine Wahrscheinlichkeit, was wir als probabilistische Falsifikation unserer Hypothese betrachten, dass Johannes Paul II ein Mensch ist. Das ist aber natürlich in diesem Fall offensichtlich unsinnig.

Ähnlich gelagert – und vielleicht etwas durchsichtiger – ist das Lottobeispiel derselben Autoren. Einen Lottohauptgewinn zu erhalten ist unwahrscheinlich, auch wenn man Lotto gespielt hat. Trotzdem spricht der Erhalt eines Hauptgewinns im Lotto sehr wohl dafür, dass der

Betreffende Lotto gespielt hat (H_0). Es kann also durchaus der Fall sein, dass $P(D|H_0)$ klein ist, aber das Auftreten von D trotzdem dafür spricht, dass H_0 wahr ist.

Nehmen wir in diesem Beispiel als Nullhypothese H_0 gerade die Vermutung: *Franz hat Lotto gespielt.* (Die Benennung als »Nullhypothese« werde ich gleich noch weiter erläutern.) Als Datum nehmen wir, *dass Franz einen Lottohauptgewinn erhalten hat.* Das sei der Gewinn, der normalerweise für sechs Richtige gezahlt wird. Nun gilt aber: P(Franz gewinnt im Lotto|H_0) = 1/14-Millionen. Also ist unsere Nullhypothese probabilistisch falsifiziert. Obwohl wir intuitiv einen Lottogewinn als gutes Indiz dafür werten würden, dass der Gewinner auch Lotto gespielt hat, sieht die Logik der probabilistischen Falsifikation das anders. Der Lottogewinn wird sogar als Indiz gewertet, dass er nicht am Spiel teilgenommen hat.

Was steckt dahinter? Nun wir haben einfach ein sehr seltenes Ereignis E beobachten können (einen Lottogewinn), aber wir kennen nur einen normalen Weg zu einem Lottogewinn, selbst wenn dieser nur sehr selten stattfindet. Trotz der Seltenheit des Ereignisses dürfen wir es in diesem Fall nicht als Falsifikationsinstanz für unsere Hypothese betrachten. Dass der Gewinner Lotto gespielt hat, ist zwar nur eine recht schwache Erklärung für seinen Gewinn, aber die richtige Frage wäre nun: Gibt es denn eine *bessere Erklärung* für den Lottogewinn, als dass der Gewinner gespielt hat? Erst wenn es die gibt, dann dürften wir so schließen wie die klassische Statistik. Doch die Alternative besteht in unserem Beispiel darin, dass jemand nicht gespielt hat und ihm irrtümlich der Hauptgewinn zugesprochen wurde. Mir sind keine definitiven Zahlen dazu bekannt, aber ich vermute, dass das noch viel unwahrscheinlicher ist als die 1 zu 14 000 000. Vermutlich hat noch nie jemand von den vielen Millionen Nicht-Lottospielen einen Hauptgewinn erhalten, die Quote ist also in jedem Fall extrem klein. Daher dürfen wir den Lottogewinn weiterhin als Indiz dafür werten, dass der Betreffende Lotto gespielt hat.

Wir sollten also für unser Verfahren der probabilistischen Falsifikation lieber den folgenden Likelihoodquotienten betrachten:

P(Franz gewinnt im Lotto|H_0) / P(Franz gewinnt im Lotto|non-H_0)

Der sollte sehr groß sein. Daher ist unsere beste Erklärung für den Lottogewinn immer noch, dass Franz gespielt hat, gegenüber der, dass er nicht gespielt hat. Das ist die Überlegung, mit der Likelihoodisten (oder auch die Bayesianer) den Fall lösen würden. Hier kommt der *komparative Aspekt* der Theorienbestätigung sehr gut zum Ausdruck. Wir haben keine »tolle« Erklärung für den Lottogewinn von Franz, denn wir können nicht sagen, warum Franz' Lottoschein gewonnen hat und nicht eine andere Zahlenkombination. Das war eben nur Zufall.

Doch unsere anderen Hypothesen dazu liefern noch deutlich schlechtere Erklärungen, weshalb das Datum klar für unsere Hypothese H_0 spricht. Wir können hier wieder einmal nur einen Vergleich zwischen der Bestätigungskraft, die ein Datum für eine Hypothese relativ zu einer anderen angibt, durchführen. Dieser Vergleich spielt im klassischen Hypothesentest allerdings deswegen keine Rolle, weil die anderen substantiellen Hypothesen typischerweise innerhalb des Experiments durch dessen Design (etwa mit Hilfe einer Zufallsstichprobe) ausgeschaltet werden.

Die klassische Falsifikation gelingt anhand des *Modus Tollens*: Gilt etwa $T \Rightarrow D$ und $\neg D$, so dürfen wir schließen auf $\neg T$. (Die erforderlichen Hilfshypothesen verkomplizierten auch diesen Schluss, aber wir lassen sie zur Vereinfachung zunächst aus dem Spiel.) Ein ähnliches Schlussverfahren – also eine Art von *probabilistischem Modus Tollens* –, wonach wir von $P(D|T)$ ist sehr groß, bzw. $P(\neg D|T)$ ist sehr klein und es gilt $\neg D$ auf $\neg T$ schließen dürften, gibt es nicht, wie Wissenschaftstheoretiker immer wieder betonen, aber auch unsere Beispiele nun eindeutig belegt haben.

Um zu einer probabilistischen Falsifikation zu gelangen, müssten wir eigentlich $P(D|H)$ mit der Wahrscheinlichkeit vergleichen, dass D bei Vorliegen anderer Hypothesen auftritt oder auftritt, wenn H falsch ist. Erst wenn hierbei $P(D|H)$ sich als sehr viel kleiner als $P(D|\neg H)$ herausstellt, haben wir einen guten Grund von einer probabilistischen Falsifikation von H auszugehen, denn erst wenn dieser *Vergleich* gegen H ausfällt, haben wir nachgewiesen, dass es gerade die Annahme H ist, die verantwortlich für die geringe Wahrscheinlichkeit von D ist. Dann erst fällt das Auftreten von D negativ auf H zurück. Anderenfalls muss H dafür keine Rolle spielen und sollte dann auch nicht durch D geschwächt werden. Daher ist mein Vorschlag dazu der folgende:

Die korrekte probabilistische Falsifikation ist komparativ:
Damit das Auftreten von D die Hypothese H probabilistisch falsifiziert, muss zumindest gelten: $P(D|H) \ll P(D|\neg H)$.

Jeder kann sich inzwischen wohl leicht weitere Beispiele für einen entsprechenden Fehler ausdenken, wenn wir nicht komparativ vorgehen. Es wird uns vielleicht nicht immer sofort gelingen, die Problematik des Beispiels ganz aufzuklären, aber es sollte uns klar sein, dass die vorgelegten Falsifikationen so nicht seriös sind.

Ähnliche probabilistische Fehlschlüsse finden wir in auch in vielen anderen Beispielen: Nehmen wir etwa einen Ziegelstein, der aus großer Höhe auf eine Betonplatte fällt und dabei etwa auf eine bestimmte Art in genau definierte 2564 Teile zerspringt (D). Aufgrund der molekularen Beschaffenheit des Ziegelsteins könnten wir ein Wahrscheinlichkeitsmodell erstellen, wonach er auch auf viele andere Weisen in viele andere Anzahlen von Teilen hätte zerspringen können. Die Wahrscheinlichkeit für D gegeben, dass *keine Magie im Spiel ist* (H), ist also sehr klein: P(D|H) ist klein. Trotzdem kämen wir nicht auf die Idee, auf das Vorliegen von Magie zu schließen, nur weil der Ziegelstein de facto auf irgendeine Art zersprungen ist.

In diesem Beispiel geht es allerdings auch noch um eine spezielle nachträgliche Berechnung der Wahrscheinlichkeit. Tatsächlich könnte das Auftreten von D eine höhere Wahrscheinlichkeit aufweisen, wenn wir die Dämonenhypothese T ins Spiel bringen: »Ein allmächtiger Dämon wollte, dass der Ziegelstein genauso zerspringt.« Dann könnte vielleicht P(D|T) tatsächlich viel größer sein als P(D|H). Also ist auch dieser Ansatz möglicherweise noch anfällig für Dämonenhypothesen (oder wir haben früher von »Klabautermanntheorien« gesprochen). Jedenfalls müssen wir in diesem Beispiel für unser Modell bestimmte Gruppierungen vornehmen, um noch zu plausiblen Schlüssen zu gelangen (s. Kap. 6.1.2).

Dass solche probabilistischen Fehlschlüsse auch reale Konsequenzen haben können, bewies leider der Fall der britischen Mutter Sally Clark, die aufgrund eines Fehlschlusses der probabilistischen Falsifikation 1999 in Großbritannien dafür verurteilt wurde, ihre beiden Kinder ermordet zu haben (s. etwa Nobles & Schiff 2005). Die beiden Kinder waren als Babys

möglicherweise den natürlichen plötzlichen Kindstod gestorben. Doch der Staatsanwalt rechnete so: Die Wahrscheinlichkeit für den plötzlichen Kindstod ist 1/8543. Außerdem nahm er an, dass das zweite Auftreten des plötzlichen Kindstods unabhängig vom ersten ist. Damit multiplizieren sich die Wahrscheinlichkeiten und es stand seiner Meinung nach 1 zu 73 Millionen, dass zweimal hintereinander ein natürlicher plötzlicher Kindstod auftritt. Da das sehr unwahrscheinlich und damit praktisch unmöglich ist, schloss er, dass die Hypothese eines natürlichen Todes nun falsifiziert sei und damit Sally Clark also eine Mörderin sein müsste.

Erst Jahre später wurde sie auf Intervention des Präsidenten der »Royal Statistical Society« Peter Green hin freigelassen. Auch Green vertrat die Ansicht, dass die genannte kleine Wahrscheinlichkeit allein keine Aussagekraft hat, sondern mit der Wahrscheinlichkeit verglichen werden müsste, dass die beiden Babys von ihrer Mutter getötet wurden. Jedenfalls ist eine einfache probabilistische Falsifikation der Hypothese, es handele sich um einen natürlichen Kindstod, wiederum nicht möglich solange die andere Wahrscheinlichkeit nicht bekannt ist.

Leider können wir im Normalfall P(D|¬H) nicht objektiv bestimmen und können damit aus Sicht der klassischen Statistik keinen entsprechenden Vergleich ziehen. Zumindest sollten wir aber versuchen, für zwei konkurrierende Hypothesen H und H* jeweils die Likelihoods P(D|H) und P(D|H*) miteinander zu vergleichen. Das wird uns als das Problem begleiten, dass Signifikanztests nicht wirklich *komparativ* sind. Die klassische Statistik muss ohne diesen Vergleich auskommen und wir werden sehen, dass das zu entsprechenden Fehlschlüssen führt.

6.1.2 Gruppieren der Daten und Teststatistik

Allerdings beginnen die Probleme schon früher. Wir müssen gewisse Gruppierungen der Daten vornehmen. Betrachten wir dazu zunächst ein einfaches Beispiel.

> Unsere Hypothese H laute zunächst, dass eine bestimmte Münze *fair* sei, d.h., es gilt: P(Kopf|H) = 1/2.

Spannende Daten für diese Münze ergeben sich aber erst bei größeren Anzahlen von Münzwürfen, nehmen wir an, wir werfen 100-mal und

unser Datum D sei eine Folge Kopf, Zahl, Zahl, Kopf, …, die sich dabei ergibt. Das erste Problem ist, dass alle einzelnen Folgen f (bzw. unsere Elementarereignisse) gleichwahrscheinlich sind, nämlich $P(\text{Folge f}|H) = (1/2)^{100} \approx 8{\cdot}10^{-31}$. Das heißt zugleich, dass jedes konkrete Einzelergebnis D (als konkrete Folge von 100 Würfen) dieselbe sehr kleine Wahrscheinlichkeit aufweist: $P(D|H) \approx 8{\cdot}10^{-31}$. Trotzdem werden manche dieser Ergebnisse die Hypothese intuitiv bestätigen und andere gegen unsere Hypothese H sprechen. Doch wie finden wir das in den Wahrscheinlichkeiten wieder? Hier müssen wir schon einen ersten Schritt vollziehen, den auch Bayesianer (und auch der Likelihoodist) mitgehen müssen. Beim Signifikanztesten betrachten wir nicht das tatsächliche Einzelereignis und seine Wahrscheinlichkeit, sondern stattdessen die Wahrscheinlichkeit dafür, dass ein Ereignis eines bestimmten Typs auftritt.

Wahrscheinlichkeitsunterschiede zwischen den Ergebnissen treten nämlich erst dann auf, wenn wir bestimmte konkrete Ergebnisse zu *Gruppen* oder *Typen von Ereignissen* zusammenfassen. Bestimmte Gruppen von Ereignissen passen besser zu unserer Hypothese – andere weniger gut. Dazu führen wir eine *Testgröße* oder *Teststatistik* oder Prüfgröße T (eine Zufallsvariable) ein, für die wir dann anhand unserer Hypothese ihre Verteilungsfunktion bestimmen. In unserem Fall werden wir als Testgröße etwa die *Summe der Köpfe* T in unserer Folge hernehmen.

Dabei ist T also eine sogenannte *Zufallsvariable*, die jeder Folge f ihre Anzahl k an enthaltenen Köpfen zuordnet. Diese Zufallsvariable ist in unserem Beispiel *binomialverteilt*, wenn wir davon ausgehen, dass es bei jedem Münzwurf immer dieselbe feste Wahrscheinlichkeit p dafür gibt, dass Kopf kommt, und die Würfe alle (kausal) *unabhängig* voneinander erfolgen. Sind die Würfe jedenfalls nach allem, was wir wissen, *kausal unabhängig*, können wir auch annehmen, sie seien *statistisch unabhängig* und dürfen die entsprechenden Wahrscheinlichkeiten für die Würfe multiplizieren. Damit erhalten wir:

Binomialverteilung zum Parameter p

$$\text{Bin}(n,k,p) = \binom{n}{k} p^k \cdot (1-p)^{n-k} \quad \text{mit} \quad \binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} \quad \text{und} \quad n! = 1 \cdot \ldots \cdot n$$

Dabei gibt »n über k« die Anzahl der Folgen in unserer Gruppe an und $p^k \cdot (1-p)^{n-k}$ gibt die Wahrscheinlichkeit einer einzelnen konkreten Folge aus der Gruppe der Würfe mit k Köpfen an. Bei einer angenommenen Binomialverteilung (etwa dem Ziehen aus einer Urne mit zwei Sorten Kugeln mit Zurücklegen) ergibt sich somit eine Wahrscheinlichkeit von Bin(n,k,p) dafür, dass wir bei n Wiederholungen des Zufallsexperiments ein Resultat erzielen mit k Kugeln der einen Sorte, wenn die Wahrscheinlichkeit in jedem Einzelfall für das Ziehen dieser Sorte Kugel gerade p ist. Dann ergibt sich für unsere Ergebnisse im Beispiel als Wahrscheinlichkeit:

$$P(T=k|H) = Bin(100,k,1/2)$$

Unsere neuen Daten E sind nun also vom Typ »T=k« (bedeutet: in unserer Folge sind k Köpfe) und damit T binomialverteilt ist, nehmen wir in unserem statistischen Modell der Situation also an, dass dieselbe Propensität p in all unseren Münzwürfen immer wieder am Werk ist. Diese Konstanz von p bzw. die Annahme, dass wir den Münzwurf immer wieder so wiederholen können, dass die relevanten Faktoren, die p beeinflussen können, i.w. konstant bleiben, die setzen wir im Sinne einer Art von transzendentalen Annahme voraus. Wenn uns das zumindest im Prinzip nicht mehr gelingen sollte, könnten wir die Experimente nämlich nicht wiederholen und verlören damit unseren Zugang zu wesentlichen Größen der Natur.

Nach unserer neuen Gruppierung mit Hilfe unserer Testgröße T fragen wir nun nicht mehr nach der Wahrscheinlichkeit, eine ganz bestimmte Folge f zu erhalten, sondern nur noch nach der Wahrscheinlichkeit, eine Folge mit einer ganz bestimmten Anzahl k an Köpfen zu erhalten. Damit teile ich den Raum aller Folgen in Gruppen oder Klassen von Folgen, die jeweils gleich viele Köpfe aufweisen. Es zählt nun nicht mehr die Wahrscheinlichkeit der konkreten Ergebnisfolge, sondern die Wahrscheinlichkeit der Gruppe, in der mein Ergebnis liegt. Dabei ist es offensichtlich, dass die Gruppen mit annähernd p% der Resultate Kopf deutlich größer sind und daher eine viel größere Wahrscheinlichkeit aufweisen als die Folgen mit ganz wenigen Köpfen. In unserem Beispiel erhalten wir so schon deutliche Unterschiede in den Likelihoods für die verschiedenen Ergebnisse:

k	0	10	20	30	40	43	45	47	50
P(T=k|H)	$8 \cdot 10^{-31}$	$1,4 \cdot 10^{-17}$	$4 \cdot 10^{-10}$	0,00002	0,01	0,03	0,05	0,07	0,08

Tabelle 6.1: Einige Werte der Binomialverteilung für p=1/2 bei 100 Würfen.

Man kann also nun schon eher erkennen, dass bestimmte Ergebnisse »T=k« viel besser zu unserer Hypothese passen als andere. Fällt mein Ergebnis in einer der größeren Klassen mit höherer Wahrscheinlichkeit, und das sind diejenigen, bei denen k in der Umgebung von n·p liegt (in unserem Beispiel also diejenigen Klassen von Folgen mit etwa 50-mal Kopf), so spricht das für die Hypothese, während ein Ergebnis weit weg von diesen noch relativ wahrscheinlichen Klassen als Datum gegen H betrachtet wird.

Dieser Übergang zu Typen entspricht übrigens dem Übergang von einzelnen *Weltzuständen* zu *Strukturbeschreibungen*, wie wir ihn bei Carnap im Rahmen der induktiven Logik kennengelernt haben. Auch dort war das Problem, dass uns Daten aus der Natur nur dann etwas für das induktive Schließen sagen können, wenn wir sie bereits in Typen von Ereignissen gruppiert haben, wonach diese Ereignisse sich in gewisser Hinsicht gleichen. Nur für die geeigneten Typen erhalten wir eine Induktionseigenschaft, die wir für das induktive Schließen benötigen.

Die Resultate oder Daten kann man so beschreiben, dass unterschiedliche Ergebnisse möglich sind:

Neue intuitive Induktionsregel
Zerlege den Ergebnisraum auf sinnvolle Weise (etwa anhand einer geeigneten Zerlegung in Typen (bzw. einer Prüfgröße) T) in Gruppen und unterscheide dann die zwei Fälle für ein beobachtetes Ergebnis x:

(1) Ergebnis fällt in wahrscheinlichere Gruppe (P(T=x|H) ist relativ groß), dann wird H bestätigt.
(2) Ergebnis fällt in unwahrscheinlichere Gruppe (P(T=x|H) ist relativ klein), dann wird H geschwächt.

Das ist so zu interpretieren, dass ein Münzwurfergebnis, das in einen größeren Bereich von konkreten Ergebnisfolgen fällt, auch intuitiv besser zu der Hypothese passt, als eines, das aus einem der Randbereiche stammt. Haben wir etwa bei 100 Münzwürfen nur 3-mal Kopf, dann

passt das kaum zu unserer Hypothese, nach der es sich um eine faire Münze bzw. einen fairen Münzwurf handelt. Man betrachte dazu die Werte der Tabelle 6.1.

Die Begründung für das Verfahren, solche Gruppen zu bilden, können wir also zunächst darin sehen, dass es unseren Intuitionen über die Bestätigung statistischer Hypothesen recht gut entspricht. Wir erwarten, bei Wahrheit von H eines der Ergebnisse mit ca. 50 Köpfen zu sehen, da die deutlich häufiger auftreten, als die aus den Randbereichen. Das sollte sich dann in unserer Bestätigungskonzeption möglichst wiederfinden. Das Beispiel belegt, wie die Zahlenverhältnisse für entsprechende Gruppen aussehen und warum oben nur von *relativ* großen Wahrscheinlichkeiten die Rede ist. Auch die größten Wahrscheinlichkeiten, die hier für eine Gruppe im Spiel sind, liegen noch unter 8%. Und die Werte würden noch viel kleiner sein, wenn wir Folgen von 1000 Würfen betrachtet hätten. Es ist daher auf diesem Wege nicht so einfach, Daten auszuwählen, die unsere Hypothese H im Sinne der Grundidee direkt deutlich stützen.

Hier droht natürlich wieder der Fehlschluss der probabilistischen Falsifikation. Um dem wenigstens teilweise zu begegnen, müssen wir von *relativ* groß bzw. klein sprechen. 50-mal Kopf (also P(T=50|H)) hat zwar auch nur eine kleine Wahrscheinlichkeit, aber natürlich spricht dieses Datum nicht dagegen, dass eine faire Münze vorliegt, sondern eindeutig dafür. Erst ein *Vergleich* der unterschiedlichen Werte zeigt, dass die Werte zwischen 40- und 60-mal Kopf deutlich größer sind als die Wahrscheinlichkeiten der Gruppen aus den Randbereichen, die nach außen hin winzig klein werden. So ist etwa $P(T=2|H) \approx 3{,}9 \cdot 10^{-27}$. Daher ist die Wahrscheinlichkeit für 50-mal Kopf doch relativ groß, jedenfalls relativ zu der für Typen von Ereignissen in den Randbereichen. Diese intuitiven Einschätzungen decken sich des Weiteren auch wieder mit unseren Likelihood-Überlegungen, die wir bereits erläutert haben. Ist die Likelihood relativ groß für eine Datum E (gegeben H) gegenüber der von einem Datum E* spricht E stärker für die Hypothese H als E*.

6.1.3 Der Zurückweisungsbereich für Hypothesen

Ein weiteres Problem ist, inwieweit H tatsächlich durch passende Ergebnisse *bestätigt* wird. Intuitiv bestätigen 52 Köpfe zwar unsere Hypothese H, aber ebenso gut werden viele konkurrierende Hypothesen bestätigt, die etwa besagen, dass die Wahrscheinlichkeit für Kopf 54% beträgt oder ähnliche Behauptungen aufstellen. Wenn bestimmte Daten jedoch sehr viele konkurrierende Hypothesen gleichzeitig bestätigen, kann die Bestätigung der einzelnen Hypothesen nicht besonders stark sein. Für eine echte Bestätigung sind die Daten letztlich zu unspezifisch. Deshalb geht man für das Hypothesentesten hier einen anderen Weg und setzt eigentlich nur auf *Falsifikationen* bzw. die eliminative Induktion in einer probabilistischen Variante.

Erhalte ich z.B. nur 2-mal Kopf, so spricht das stark gegen unsere Hypothese. Es spricht natürlich auch stark gegen viele andere Hypothesen wie etwa, dass die Wahrscheinlichkeit für Kopf 88% sei, aber das mindert in diesem Fall nicht die schwächende Wirkung für unsere ursprüngliche Hypothese H. Sollte sich also ein Ergebnis E einstellen, das bei Vorliegen von H als sehr unwahrscheinlich anzusehen ist, so betrachten wir ab einem bestimmten Punkt H als probabilistisch *falsifiziert*. Unser Datum kann dabei ohne weiteres eine Vielzahl von Hypothesen zugleich in überzeugender Weise falsifizieren. Doch ab wann genau betrachten wir H als probabilistisch falsifiziert?

Dazu wird eine weitere Gruppierung eingesetzt – Greco (2011) spricht hier vom *Abschwächen der Daten* – indem ein *Ablehnungs-* oder *Zurückweisungsbereich* Z und ein *Akzeptanzbereich* A für unsere Hypothese H und die Folge von 100 Münzwürfen festgelegt wird. In den Ablehnungsbereich Z sollten nur bestimmte Gruppen aus dem Randbereich aufgenommen werden. Dazu legen wir eine Wahrscheinlichkeitsschwelle für bestimmte Fehler fest, die wir noch akzeptieren wollen, bevor wir von einer Falsifikation sprechen. Meist wird hier 5% als sogenanntes Signifikanzniveau gewählt, was bedeutet, dass jedes Ergebnis, das in eine Gruppe G fällt, deren Wahrscheinlichkeit insgesamt kleiner ist als 5%, als Falsifikationsinstanz für H betrachtet wird. Wir wählen also A und Z so, dass gerade noch gilt: $P(E \in Z|H) \leq 5\%$.

Das ist so zu verstehen, dass wir bereit sind, unsere Hypothese H in maximal 5% der Fälle irrtümlich als probabilistisch falsifiziert einzustufen, selbst wenn sie wahr ist. Man nennt das für Nullhypothesen (s.u.) auch den *Fehler erster Art* oder α-*Fehler*. Da wir aus den Daten nicht eindeutig auf die wahren Hypothesen zurückschließen können, müssen wir ein gewisses *Irrtumsrisiko* eingehen. Das spezielle Risiko eine Nullhypothese fälschlicherweise abzulehnen, soll allerdings begrenzt werden.

In unserem Beispiel wird man den Ablehnungsbereich *zweiseitig* und relativ symmetrisch aus den beiden Randbereichen zusammensetzen. So wählen wir etwa Z = {0,1,...,39,40} ∪ {61,62,...,100} und damit A = {41,...,60}. Erhalten wir nun eine Kopfanzahl in dem Bereich Z, betrachten wir H als probabilistisch falsifiziert, da die Wahrscheinlichkeit für ein Ergebnis aus diesem Bereich insgesamt unter 5% liegt, wenn die Hypothese H wahr sein sollte.

Beispiel: Modifizieren wir das Beispiel etwas und wählen als Hypothese nun, dass unsere Münze *höchstens* eine Wahrscheinlichkeit von 0,5 für Kopf aufweist, so erhalten wir eine *einseitige* Ablehnungsmenge.

Unsere Hypothese H* laute nun, dass eine bestimmte Münze eine Wahrscheinlichkeit von höchstens 0,5 für Kopf aufweise:
$P(\text{Kopf}|H^*) \leq 1/2$.

Nun führen ganz kleine Anzahlen von Kopf nicht mehr zur Falsifikation unserer Hypothese H*, sondern nur noch recht große Werte. Damit konzentriert sich der Ablehnungsbereich ganz auf die größeren Werte, wird aber tendenziell dort auch etwas größer, da nun für diesen Bereich die gesamte *Irrtumswahrscheinlichkeit* von 5% zur Verfügung steht, während sie sich im zweiseitigen Fall auf beide Enden mit je etwa 2,5% aufgeteilt hat. Sehr groß sind die Unterschiede allerdings nicht. Es ergibt sich als Zurückweisungsbereich Z = {59,60,...,100} für unsere Hypothese H*. Die allgemeine Falsifikationsregel lautet dann jeweils:

Probabilistische Falsifikation für klassische Hypothesentests
Falls E ∈ Z, dann ist H probabilistisch falsifiziert.
Falls E ∈ A, dann ist H (sehr) schwach bestätigt.

Die Wahl von 5% ist natürlich mit einer gewissen Willkür behaftet und ein anderer Wert wäre ebenso geeignet. Früher hat man kleinere Werte versucht, aber dann zu wenige Falsifikationen in der wissenschaftlichen Praxis erhalten, weshalb man meist bei 5% bleibt. Hier wird wieder eine Verrechnung von Sicherheit und Gehalt unserer Hypothesen deutlich, die wir schon in Kapitel 2 angesprochen haben. Erhöhen wir die Sicherheit auf 1% oder weniger, erhalten wir auch weniger bestätigte Forschungshypothesen.

Die Irrtumswahrscheinlichkeit von 5% wird auch als *Signifikanzniveau* bezeichnet und gibt uns so etwas wie die Wahrscheinlichkeit an, mit der wir fälschlicherweise H verwerfen könnten, obwohl H wahr ist. Frequentistisch interpretiert können wir sagen: Selbst wenn H wahr ist, kann sich natürlich (durch Zufall) ein Ergebnis aus dem Ablehnungsbereich Z einstellen, und wird das auch immer wieder einmal tun, und zwar in ungefähr 5% aller Fälle. Das heißt: Wenn wir 1000 wahre Hypothesen diesem Testverfahren unterwerfen, werden dabei trotzdem ca. 50 probabilistisch falsifiziert. Das ist schon ein wesentlicher Unterschied zu echten deterministischen Falsifikationen, bei denen das nicht vorkommen kann, der uns immer wieder beschäftigen wird. Noch einmal: Wenn ich 100 wahre Nullhypothesen diesem Testverfahren unterwerfen werde, werden dabei ca. 5 fälschlicherweise falsifiziert. Meine Irrtumswahrscheinlichkeit für die probabilistische Falsifikation wird daher gerade durch die 5% oder auch eine andere Zahl etwa 1%, die wir ansonsten als Signifikanzniveau wählen können, angegeben.

Über die falschen Hypothesen, deren Ergebnisse trotzdem im Annahmebereich liegen – und das können durchaus recht viele sein – machen wir uns übrigens an dieser Stelle keine großen Sorgen, denn wir wollen diese Art von Bestätigung einer Theorie nicht weiter ernst nehmen, sondern setzen ganz auf die Falsifikationen. Nur wenn die fehlerhaft sind, begehen wir hier also einen ernstzunehmenden Fehler, den wir deshalb genauer beziffern und nach oben hin durch die 5% begrenzen.

Ein Problem dabei ist – was z.B. Kritiker wie Howson & Urbach (1993) einwenden –, dass die Wahl der Teststatistik nicht einfach vorgegeben ist, obwohl die hier gewählte zugegebenermaßen sehr natürlich wirkt, doch das muss in anderen Beispielen nicht ganz so einfach sein. Wir könnten jedenfalls andere Gruppen und die dazugehörigen Testgrößen

bilden, indem wir z.B. die Ergebnisse in einer Gruppe vereinigen, die 32-mal Kopf und die, die 52-mal Kopf aufweisen. Diese Gruppe hat dann eine Wahrscheinlichkeit von ca. 7,41%. Ein Wurfresultat mit 32-mal Kopf gehört somit zu dieser Gruppe und würde zu einer relativ hohen Wahrscheinlichkeit führen, doch unsere Hypothese würde intuitiv durch das Ergebnis nicht wirklich gestützt. So erhalten wir also keine geeignete Testgröße, doch was sind dann unsere genauen Kriterien für zulässige Testgrößen? Letztlich muss auch die klassische Statistik hier mit Grundintuitionen zur Bestätigung von Theorien starten und dann versuchen, diese möglichst überzeugend in statistische Modelle umzusetzen.

Max Albert (1992) hat dazu ein spezielles Argument entwickelt, warum wir uns zumindest auf die Randbereiche für die Menge Z beschränken sollten, wenn wir die Daten abschwächen und zu bestimmten Gruppierungen übergehen. Er plädiert dafür, den Ablehnungsbereich ganz im Sinne von Popper möglichst groß zu gestalten. Wenn wir H so auffassen, dass wir Z aus möglichst vielen Werten aus den Randbereichen bilden, erhalten wir so mehr potentielle Falsifikatoren für H, als wenn wir einen zentraleren Bereich gewählt hätten, und H hat somit einen größeren Gehalt, weil es mehr Werte verbietet als eine Hypothese mit einer Zurückweisungsmenge im mittleren Bereich, die nur durch bestimmte Tricks zustande gekommen ist. Wir verstehen unsere Hypothesen danach am besten so, dass sie solche »Randergebnisse« verbieten, weil sie dadurch im Sinne Poppers den größten empirischen Gehalt aufweisen.

Popper tritt zumindest für solche Hypothesen mit größtmöglichem Gehalt ein und der Schluss auf die beste Erklärung geht in dieselbe Richtung. Dadurch würden die Spielräume für die Wahl von Z zumindest deutlich verkleinert und wären in unserem Beispiel eher unbedeutend. Die 43 wäre damit jedenfalls ausgeschlossen, denn für sie können wir viele andere Werte aus den Randbereichen in Z aufnehmen, also deutlich mehr Falsifikatoren erhalten. Außerdem entspricht das Verlegen des Zurückweisungsbereichs in die Randbereiche natürlich wiederum unseren intuitiven Einschätzungen. Ergebnisse in den Randbereichen passen intuitiv offensichtlich schlechter zu unserer Nullhypothese als solche in den Innenbereichen.

Um zu Signifikanztests zu gelangen, müssen wir aber nun noch einen weiteren Schritt gehen, denn wir hatten schon in der Diskussion des Falsifikationismus darauf hingewiesen, dass Falsifikationen allein uns noch nicht viel weiter bringen. Erst mit eliminativen Induktionen gewinnen wir zugleich positive Aussagen über bestimmte Hypothesen. Das ist genau der Weg, den auch die klassische Statistik einschlägt. Es geht uns überhaupt nicht darum, unsere eigentlichen Hypothesen zu falsifizieren oder durch gescheiterte Falsifikationsversuche als bewährt zu betrachten. Da lag Popper falsch. Sondern es geht vielmehr darum, mögliche Konkurrenten unserer eigentlichen Hypothesen durch Falsifikationen auszuschalten, ganz im Sinne der eliminativen Induktion. Unsere eigentliche Forschungshypothese, die wir gerne stützen möchten, wäre in unserem Beispiel dann etwa:

H*: Die Münze ist gefälscht (d.h. P(Kopf) $\neq 1/2$).

Zu der Hypothese H* wählen wir dann als *Gegenhypothese* – auch *Nullhypothese* genannt – unsere Hypothese H, die Münze sei fair. Indem wir sie falsifizieren, haben wir damit H* bestätigt. Das ist die Vorgehensweise bei Hypothesentests im Sinne von *klassischen Signifikanztests*. Es handelt sich also um ein Verfahren der *eliminativen Induktion* mit Hilfe einer probabilistischen Falsifikation. Die Hypothese H wird nur als »Dummy-Hypothese« aufgestellt, um sie anhand der Daten zu verwerfen. Man nennt solche Hypothesen H deshalb *Nullhypothesen* (und bezeichnet sie meist mit H_0), weil sie in der Regel behaupten, dass kein wirklicher Effekt vorliegt, sondern unsere Ergebnisse noch kompatibel mit der Nullhypothese sind, dass es sich also doch um eine faire Münze handelt.

Wenn wir also z.B. 35-mal Kopf erhalten, dann schauen wir nach, ob die Hypothese H, nach der kein wirklicher Effekt (etwa einer verbogenen Münze) vorliegt, sondern die Münze fair ist und wir es tatsächlich nur mit einer Zufallsschwankung zu tun haben, dieses Ergebnis noch erklären kann. Dazu müssen wir allerdings eine Grenze ziehen, bis wohin wir noch davon sprechen möchten, dass die Nullhypothese das Ergebnis erklären kann oder bis wohin wir sie jedenfalls noch als *kompatibel* mit dem Ergebnis betrachten möchten. Die hatten wir anhand der Irrtumswahrscheinlichkeit von 5% und dem zweiseitigen symmetrischen Ablehnungsbereich gesetzt. In unserem Beispiel müssten

wir also konstatieren, dass eine Anzahl von 35-mal Kopf nicht mehr durch unsere Nullhypothese H abgedeckt würde und wir daher H verwerfen müssen. Damit wurde dann H* bestätigt.

6.2 Die Logik kontrollierter Experimente

Ein Problem ist bei dem bisherigen Verfahren allerdings, ob wir wirklich schon alle möglichen Konkurrenten von H* ausgeschaltet haben, wenn wir H verworfen haben. Das ist in unserem Beispiel nicht so klar. Zum Beispiel haben wir noch nicht die Dämonenhypothese ausgeschlossen, dass ein böswilliger Dämon (oder irgendein anderer Effekt unseres Experimentierens) genau in dem Moment unser Experimentalergebnis verfälscht hat, in dem wir unsere Münze testen wollten, eigentlich aber H wahr ist. Solche Hypothesen sollten wir als zu unwahrscheinlich von Anfang an verwerfen, sagen Sie? Das ist wohl richtig, aber damit sind wir auch wieder auf unsere erste Liste von möglichen Hypothesen im Sinne der eliminativen Induktion angewiesen und auf möglicherweise subjektive Einschätzungen, welche Effekte und kausalen Zusammenhänge überhaupt plausibel sind.

Das zeigt sich sogleich in etwas realistischeren Beispielen, in denen wir mehrere realistische Hypothesen als Antworten zu der Frage kennen, wie ein bestimmtes auffälliges Ergebnis zustande gekommen sein kann. Nehmen wir als Hypothese H etwa an, dass *die regelmäßige Einnahme eines Pausenbrots zu besseren Schulleistungen führt*. Dazu untersuchen wir dann zwei Gruppen von Schülern (A) 30 Schüler, die ein Pausenbrot essen und (B) 30 Schüler, die kein Pausenbrot essen, und erhalten als Datum (E), dass die Schulleistungen der Schüler aus der Gruppe (A) tatsächlich besser waren. Das Datum E könnte etwa lauten: Während die Schüler aus Gruppe (A) in einem entsprechenden Test im Durchschnitt 65 Punkte (von 100 möglichen) erhielten, haben die Schüler aus Gruppe (B) nur 60 Punkte im Durchschnitt erzielt.

Die Frage ist nun, was die beste Erklärung für unser Datum E ist. Eine Möglichkeit finden wir in unserer Hypothese H; eine andere findet sich in der Nullhypothese H_0, wonach die Leistungsfähigkeit der Kinder aus beiden Gruppen eigentlich gleich ist, aber heute eben durch Zufall die

Schüler aus der Gruppe (A) etwas besser abgeschnitten haben. Um diese Nullhypothese zu widerlegen, müssen wir ein probabilistisches Modell der Situation entwerfen und anhand des Modells zeigen, dass im Lichte von H_0 das Resultat sehr unwahrscheinlich ist, so dass wir H_0 damit als probabilistisch falsifiziert betrachten dürfen.

H_0 besagt hier etwa, dass die Mittelwerte μ_A und μ_B der Leistungen aus beiden Grundgesamtheiten gleich sind, bzw. ihre Differenz gleich 0 ist. Dabei denken wir uns die großen Gruppen aller Schüler mit Pausenbrot und aller ohne Pausenbrot und eine große Anzahl (fiktiver) entsprechender Tests, für die μ_A und μ_B die jeweiligen Mittelwerte sind. Statt einer solchen Häufigkeitsinterpretation könnten wir die Mittelwerte μ_A und μ_B natürlich auch als die tatsächlich durchschnittliche Leistungsfähigkeit der Schüler in den beiden Gruppen denken, ganz im Sinne einer Propensitäteninterpretation dieser Größen. Unser statistisches Modell kann also unterschiedliche Interpretationen besitzen. Jedenfalls müssen wir erläutern, welche Werte wir in unserem kleinen Experiment nun schätzen oder anderweitig ermitteln wollen. Auf die Bildung der statistischen Modelle gehen wir gleich noch einmal ein.

Doch die statistische Beurteilung ist nicht die ganze Geschichte, denn es gibt natürlich eine Vielzahl weiterer Hypothesen, die wir ebenfalls ausschließen müssen, ehe wir H als begründet ansehen dürfen. Insbesondere wollen wir mit H nicht einfach nur behaupten, dass die Schüler mit Pausenbrot aus unerfindlichen Gründen signifikant bessere Schulleistungen aufweisen, sondern, dass gerade das Pausenbrot dafür die Ursache ist. Das ist für uns die spannende Aussage, denn nur sie gestattet es, verbessernd in die Schulleistungen der Schüler einzugreifen (vgl. auch Kap. 7). Dann gibt es aber viele Konkurrenzhypothesen, die wir ebenfalls zurückweisen müssen; etwa T_1, wonach die Schüler mit Pausenbrot aus reicheren Elternhäusern stammen (das Brot wird vielleicht sogar von der Haushaltshilfe geschmiert) und in diesen Elternhäusern ebenfalls mehr Geld für Nachhilfe ausgegeben wird. Diese Nachhilfe ist dann aber womöglich die eigentliche Ursache der besseren Leistungen.

So lassen sich weitere Hypothesen $T_2,\ldots,T_n$ finden, die andere Faktoren benennen, die aus irgendwelchen Gründen mit dem Verzehr eines Pausenbrots korreliert sind und einen positiven kausalen Einfluss auf die Schulleistung haben könnten. Die müssen wir alle für unser Expe-

riment ausschließen, doch das geht nicht allein mit Hilfe der Statistik, sondern nur, indem wir ein *kontrolliertes Experiment* durchführen, in dem wir alle möglichen Störfaktoren kontrollieren und ihre Wirkung ausschalten können. Da diese Faktoren in einem solchen Experiment jeweils in derselben Form in der Versuchs- wie in der Kontrollgruppe vorkommen, können sie nicht verantwortlich sein für den Unterschied zwischen den beiden Gruppen. Sie kommen dann also als Erklärung der besseren Leistungen der Gruppe (A) nicht mehr in Frage, d.h., die betreffenden Hypothesen werden somit *eliminiert*. Zu diesem Zweck sind statistische Hypothesentests mit bestimmten Ideen für das kontrollierte Experimentieren zu kombinieren, wie sie insbesondere schon von Fisher (1951) formuliert wurden.

Ein typischer Vorschlag wäre etwa, mit einer Gruppe von 60 Schülern zu beginnen, die alle bisher kein Pausenbrot essen und daraus per *Zufallsauswahl* die Gruppen (A) und (B) von je 30 Schülern auszuwählen und den Schülern in (A) nun ein Pausenbrot zu verordnen (Stichwort: *Randomisierung*), um zu sehen, ob das auf längere Sicht einen Einfluss auf die Schulleistungen hat. Mit dieser Art von Randomisierung hofft man, die alternativen Hypothesen $T_1,\ldots,T_n$ ausschalten zu können, denn die in den Theorien genannten Faktoren sollten sich durch die Randomisierung einigermaßen gleichmäßig auf die beiden Gruppen verteilt haben, so dass ein Unterschied zwischen den Gruppen nicht auf einen dieser Faktoren zurückzuführen ist. Wir vertrauen an dieser Stelle auf die *Methode der Differenz von Mill*, nach der ein Unterschied (in den Wirkungen) in den beiden Gruppen nur durch einen Faktor verursacht werden kann, in dem sich die beiden Gruppen außerdem noch unterscheiden (vgl. Kap 7.2.3). Wie man solche Methoden zur Ermittlung von Ursachen ausbauen und systematisieren kann, wird genauer im nächsten Kapitel zu untersuchen sein.

Um sicherzustellen, dass die beiden Gruppen sich nicht bereits zu Beginn des Experiments in puncto Leistungsfähigkeit unterscheiden, könnte man natürlich auch zu Beginn und am Ende des Experiments zusätzlich Leistungstests in beiden Gruppen durchführen, um die jeweilige Entwicklung genauer zu beobachten. Doch das würde unser Experiment weiter verkomplizieren (und verteuern?) und wir verzichten daher darauf. Oder wir könnten weitere Informationen über die Verteilung der anderen

möglicherweise relevanten Faktoren für den Schulerfolg in unseren beiden Stichproben einholen, um sicherzustellen, dass sich die Gruppen (A) und (B) darin tatsächlich nicht wesentlich unterscheiden, sondern homogen sind bis auf die Gabe des Pausenbrots.

Es handelt sich dabei nach meiner Ansicht immer um Maßnahmen, die genau dazu dienen, *andere Erklärungen* als recht unwahrscheinlich erscheinen zu lassen, so dass zum Schluss nur eine Erklärung übrigbleibt, nämlich die durch Hypothese H. Die Erklärung durch unsere Nullhypothese H_0 besagt dabei etwa, dass die zu beobachtenden Unterschiede nur Zufall waren. Das ist eine Erklärung, die man natürlich für probabilistische Phänomene immer in Betracht ziehen muss. Sie soll durch den statistischen Test selbst widerlegt werden, der aufzeigt, dass bei eigentlich gleicher Leistungsfähigkeit eine solche Abweichung doch sehr unwahrscheinlich gewesen wäre, und wir die Ergebnisse daher als Indiz dafür deuten dürfen, dass ein echter Effekt vorliegt, der dann durch H erklärt wird, weil wir die anderen potentiell erklärenden Faktoren für die beobachtete Differenz alle durch das Experimentaldesign ausschließen konnten. Wir haben es hier also mit einer typischen *eliminativen Induktion* zu tun (oder einem Schluss auf die beste Erklärung), bei der allerdings eine ganze Reihe von Hypothesen auf einen Schlag (durch die Gestaltung des Experiments) widerlegt werden soll.

Die Logik von Signifikanztests innerhalb eines kontrollierten Experiments: Zu einem Datum E gibt es eine Liste von möglichen erklärenden Hypothesen H, H_0, T_1,…,T_n, wobei durch ein spezielles (randomisiertes) Experiment, in dem E als Ergebnis auftritt, die Hypothesen T_1,…,T_n als mögliche Erklärungen ausgeschieden werden sollen. Dann wird die Hypothese H_0 durch einen statistischen Test zum Signifikanzniveau 5% widerlegt, weil gilt: $E \in Z$ und $P(Z|H_0) \leq 5\%$, wobei Z den Zurückweisungsbereich zu H_0 darstellt. Gemäß der eliminativen Induktion haben wir somit H bestätigt, weil alle anderen Hypothesen eliminiert wurden.

Dabei bietet das Niveau von 5% zugleich unsere Irrtumswahrscheinlichkeit dafür an, dass wir eine an sich wahre Nullhypothese H_0 trotzdem

verwerfen. Dann ist natürlich unsere Bestätigung von H irreführend. Es wird also mit 5% Wahrscheinlichkeit eine falsche Hypothese H bei einem solchen Signifikanztest bestätigt. Dazu kommen allerdings noch die möglichen Fehler, die aus den irrtümlichen Zurückweisungen der Hypothesen $T_1,\ldots,T_n$ als Erklärungen von E resultieren. Leider können wir auch von einer Randomisierung keine Wunder erwarten (vgl. Kap. 7). Das Ausmaß dieser Fehler können wir aber nicht so leicht beziffern.

6.3 Die statistische Modellbildung am Beispiel: Der t-Test für zwei Stichproben

Wie wird in unserem Pausenbrotbeispiel nun im positiven Fall die Zufallshypothese H_0 widerlegt? Kommen wir zumindest kurz auf das statistische Modell zu sprechen, das dabei benutzt wird. Es geht hier um eine Erklärung für die Differenz zwischen zwei Gruppen. Das ist die Besonderheit unseres Experiments, aber ansonsten ist die Vorgehensweise wieder wie oben im einfachen Signifikanztest.

Als Testgröße T können wir hier z.B. die *Differenz der Mittelwerte* in unseren Stichproben wählen. Die Nullhypothese besagt in unserem Beispiel, dass die Mittelwerte der Grundgesamtheiten gleich sind und daher unsere Testgröße $T = d_A – d_B$ mit den Stichprobenmittelwerten d_A und d_B eigentlich einen Mittelwert von 0 hat und wir nur eine Zufallsabweichung in unserem Experiment beobachten. Erst wenn die Zufallsvariable T also zu stark von der Null abweicht, gilt die Nullhypothese als widerlegt (vgl. Diez et al. 2012, Kap. 5.4). Dabei wird die Mittelwertdifferenz der Stichproben noch normiert, indem man sie durch den Standardfehler der Mittelwertdifferenz teilt. Für die so resultierende Größe lässt sich dann zeigen, dass sie t-verteilt ist (mit einem bestimmten Freiheitsgrad). Mit Hilfe der t-Verteilung können wir dann einen Schwellenwert w bestimmen, der uns den Zurückweisungsbereich Z (hier einseitig) liefert: $Z = \{x; x>w\}$. Wird T also größer als w, lehnen wir die Nullhypothese ab und akzeptieren die Forschungshypothese H_1 ganz im Sinne einer einfachen eliminativen Induktion mit den zwei Hypothesen H_0 und H_1.

Die Logik einfacher Signifikanztests: Wir konstruieren zu einer *Nullhypothese* H_0 und einer dazu gegensätzlichen *Forschungshypothese* H_1 eine Testgröße T und ermitteln zum Signifikanzniveau α einen Zurückweisungsbereich Z (mit $P(T \in Z|H_0) < \alpha$). Fällt unser Testergebnis T=r dann in den Ablehnungsbereich Z, so betrachten wir die Nullhypothese als *probabilistisch falsifiziert* und die Forschungshypothese demnach als durch E bestätigt.

Das klingt alles recht einfach und naheliegend, aber wir müssen eine ganze Reihe von *Annahmen* unterschreiben, um diesen Weg so beschreiten zu können und u.a. zu einer konkreten Verteilungsfunktion für T zu gelangen. Einige der typischen Annahmen für eine solche Modellbildung wollen wir uns nun kurz ansehen. Eine Grundidee ist dabei, dass wir eine Grundgesamtheit G haben, aus der wir eine Stichprobe S ziehen, die für diese Grundgesamtheit *repräsentativ* sein soll, d.h., sie soll der Grundgesamtheit möglichst ähnlich sein. Zumindest sollte sie der Grundgesamtheit bzgl. aller *relevanten Merkmale* möglichst ähnlich sein. In unserem Beispiel, in dem es darum geht, herauszufinden, ob bestimmte Faktoren einen Einfluss auf die Schulleistung haben, sollte die Stichprobe eine möglichst ähnliche Verteilung der Faktoren aufweisen, die tatsächlich oder möglicherweise über die Schulleistung mitbestimmen. Ist etwa der Anteil der Kinder mit IQ zwischen 120 und 130 in G ca. 7%, so sollte er das auch in der Stichprobe sein etc.

Wie kann man das erreichen? Wenn wir bereits die meisten der Faktoren und ihre Verteilung in G kennen würden, könnten wir eine Stichprobe S entsprechend der Grundgesamtheit konstruieren. Da das meistens nicht der Fall ist, müssen wir uns normalerweise auf das Verfahren einer *Zufallsauswahl* aus der Stichprobe stützen. Dabei hat idealer Weise jedes Element aus G dieselbe Wahrscheinlichkeit ausgewählt zu werden. Mit dieser *Randomisierung* verbinden wir dann die Hoffnung, dass zumindest die meisten unserer Stichproben ziemlich repräsentativ für die Grundgesamtheit sind, denn die Merkmale, die häufiger in G auftreten, haben so auch eine größere Chance für S ausgewählt zu werden und umgekehrt. Im Idealfall erhalten wir so eine repräsentative Stichprobe. Da es sich allerdings um einen Zufallsprozess handelt und

eventuell viele Merkmale und entsprechende Merkmalskombinationen bei bestimmten Schülern im Spiel sind, dürfen wir das keineswegs als sicheren Weg zu repräsentativen Stichproben betrachten. Je größer die Zufallsstichproben sind, umso eher dürfen wir ceteris paribus natürlich hoffen, eine repräsentative Stichprobe zu erhalten. Sollten wir uns einer Vollerhebung nähern, sollte das deutlich sein. In unserem Beispiel sind wir von solchen Anzahlen aber weit entfernt.

Dazu kommt das Problem, dass einige der zufällig ausgewählten Schüler sich vielleicht weigern, an der Studie teilzunehmen. Tatsächlich sind etwa die Antwortquoten bei entsprechend ausgewählten Interviewpartnern in ähnlichen Studien oft nicht sehr hoch und liegen manchmal sogar unter 50%. Dann müssen Ersatzkandidaten ausgewählt werden und wir wissen nicht, ob die Verweigerer nicht eine spezielle Untergruppe mit besonderen relevanten Merkmalen darstellen, die dann in unserer Stichprobe eben nicht mehr angemessen repräsentiert werden.

Letztlich geht es uns in unserem Beispiel sogar darum, zwei Stichproben aus zwei Grundgesamtheiten G und G* auszuwählen, deren Schüler ähnliche Schulleistungen aufweisen. Dabei ist G unsere eigentliche Grundgesamtheit der Schüler, die bisher kein Pausenbrot zu sich nahmen und das beibehalten und G* ist die neu gebildete bzw. und damit noch teilweise fiktive Grundgesamtheit der Schüler, die bisher kein Pausenbrot zu sich nahmen, aber nun dazu gebracht werden, es zu tun. So wollen wir testen, welchen Einfluss in der Grundgesamtheit G die Einnahme eines Pausenbrots hätte. Das Ergebnis könnte dann eine Hilfestellung für die Schulpolitik bieten und etwa dafür sprechen, dass die Bereitstellung eines solchen Pausenbrots von Seiten der Schule etwa dort geschehen sollte, wo die Eltern das nicht leisten.

Damit zeigen sich aber sogleich praktische Probleme. Damit das Experiment sinnvoll funktionieren kann, müssen wir es über einen längeren Zeitraum – möglichst einige Jahre – laufen lassen. Wer stellt dabei jedoch sicher, dass nicht einige der Schüler (bzw. ihre Eltern) aus der Gruppe mit Pausenbrot wieder in die Gewohnheiten der früheren Jahre zurückfallen und sich klammheimlich wieder in die Gruppe ohne Pausenbrot einreihen? Hier lauern viele praktische Probleme bei der Umsetzung solcher Experimente.

Die nächste Frage ist natürlich, wer gehört genau zur Grundgesamtheit G (die auch unsere Gesamtheit G* festlegt) und haben wir einen Überblick über G, der eine solide Zufallsauswahl ermöglicht? Ein Problem ist, dass ebenso zukünftige Schüler mit vielleicht neuen Ernährungsgewohnheiten und neuen Anforderungen in der Schule zu berücksichtigen wären, doch die können wir offensichtlich noch nicht einbeziehen. Für sie führen wir aber das Experiment gerade durch. Da können wir nur hoffen, dass sie unseren heutigen Schülern hinreichend ähneln. Außerdem können wir vielleicht speziell dafür sorgen, dass bestimmte Schultypen angemessen vertreten sind. Das betrifft auch Typen wie etwa das Gymnasium, das in einer »reicheren Gegend« liegt etc.

Unser statistisches Modell trifft aber meist noch weitergehende inhaltlich Annahmen. Woher rührt etwa die Unsicherheit in unserer Bestimmung der durchschnittlichen Schulleistung in unserer Grundgesamtheit, die wir statistisch zu behandeln gedenken? Eine erste Unsicherheit finden wir in der Zufallsauswahl. Wir könnten einen besseren oder einen der schlechteren Schüler auswählen. Das tatsächlich zu beobachtende Schulleistungsergebnis der n einzelnen, gewählten Schüler wird nun jeweils durch eine Zufallsvariable $X_1,...,X_n$ beschrieben. Wie bestimmen wir dann die Schulleistung? Dafür könnten wir einfach die aktuellen Schulnoten der Schüler hernehmen oder aber wir entwerfen einen allgemeinen Schulleistungstest, in dem die Schüler zwischen 0 und 100 Punkten erhalten können. Davon wollen wir hier ausgehen. Die Zufallsvariablen X_i liefern also alle einen Wert zwischen 0 und 100. Da jeder Schüler dieselbe Chance hat, gewählt zu werden, nehmen wir einfach an, dass die Verteilung auf die 100 Punkte für alle Zufallsvariablen dieselbe ist, d.h., insbesondere sind auch ihre Erwartungswerte μ und ihre Varianzen σ^2 innerhalb einer Gruppe dieselben, nur dass wir sie noch nicht kennen. Außerdem gehen wir davon aus, dass die Zufallsvariablen X_i stochastisch unabhängig voneinander sind. Bei einer Zufallswahl beeinflusst demnach die Wahl eines Schülers nicht die der anderen. Eine wesentliche idealisierende Annahme für unsere Stichproben ist jedenfalls, dass ein grundlegendes normalverteiltes Merkmal X gemessen werden soll und all unsere Variablen $X_1,...,X_n$ *identisch* zu einem X verteilt sind.

Das stimmt in unserem Beispiel nicht ganz, da wir hier praktisch »Ziehen ohne Zurückzulegen«. Wenn wir einen besonders guten Schüler ausgewählt haben, sinken die Chancen bei der nächsten Auswahl etwas, wieder einen besonders guten Schüler zu wählen. Doch bei sehr großen Grundgesamtheiten gehen wir davon aus, dass dieser Effekt vernachlässigbar klein ist.

Durch die Wahrscheinlichkeitsverteilungen sind auch noch weitere Unsicherheiten abgedeckt, wie etwa die *Tagesform der Schüler* bei unserem Schulleistungstest und ebenfalls die *Messfehler bei der Testauswertung* (wir wissen schließlich genau, dass auch Lehrer keineswegs einheitliche Noten für dieselben Leistungen verteilen).

Für die zweite Gruppe erwarten wir Entsprechendes für die Variablen $Y_1,\ldots,Y_m$. Unser durchschnittliches Testergebnis ist dann: $d_X = 1/n \sum X_i$ und $d_Y = 1/m \sum X_j$. Das reicht aber oft noch nicht, um nun eine Verteilung für die Testgröße $T = d_X - d_Y$ zu bestimmen. Deshalb nehmen wir etwa an, dass die Verteilungen auf die Punkte im Wesentlichen einer *Normalverteilung* folgen, d.h., die X_i sind alle nach $N(\mu_A,(\sigma_A)^2)$ verteilt und die Y_i nach $N(\mu_B,(\sigma_B)^2)$ verteilt. Oft sind natürliche Größen tatsächlich zumindest annähernd normalverteilt und der zentrale Grenzwertsatz erläutert dazu, warum das so ist. Außerdem können wir unsere Arbeit noch weiter vereinfachen, indem wir davon ausgehen, dass die zwei Varianzen gleich groß sind: $(\sigma_A)^2 = (\sigma_B)^2$, was vermutlich annähernd der Fall sein wird, da beide Gruppen aus der Gruppe G hervorgehen. Unsere Nullhypothese lautet damit also H_0: $\mu_A=\mu_B$. Für dieses Modell können wir dann unsere Testgröße T einführen (wie oben schon beschrieben), in der die Mittelwerte der Stichproben d_X und d_Y miteinander verglichen werden und wir die Stichprobenstreuungen s_Y und s_X zur Normierung mit heranziehen müssen:

$$T = \frac{d_X - d_Y}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

Dieses T ist nun approximativ t-verteilt mit n+m–2 Freiheitsgraden (vgl. Genschel & Becker 2005, 247 ff.). Unsere Entscheidungsregel lautet damit:

Verwerfe H_0, wenn $T > t_{n+m-2;1-\alpha}$

Dabei ist $t_{n+m-2;1-\alpha}$ das sogenannte (1–α)-Quantil der t-Verteilung mit n+m–2 Freiheitsgraden. In unserem Fall ist 1–α gerade 95%. (Dabei lassen wir hier die Möglichkeit einfach einmal außer Acht, dass das Pausenbrot die Schulleistungen verringert, sonst müssten wir einen entsprechenden zweiseitigen Test entwickeln, in dem T auch negative Werte annehmen könnte und wenn sie zu negativ (also zu klein) würden, wäre eine entsprechende Nullhypothese ebenfalls widerlegt.) Dabei sieht die t-Verteilung ähnlich wie die Standardnormalverteilung aus und liegt also symmetrisch um die Null herum. Für größere Freiheitsgrade nähert sie sich der Standardnormalverteilung, die sogar schon ab Freiheitsgraden von 30 als gute Näherung betrachtet wird.

Um für unser *fiktives Pausenbrotbeispiel* konkret ausrechnen zu können, ob eine signifikante Abweichung der Testgröße T nach oben hin vorliegt, müssen wir also zunächst die Stichprobenstreuung s bzw. die Varianz s^2 der Stichproben ermitteln. Da wir hier keine echten Stichproben haben, dürfen wir sie hier einigermaßen plausibel zu unseren angenommenen durchschnittlichen Schulleistungsergebnissen $d_X = 65$ (mit Pausenbrot) und $d_Y = 60$ (ohne Pausenbrot) passend wählen. Nehmen wir etwa an, dass beide gleich sind: $s_X = s_Y = 10$, dann erhalten wir: $T = 5 \,/\, \sqrt{\frac{100}{30} + \frac{100}{30}} \approx 1{,}9$. Dazu sehen wir in Tabellen der t-Verteilung nach und finden $t_{58;95\%} \approx 1{,}67$ und $t_{58;99\%} \approx 2{,}39$ (zur Standardnormalverteilung gibt es nur noch kleine, aber erkennbare Abweichungen). Damit zeigt sich, dass unser Ergebnis bei einer Irrtumswahrscheinlichkeit von 5% bereits signifikant wäre, unsere Pausenbrothypothese wäre also auf diesem Niveau bestätigt worden, aber nicht mehr bei der kleineren Irrtumswahrscheinlichkeit von 1%. Der p-Wert (s.u.), also die Irrtumswahrscheinlichkeit, bei der unsere Nullhypothese noch gerade so falsifiziert würde, liegt ungefähr bei 3%. Wir sehen aber auch, wie das Ergebnis von der Streuung abhängt. Bei größerer Streuung müssen auch die Abweichungen größer werden, ehe wir ein signifikantes Ergebnis erhalten, und das klingt wiederum recht plausibel.

Der t-Test gehört vermutlich zu den in den Sozialwissenschaften am meisten eingesetzten Testverfahren. Es gibt ihn in unterschiedlichen Varianten. Die t-Verteilung ähnelt zwar der Standardnormalverteilung, ist aber – je nach Freiheitsgrad – etwas breiter verteilt. Wir haben es hier

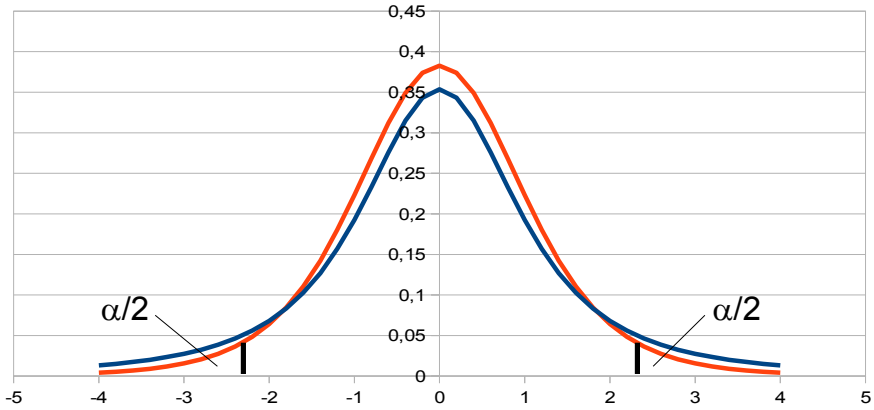
mit etwas größeren Unsicherheiten zu tun. Gegenüber einem *Gaußtest* kennen wir insbesondere die Varianzen nicht genau, sondern müssen sie anhand der Stichprobenstreuungen schätzen. Im einfachsten Fall des t-Tests haben wir nur eine Stichprobe. Wir kennen z.B. schon den bisherigen Mittelwert der Leistungsstärke μ_0 der Schüler, die kein Pausenbrot essen und möchten nur noch testen, ob unsere neue Gruppe von Pausenbrotessern bei Schulleistungstests tatsächlich besser abschneidet. Dann vereinfacht sich die t-verteilte Testgröße etwas und wir erhalten:

$$T = \sqrt{n} \cdot \frac{d_X - \mu_0}{s_X}, \text{ mit } s_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - d_X)^2}$$

als der Stichprobenstandardabweichung. Dabei ist die Testvariable T wiederum t-verteilt, aber diesmal mit n–1 Freiheitsgraden, und für unseren Signifikanztest gilt dann, dass wir die Nullhypothese verwerfen können, wenn die Größe T in den Zurückweisungsbereich fällt, also oberhalb oder unterhalb bestimmter Grenzen liegt. Je nachdem, ob dieser Test nun zweiseitig oder einseitig ist, wird unser Verwerfen der Nullhypothese zum Niveau α gerade dann stattfinden, wenn gilt:

(1) $|T| > t_{n-1;1-\alpha/2}$ (zweiseitiger Test) oder
(2) $T > t_{n-1;1-\alpha}$ (rechtsseitiger Test) oder
(3) $T < t_{n-1;\alpha} = -t_{n-1;1-\alpha}$ (linksseitiger Test)

mit $t_{n-1;1-\alpha}$ als dem kritischen Wert, der das $(1–\alpha)$-Quantil der t-Verteilung mit n–1 Freiheitsgraden angibt, also dem Wert für T, so dass gilt: $P(0 \leq T \leq t_{n-1;1-\alpha}|H_0) = 1 - \alpha$. Den Wert können wir einer entsprechenden Tabelle entnehmen oder mit einer Tabellenkalkulation ermitteln. Die t-Verteilungsdichten sind mit Hilfe der Gammafunktion zwar auch direkt beschreibbar, sind aber recht unübersichtlich und ihre Formel hilft uns kaum weiter. In der Grafik sieht man besser, wie die Verteilungsdichten aussehen und wo in den Randbereichen die Werte für die Zurückweisungsmenge zu finden sind:



Grafik 6.1: t-Verteilungsdichten für 2 und 6 Freiheitsgrade

Dabei müssen für den Test wieder die oben genannten Voraussetzungen erfüllt sein, mit deren Hilfe man die Verteilung für T ableiten kann, wonach für die X_i jeweils gilt:

(1) die X_i sind intervallskaliert
(2) jeweils identisch normalverteilt
(3) und voneinander statistisch unabhängig

Dem sogenannten »p-Wert« werden wir beim Signifikanztesten immer wieder begegnen. Er gibt zu einem Datum E an, wie groß die Wahrscheinlichkeit dafür ist, dass dieses Datum oder ein noch unwahrscheinlicheres auftritt, wenn wir von der Nullhypothese ausgehen. Für unsere Testgröße T und einen rechtsseitigen t-Test und ein Testresultat T=r, erhalten wir als p-Wert also gerade $P(T \geq r \mid H_0)$ bzw. für den zweiseitigen Test $P(|T| \geq r \mid H_0)$. Ist dieser Wert dann kleiner α, so wird die Nullhypothese abgelehnt. Das entspricht der oben genannten Regel.

Die klassische Statistik liefert uns nun für viele Situationen geeignete Testgrößen und gibt dazu die Wahrscheinlichkeitsverteilungen dieser Testgrößen an. Sie erlaubt es so, zu jedem Testresultat den p-Wert zu bestimmen und somit eine Testentscheidung zu fällen. (Allgemein lässt sich der p-Wert wie folgt bestimmen: Zunächst bestimmen wir die Menge K der Resultate, die noch unwahrscheinlicher sind als T=r mit $K = x|P(T=x|H_0) \geq P(T=r|H_0)$ und dann finden wir $p=P(T \in K|H_0)$.

Dabei wird meistens angenommen – und das sieht auf den ersten Blick auch plausibel aus –, dass die alternative Hypothese umso besser bestätigt wurde, umso kleiner der p-Wert ausfällt. Diese Behauptung finden wir schon bei Fisher (1958, 60) und immer wieder in der Statistik-Literatur (vgl. dazu Wagenmakers 2007, 787), so wird der p-Wert u.a. bei Sachs & Hedderich(2006, 323) als »ein nützliches und informatives Maß für die Evidenz einer Hypothese« bezeichnet. Wir werden diese Behauptung später aber noch kritisch diskutieren.

Das p-Wert Postulat: Je kleiner der p-Wert der Daten E relativ zu H_0 ist, umso *stärker* spricht E gegen die Nullhypothese und damit wird unsere Forschungshypothese umso *stärker bestätigt*.

Was bedeutet ein nichtsignifikantes Resultat? Was passiert, wenn der p-Wert oberhalb von 5% liegt? Dann ist unser Testresultat *nicht signifikant* und wir ziehen daraus genaugenommen keine weiteren Schlussfolgerungen. Insbesondere sagen wir nicht, dass die Nullhypothese dadurch *bestätigt* würde. Wir hatten schon gesehen, warum wir in diesem Rahmen davon Abstand nehmen, von positiven probabilistischen Bestätigungen zu sprechen. Es würden zu viele konkurrierende Hypothesen genauso mitbestätigt. Die Praktiker nehmen es damit aber nicht immer so genau und sprechen auch gerne mal davon, dass unsere Resultat doch die Nullhypothese bestätigt hätte.

Doch bleiben wir strikt: Ein nichtsignifikantes Ergebnis führt also nicht zu einer Bestätigung der Null- und damit einer Schwächung unserer Forschungshypothese. Das ist schon ein recht signifikanter Aspekt des Signifikanztestens, der Anlass zu einer kleinen Anekdote gibt: Der empirische Forscher Professor Ehrgeizig kommt zum Wissenschaftstheoretiker und bittet ihn um einen kleinen Gefallen. Er hat eine tolle neue Theorie T aufgestellt, die er gerne bestätigen würde, sich aber auf keinen Fall von irgendwelchen widerspenstigen Daten beschädigen lassen möchte. Also möchte er von dem Wissenschaftstheoretiker gerne ein Testverfahren für seine schöne Theorie kennenlernen, das sie bestätigen kann, das sie aber nicht schwächen kann, falls die Natur doch nicht wie nach T zu erwarten ist auf seine Fragen antwortet.

Da erwidert der Wissenschaftstheoretiker zunächst, dass er damit leider nicht dienen könne, denn bei jedem Test einer Theorie könnte es natürlich auch passieren, dass sich der Test gegen die Theorie richtet, denn sonst wäre es doch im strikten Sinne des Wortes kein *Test* der Theorie. Doch dann besinnt er sich und verkündet die frohe Botschaft. Genaugenommen finden wir das gesuchte Testverfahren im Signifikanztest, denn nichtsignifikante Ergebnisse sprechen schließlich nicht gegen die Forschungshypothese. Das Testverfahren kennt der Wissenschaftler zudem schon sehr gut, und es ist leicht anwendbar, gerade für einfache eher unspezifische Hypothesen. Fall gelöst!

Allerdings bringt das nun wieder den Wissenschaftstheoretiker etwas ins Grübeln. Was ist das eigentlich für ein Test, der nur einseitig für die Hypothese ausgehen kann? Den kann man dann selbst mit den einseitigen Vorgaben von Professor Ehrgeizig bedenkenlos immer wieder anwenden, um damit dieselbe Hypothese wiederholt zu testen. Irgendwann wird ein solcher Test sicher positiv ausgehen und ein signifikantes Ergebnis bringen. Die insignifikanten Ergebnisse bedeuten dagegen nichts. Das sieht nach einer recht einseitigen Bevorzugung der Forschungshypothesen und einer Vernachlässigung der Alternativen aus. Das sollte uns schon zu denken geben, ob das ein gutes Testverfahren für die Wissenschaft darstellt. Wir werden das noch genauer untersuchen müssen.

Resümee zum Beispiel t-Test. Doch was kann diese kurze Debatte eines Beispiels aus der Statistik uns bisher lehren? Denken wir an die Abgrenzung der klassischen Statistik gegenüber den Bayesianern. Einer der Haupteinwände gegen den Bayesianismus ist, dass wir dort insbesondere für die Angabe plausibler Startwahrscheinlichkeiten ganz auf die Kunst und Expertise der jeweiligen Wissenschaftler angewiesen sind, und das sei ein zu subjektives und unzuverlässiges Element in der wissenschaftlichen Bewertung von Hypothesen. Die klassische Statistik liefere dagegen ein quasi-automatisches Testverfahren. Das erscheint uns jedenfalls so, weil wir uns an die Anwendung der klassischen Statistik bereits gewöhnt haben. Doch der Schritt von einer empirischen Situation, die wir beschreiben wollen, hin zu einem vollen statistischen Modell, mit dem wir dann rechnen können, geht keineswegs automatisch vonstatten, sondern bedarf ebenso der Expertise und Kunstfertigkeit des

klassischen Statistikers. Dafür sollte die kurze Beispieldarstellung schon erste Anhaltspunkte aufzeigen.

Wir bestimmen zunächst Zufallsvariablen als Repräsentationen von Merkmalen der Situation. Dabei legen wir bereits eine Liste von Merkmalen fest, die wir für relevant für unser Ergebnis halten. Insbesondere müssen wir dann eine sinnvolle Testgröße konstruieren. Um für sie eine Verteilung zu erhalten, mit der wir erfolgreich weiterrechnen können, müssen wir darüber hinaus etliche Annahmen bzw. Idealisierungen über die Verteilungen unserer grundlegenden Zufallsvariablen vornehmen. Implizit beziehen wir uns dabei auf meist nicht scharf abgegrenzte oder sogar fiktive Grundgesamtheiten. Auch die Wahl einer Irrtumswahrscheinlichkeit und die Wahl eines dazugehörigen Ablehnungsbereichs unterliegen jeweils einer gewissen Willkür. Die Bayesianer können also zu Recht schon an dieser Stelle darauf verweisen, dass auch die Anwendung der klassischen Statistik nicht so trivial und frei von problematisierbaren Entscheidungen ist, wie es manchmal dargestellt wird.

Allgemeiner würde ich behaupten, dass alles induktive Schließen letztlich eine Kunst darstellt, die man in der Praxis erlernen muss und die alles andere als automatisiert abläuft. Dazu muss man sich immer wieder der eingesetzten (idealisierenden) Annahmen vergewissern und vielleicht sogar unterschiedliche Ansätze nebeneinander anwenden. Natürlich kann die klassische Statistik darauf verweisen, dass wir auf einige der Annahmen in unserem Beispiel letztlich wieder verzichten können und etwa mit nichtparametrischen Verfahren arbeiten könnten. Doch einige grundlegende Annahmen und Idealisierungen müssen wir in jedem Fall vornehmen, um zu einem statistischen Modell zu gelangen, mit dem wir dann konkrete Schlussfolgerungen ziehen können. Vereinfacht ausgedrückt könnten wir das Verfahren so beschreiben: Zu einer Situation mit bestimmten Merkmalen, die sie unserer Meinung nach i.w. charakterisieren, suchen wir nach einem Urnenmodell, dass die probabilistischen Anteile der Situationsbeschreibung unserer Meinung nach am besten wiedergibt. Mit Hilfe dieses Modells legen wir dann den Zurückweisungsbereich für bestimmte Hypothesen fest.

6.4 Bayesianische Kritiken am Signifikanztesten

Eine mögliche erste Kritik der Bayesianer am Signifikanztesten hatten wir schon im Dschungelfieberbeispiel kennengelernt. Selbst wenn wir zugestehen, dass ein signifikantes Datum (das ist ab jetzt immer im Sinne eines bestandenen Signifikanztests gemeint) eine *inkrementelle* Bestätigung für unsere Forschungshypothese darstellt, dann bleibt damit aber noch offen, wie stark die Bestätigung ist. Wir wissen nicht, ob sie überhaupt einen nennenswerten Zuwachs der Wahrscheinlichkeit unserer Forschungshypothese bieten würde und wir wissen nicht, in welchem Bereich die anschließend liegen würde, da wir keine Vorher-Wahrscheinlichkeiten dafür kennen. Im Dschungelfieberbeispiel fanden wir ein bemerkenswertes Testresultat, das dagegen sprach, dass der Patient noch gesund ist. Dadurch gab es auch aus bayesianischer Sicht einen großen Wahrscheinlichkeitszuwachs für die Dschungelfieberhypothese, aber trotzdem blieb sie relativ unwahrscheinlich, weil ihre Ausgangswahrscheinlichkeit so gering war. Dass die nicht berücksichtigt wurde, nennen wir den *Basisratenfehlschluss*.

Dazu kommen unsere Beispiele gegen eine probabilistische Falsifikation, die zeigten, dass das Auftreten eines unwahrscheinlichen Datums noch nicht unbedingt gegen eine Nullhypothese spricht. Wir stießen dabei auf den Fehlschluss der probabilistischen Falsifikation. Das sind alles gute Gründe, beim klassischen Statistiker nachzufragen, warum er trotzdem glaubt, dass ein bestandener Signifikanztest in relevanter Weise unsere Forschungshypothese bestätigt? Verstärkt werden diese Zweifel am Signifikanztesten durch bestimmte Anwendungen, in denen scheinbar mühelos immer wieder fragwürdige Behauptungen wie die Existenz von Psi-Phänomenen signifikant bestätigt werden. Ein Beispiel aus der letzten Zeit, das Bayesianer wieder zum Anlass für weitere Kritik am Signifikanztesten genommen haben, ist das folgende.

6.4.1 Ein Beispiel: Hellsehen

Der Psychologe Daryl Bem hat (2011) in neun Experimenten mit über 1000 Teilnehmern mögliche Phänomene von Präkognition untersucht. Dabei sollten die Probanden vorhersagen auf welcher Seite Bilder,

die kurz danach rechts oder links auf einem Bildschirm erschienen, auftauchen würden. Tatsächlich gelang den Probanden das mit einer Quote von 53,1% (p-Wert 0,011) und für die Teilmenge der erotischen Bilder sogar mit einer Trefferquote von 57,6% (p-Wert 0.00008) gegenüber der Nullhypothese, dass wir in ca. 50% der Fälle richtig treffen. Acht der neun Experimente bestätigen signifikant und z.T. sogar hochsignifikant die Forschungshypothese, dass es Fälle von Präkognition gibt.

Die Fachkollegen waren über diese Ergebnisse verständlicherweise nicht sehr erfreut und bemühten sich, in einer lebhaften Debatte darum, Bem methodische Fehler in seiner Vorgehensweise nachzuweisen (z.B. Alcott 2011). Da wurde u.a. bemängelt, dass Bem nur einseitige Tests gewählt hat oder dass er explorative Daten zur Bestätigung seiner Hypothesen eingesetzt hat oder dass er mehrfache t-Tests ohne entsprechende Korrekturen durchgeführt hätte. Bem bestritt das, und er ist ein erfahrener und renommierter Wissenschaftler, der natürlich genau weiß, wie solche Experimente und Tests korrekt durchzuführen sind. Die meisten dieser Einwände halte ich nicht für zwingend und Bems Vorgehensweise unterscheidet sich nicht wirklich von der üblichen Testpraxis in den Sozialwissenschaften oder der Medizin.

Ein interessantes Problem aus der Debatte – nämlich die Unterscheidung von explorativen und testenden Daten – möchte ich nur kurz erwähnen. Werden Daten bereits für die Aufstellung einer Hypothese verwandt, kann es in bestimmten Fällen passieren, dass sie keinen wirklichen Testfall für die Hypothese mehr darstellen und damit auch keine Bestätigung dieser Hypothese liefern können. Schurz zeigt das in (2104a), worauf wir kurz im Kapitel 3 zu sprechen kamen. Die Daten werden etwa dafür genutzt, bestimmte Parameter einer Hypothese festzulegen und können dann nicht mehr als Testfall für diese Parameter dienen, an denen die Hypothese scheitern könnte. Dann sollten wir sie auch nicht als bestätigend ansehen. Doch ein derartiger Fall liegt hier nicht vor und wir haben für das Signifikanztesten sogar schon allgemein eine ähnliche Einseitigkeit festgestellt. Es wäre daher kaum begründbar, das nun zu einem entscheidenden Einwand gegen Bems Studie zu erheben.

Andererseits sind wir nicht geneigt, nach Bems Studie auch nur vorläufig an Hellseherei zu glauben. Dazu bieten die Studie und andere

ähnliche Ergebnisse intuitiv keineswegs hinreichend überzeugende Daten und wir haben schließlich sehr gute Gründe, die dagegen sprechen. Präkognition würde bedeuten, dass zukünftige Ereignisse heutige Vorhersagen beeinflussen könnten. Das wäre ein Fall von Rückwärtskausalität mit all seinen seltsamen Konsequenzen, die bis hin zu gewissen Paradoxien reichen können. Zumindest sagt uns die bisherige Physik, dass keine Fälle von Rückwärtskausalität bekannt sind, und diese Vorstellung passt einfach nicht kohärent in unser Weltbild. Es müssten also schon sehr überzeugende Belege auf den Tisch gelegt werden, ehe wir bereit wären, wenigstens vorläufig davon auszugehen, dass es Hellseherei gibt.

Was sind dann die Konsequenzen, die wir aus diesen und ähnlichen Ergebnissen im Hinblick auf Psi-Phänomene ziehen sollten? Die Bayesianer deuten das zumindest wieder als guten Beleg dafür, dass die Signifikanztests zu schnell zu einer scheinbaren Bestätigung unserer Hypothesen führen, diese aber zumindest sehr schwach ist. Wir sollten nach ihrer Meinung eher auf bayesianische Verfahren umsteigen. Dem klassischen Statistiker sollten die vielen signifikanten Bestätigungen von Psi-Hypothesen andererseits zu denken geben und vielleicht ein Anstoß und eine Motivation sein, zumindest genauer zu bestimmen, wie stark die signifikanten Bestätigungen tatsächlich sind.

6.4.2 Eine bayesianische Hypothesenwahl

Bayesianer versuchen genau das zu quantifizieren. Wir haben aber in Kapitel 5 schon gesehen, dass es unterschiedliche bayesianische Maße für die Bestätigungstärke gibt. Für den Test zweier Hypothesen H_0 und H_1 (oder zweier Modelle M_0 und M_1) gegeneinander setzen viele Bayesianer seit Jeffreys (1961) gern auf den sogenannten *Bayes-Faktor*. (Statistiker sprechen dabei oft von »Modellen« statt »Hypothesen«, wenn sie ein bestimmtes Phänomen anhand konkreter Parameter oder konkreter Wahrscheinlichkeitsverteilungen beschreiben.)

Gerade um das Problem mit den eher subjektiven Vorher-Wahrscheinlichkeiten für die Hypothesen zu vermeiden, bietet es sich an, den folgenden Bayes-Faktor BF_{01} gegeben ein Datum E zu betrachten (vgl. a. Wagenmakers 2007, 2015):

Der Bayes-Faktor: $BF_{01} = \frac{P(E|H_0)}{P(E|H_1)}$

Der Faktor besagt zunächst, um wieviel wahrscheinlicher das Datum unter der Annahme H_0 ist gegenüber der Annahme H_1. Das Verhältnis entspricht auch dem Maß der Likelihoodisten, das etwa Royall (1997) propagiert. Doch die Likelihoodisten akzeptieren diesen Wert nur, wenn die beteiligten Likelihoods objektiven Charakter haben, was für die Bayesianer nicht unbedingt der Fall sein muss. Bayesianer müssen noch nicht einmal dieselben Werte erhalten wie die Likelihoodisten, denn die letzteren sind auf die Likelihoodanbindung angewiesen (vgl. Kapitel 5.3.11), wonach nur die Likelihoods akzeptiert werden, die direkt von unseren Hypothesen vorgegeben werden, während die Bayesianer das nicht sind. Ich habe zwar in Kapitel 5.3.11 dafür argumentiert, dass die Bayesianer ebenfalls die Likelihoodanbindung akzeptieren sollten, aber die Beispiele von Hawthorne zeigten, dass das nicht immer zu den üblichen Updateverfahren der Bayesianer passt und daher für sie nur schwer einzuhalten ist. Gelingt das nicht, unterscheiden sich Likelihoodquotient und Bayes-Faktor sogar zahlenmäßig. Das werden zumindest objektive Bayesianer natürlich tunlichst zu vermeiden suchen.

Werte des Bayes-Faktors größer als 1 sprechen dann jedenfalls für die Nullhypothese und Werte kleiner 1 in entsprechender Weise für die alternative Hypothese. Die Bedeutung des Bayes-Faktors ist zunächst schon durch seine Rolle als *Update-Faktor* für das Verhältnis der Vorher-Wahrscheinlichkeiten zu den Nachher-Wahrscheinlichkeiten definiert:

$$\frac{P(H_0|E)}{P(H_1|E)} = \frac{P(E|H_0)}{P(E|H_1)} \cdot \frac{P(H_0)}{P(H_1)} = BF_{01} \cdot \frac{P(H_0)}{P(H_1)}$$

Außerdem ergibt sich daraus eine einfache Berechnung der Nachher-Wahrscheinlichkeiten, wenn wir mit zwei gleichberechtigten Hypothesen starten, für die etwa gilt: $P(H_0) = 0{,}5 = P(H_1)$. Wenn wir noch bedenken, dass $P(H_0|E) = 1–P(H_1|E)$ gilt, ergibt sich für die Nachher-Wahrscheinlichkeit nach dem Updaten mit E:

$$P(H_0|E) = \frac{BF_{01}}{1 + BF_{01}} \text{ und } P(H_1|E) = \frac{BF_{10}}{1 + BF_{10}}$$

mit einem entsprechenden Bayes-Faktor $BF_{10} = 1/BF_{01}$. Damit erhalten wir in dieser Situation z.B. für $BF_{10} = 3$ gerade $P(H_0|E) = 0{,}75$ und für $BF_{10} = 10$ gerade $P(H_0|E) = 0{,}91$. Das ist hilfreich für die Interpretation des Bayes-Faktors.

BF_{01}	0,005	0,01	0,033	0,1	0,33	1	3	10	30	100	200
$P(H_0\|E)$	0,005	0,01	0,03	0,09	0,25	0,5	0,75	0,91	0,97	0,99	0,995

Tabelle 6.2: Die Nachher-Wahrscheinlichkeiten für die Nullhypothese für unterschiedliche Bayes-Faktoren bei $P(H_0) = 0{,}5 = P(H_1)$.

Das hilft uns zu verstehen, wie die Bewertung der verschiedenen Bayes-Faktoren im Hinblick auf ihre Bestätigungsstärke ausfällt. Die Einstufungen unterscheiden sich bei unterschiedlichen Autoren ein wenig, erfolgen aber doch weitgehend einhellig (vgl. Wagenmakers et al. 2011 oder Wagenmakers 2015 mit weiteren Literaturhinweisen):

BF_{01}	Interpretation:
1	keine Bestätigung von H_0
1–3	anekdotische (unbedeutende) Bestätigung von H_0
3–10	schwache Bestätigung von H_0
10–30	starke Bestätigung von H_0
30–100	sehr starke Bestätigung von H_0
100–200	extrem starke Bestätigung von H_0

In der anderen Richtung mit Bayes-Faktoren unter 1 finden wir entsprechende Bestätigungen der alternativen Hypothese H_1. Das ist anders als für den Signifikanztest ein echter *Vergleich* der Hypothesen, der symmetrisch ausfällt und insbesondere auch zur Bestätigung der Nullhypothese führen kann, wie sie die Tabelle zeigt.

Dabei entspricht das Vorgehen auch wieder den Vorgaben des Schlusses auf die beste Erklärung, denn ein wichtiger Parameter der Erklärungsstärke, war gerade, wie stark die Theorien die Daten vorhersagen. Wenn wir nur einfache Hypothesen miteinander vergleichen, die sich in den anderen Dimensionen der Erklärungskraft nicht wesentlich unterscheiden, steht der Likelihoodquotient bzw. der Bayes-Faktor für einen *Vergleich der Erklärungsstärke* im Rahmen des abduktiven Schließens ebenfalls ganz im Vordergrund.

Ein Problem dabei ist allerdings, wie die Likelihoods jeweils bestimmt werden können. Denken wir uns als Nullhypothese etwa die Hypothese, dass unsere Münze fair ist, d.h. die konkrete Wahrscheinlichkeit $P(\text{Kopf}|H_0) = \theta = 0{,}5$ ist. Unsere (einseitige) alternative Hypothese H_1 besage, dass die gesuchte Propensität $\theta > 0{,}5$ sei. Für die Nullhypothese können wir dann für ein Datum E (eine Anzahl von k-mal Kopf bei n Würfen) mit Hilfe der Binomialverteilung die Likelihood $P(E|H_0) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$ bestimmen. Das ist leider für die alternative Hypothese nicht so einfach, denn wir haben kein festes θ, sondern nur einen Bereich dafür zwischen 0,5 und 1. Man bezeichne den jeweiligen Bereich für den Parameter θ mit Θ. Dann benötigen wir eine *Vorher-Wahrscheinlichkeitsdichte* $p(\theta)$ auf diesem Bereich, die angibt, wie wahrscheinlich die jeweiligen Werte θ sind.

Sie gehört mit zu unsere alternativen Hypothese und soll unsere weiteren Vorkenntnisse darüber zum Ausdruck bringen, wo θ vermutlich liegen wird. Wissen wir darüber nichts, bietet sich eine möglichst informationsarme Dichte an. Im einfachsten Fall eine konstante Dichtefunktion. Für den Fall der Münze bietet sich das nicht gerade an, denn wir wissen schon, dass eine normal geworfene Münze vermutlich keine Wahrscheinlichkeit nahe 1 hat, dass Kopf kommt. Also sollte die Dichte auch größer im Bereich nahe 0,5 sein. Wie sie genau zu gestalten ist, ist allerdings meistens nicht festgelegt und hier kritisieren die klassischen Statistiker einen Spielraum, den der Bayesianer subjektiv ausfüllen muss.

Jedenfalls gehen wir dann davon aus, dass es eine Likelihoodfunktion $f(E|\theta) = P(E|\theta,H_1)$ gibt, die zu jedem θ eine Likelihood für unser Datum E festlegt. In unserem Beispiel wird das durch die oben genannte Binomialverteilung vorgegeben $f(E|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$. Nun müssen wir nur noch anhand der Dichte $p(\theta)$ über die θ mitteln (bzw. das Integral bilden) und erhalten so die gesuchte (marginale) Likelihood:

$$P(E|H_1) = \int_\Theta f(E|\theta)\, p(\theta)\, d\theta$$

Dieses Integral ist allerdings meistens nicht leicht auszurechnen und verlangt moderne Methoden sowie den Computereinsatz.

In der bayesianischen Statistik wird noch ein anderer Zusammenhang beschrieben, nämlich der zum sogenannten bayesianischen (oder

schwarzschen) Informationskriterium BIC (vgl. Held 2008, 216 ff. für die Herleitung des Zusammenhangs). Wir können nämlich auch beurteilen, wie gut eine bestimmte Hypothese oder ein Modell M zu den Daten passt, indem wir uns ansehen, wie groß die Likelihood für den Parameter $\hat{\theta} \in \Theta$ ausfällt, für den sie maximal wird. Es sei $L_1 = P(E|\hat{\theta},H_1) = f(E|\hat{\theta})$. Dann ist L_1 schon eine erste Annäherung daran, wie gut H_1 zu den Daten E passt. Sie gibt die maximale Wahrscheinlichkeit an, die nach L_1 den Daten zukommen kann. Wir könnten auch diese maximalen Likelihoods nun zu einem Hypothesen- oder Modellvergleich heranziehen und sollten die Hypothese bevorzugen, die dabei den größeren Wert aufweist.

Es gibt allerdings ein Problem dabei, auf das wir noch etwas ausführlicher in Kapitel 7 zu sprechen kommen, wenn wir das Akaike Kriterium (AIC) diskutieren, zu dem das BIC das bayesianische Pendant darstellt. Wenn wir ein Modell nämlich komplexer gestalten und mehr freie Parameter einbauen, mit deren Hilfe wir die Daten besser reproduzieren können, wird sich auch die maximale Likelihood vergrößern. Unser (möglicherweise vektorieller) Modellparameter $\hat{\theta}$ kann dann flexibler so gewählt werden, dass L_1 größer wird. Damit wird unsere Theorie aber nicht wirklich besser. Ihr Gehalt sinkt, und wir hatten schon in Kapitel 2 gesehen, dass das keineswegs ein Fortschritt sein muss. Deshalb wird für komplexere Modelle noch ein Strafterm für die Komplexität der Hypothese eingeführt, der von der Anzahl q der freien Parameter in H_1 abhängt. Das Kriterium ergibt dann:

$$BIC(H_i) = 2 \log(L_i) - q_i \cdot \log(n)$$

Dabei ist n die Anzahl der Beobachtungen und die anderen Größen sind wie oben definiert zu verstehen. Tatsächlich ergibt sich nun eine Approximation für die marginale Likelihood, wonach gilt: $P(E|H_1) \approx \exp(BIC(H_1)/2)$. Da der BIC-Wert einer Hypothese viel einfacher zu bestimmen ist als die marginale Likelihood und auch objektiver erscheint, weil keine Vorher-Dichte für θ mehr eingeht, liefert das ebenfalls eine hilfreiche Approximation für den Bayes-Faktor. Man muss für jede Hypothese jeweils nur noch ein Maximum bestimmen und dafür die Likelihood berechnen. Hinweise zu den Vorteilen und Problemen dieser Approximation finden sich etwa bei Wagenmakers (2007, 796 ff.). Doch für diese eher technischen Probleme muss ich auf die Literatur

der bayesianischen Statistik verweisen. Uns wird im Folgenden mehr interessieren, wie ein Vergleich zu den klassischen Hypothesentests ausfällt.

Von einer vorläufigen absoluten Bestätigung einer Hypothese H_1 gegenüber der Nullhypothese kann man jedenfalls erst ab Werten von 10 für den Bayes-Faktor BF_{10} sprechen. So erhalten wir ein erstes brauchbares Kriterium für die bayesianische Modellwahl und damit für unseren Vergleich mit der klassischen Statistik:

> **Bayesianische Hypothesenwahl**: Die Hypothese H_1 ist gegenüber der Nullhypothese genau dann zu wählen, wenn gilt: $BF_{10} \geq 10$.

Das gilt natürlich auch nur unter dem Vorbehalt, dass die Vorher-Wahrscheinlichkeiten noch einigermaßen nahe bei 0,5 liegen, weil nur dann damit die Nachher-Wahrscheinlichkeit der Hypothese H_1 auf über 90% steigt. Das wissen wir zwar im Normalfall nicht, aber wir werden das immer wieder als gute Ausgangsannahme betrachten. Außerdem wird sich zeigen, dass auch der klassische Statistiker auf eine ganz ähnliche Annahme angewiesen ist.

Die Anwendung auf Beispiele. Der Einsatz dieser Idee zur Analyse konkrete Beispiele aus der psychologischen Forschung bietet weitere Kritikpunkte der Bayesianer an den klassischen Signifikanztests. Zunächst ist deutlich geworden, dass der Bayes-Faktor einen symmetrischen Zugang zur Null- und zur Forschungshypothese bietet und wir damit das genannte Problem loswerden, dass nur einseitig die Nullhypothese falsifiziert werden kann, aber sich die Daten nie gegen die Forschungshypothese richten können, und wir somit eine einseitige Bevorzugung der Nullhypothese erhalten. Das ist nun kein Problem mehr. Das Verfahren der Bayesianer ist tatsächlich komparativ.

Außerdem haben Bayesianer wie Wetzels et al. (2011) das Verfahren eingesetzt, um damit 855 Beispiele von signifikanten t-Tests zu überprüfen. Sie kamen zu dem Ergebnis, dass man in etwa 70% der Fälle davon sprechen kann, dass gemäß dem bayesschen Verfahren nur eine »anekdotische Bestätigung« mit einem Bayes-Faktor von unter 3 vorliegt, die keine echte Bestätigung der Forschungshypothese darstellt.

Also wird von den klassischen Statistikern in all diesen Fällen die Bestätigungsstärke systematisch überschätzt und das bayesianische Verfahren der Modellwahl wäre demnach vorzuziehen.

Wagenmakers et al. (2011) halten das auch für eine angemessene Analyse des Bem-Beispiels. Sie haben die Daten mit Hilfe des genannten bayesianischen Ansatzes nachanalysiert und kamen zu dem Ergebnis, dass nur in drei Fällen, der Bayes-Faktor überhaupt größer als 3 war. Die beiden besten Werte waren aber auch nur 7,61 und 3,49. Ist damit die Debatte um die Hellseherei beendet? Gibt es einen klaren Punktsieg der Bayesianer gegenüber den Signifikanztestern?

Leider ist auch diese Problematik wieder einmal nicht ganz so einfach, wie die schnelle Replik von Bem et al. (2011) gezeigt hat. Bem et al. stürzen sich vor allem auf die Achillesferse der Bayesianer, nämlich die Vorherdichte für die Trefferquote θ. Die Nullhypothese besagt, dass $\theta = 50\%$ ist, während die Forschungshypothese (hier die Hellseherhypothese) H_1 behauptet, dass $\theta > 50\%$ sei. Bem argumentiert nun, dass wir für mögliche Psi-Phänomene (aber auch für andere psychologische Effekte) schon wissen, dass es sich nur um *kleine Effekte* handelt. Da Wagenmakers et al. (2011) von einer informationsarmen gleichmäßigen Vorher-Dichte für θ ausgegangen sind, die auch Werte nahe 1 mit einbezieht, haben sie den Test nach Bem et al. damit zugunsten der Nullhypothese gestaltet, denn natürlich sind bei höheren Werten von θ die Likelihoods $f(E|\theta)$ für unser Resultat E sehr klein, da die zu beobachtende Trefferquote mit ca 53% weit von den Werten nahe 1 entfernt ist.

Doch das sieht sogleich anders aus, wenn wir unser Vorwissen in entsprechende *informierte Vorher-Dichten* einfließen lassen und damit unsere Hypothese H_1 realistischer gestaltet ist. Dann erhalten die größeren θ und damit die ganz kleinen Likelihoodwerte ein kleineres Gewicht in der Durchschnittsbildung (unserem Integral) und die marginale Likelihood steigt deutlich an.

Bems Vorwurf an die bayesianischen Kritiker ist also, dass sie die Nullhypothese durch eine unrealistische Ausgestaltung der Forschungshypothese begünstigt haben, wir aber, sobald wir aber zu einer realistischeren Modellierung übergehen, zu ähnlichen Ergebnissen kommen wie bei den klassischen Signifikanztests.

Ein weiteres Problem für den bayesianischen Vergleich besteht darin, dass eine Punkthypothese ($\theta = 0$) mit einer Intervallhypothese ($\theta > 0$) verglichen wird. Um der Problematik zu entgehen, werde ich später nur Szenarien miteinander vergleichen, in denen ausschließlich Punkthypothesen darum konkurrieren, die Daten zu erklären. Tatsächlich ergeben die Werte für die Bayes-Faktoren der realistischeren Hypothesen wiederum erste Bestätigungen der Forschungshypothese: Wir erhalten (nach Bem et al. 2011):

Experiment	p-Wert	BF (Bem)	BF (Wagenmakers)
1	0,014	4,94	1,64
2	0,018	3,45	1,05
3	0,014	5,35	1,82
4	0,028	1,76	0,58
5	0,028	2,74	0,88
6	0,018	3,78	1,10
7	0,19	0,50	0,13
8	0,058	1,62	0,47
9	0,004	10,12	5,88

Tabelle 6.3: Unterschiedliche Auswertungen von Bems 9 Experimenten.

Bem et al. berechnen nun die p-Werte für zweiseitige t-Tests (statt wie ursprünglich für einseitige), so dass nur noch 7 der 9 Ergebnisse signifikant sind. Dazu berechnen sie noch den BF-Wert für das kombinierte Gesamtexperiment. Ein großer Vorteil des bayesianischen Ansatzes ist es, dass wir hier klare Regeln für das mehrfache Updaten mit neuen Daten haben. Als Bayes-Faktor ergibt sich dabei: 13,7. Außerdem rechnet Bem noch aus, wie hoch die Wahrscheinlichkeit dafür ist, dass in allen Fällen die Nullhypothese richtig ist, wenn wir dafür die Vorher-Wahrscheinlichkeit 0,5 vergeben. Es ergibt sich: $7{,}3 \cdot 10^{-5}$.

Also stützen Bems Ergebnisse auch bei einer bayesianischen Auswertung die Hellseherhypothese. Damit müssen wir wohl leben. Das induktive Schließen ist nicht irrtumssicher. Was können wir in diesem konkreten Beispiel daraus schließen? Für einen Bayesianer gilt es zunächst abzuwägen, wie hoch die Vorher-Wahrscheinlichkeit der Forschungshypothese ist. Wagenmakers et al. (2011, 428) geben sie probehalber mit 10^{-20} an, wodurch sie also selbst nach dem Updaten

mit Bems Daten immer noch sehr klein bliebe. Man denke aber auch an die kleine Anekdote in Kapitel 5 zur Handynutzung und Hirnerweichung. Wähle ich die Vorher-Wahrscheinlichkeit nur klein genug, haben wir mit realen Daten keine Chance mehr, die Hypothese jemals ernsthaft zu erwägen. Diese Gefahr einer Immunisierung unserer Vorannahmen in der anderen Richtung darf man nicht außer Acht lassen.

Auf jeden Fall läge dann die Lösung für das Bem-Beispiel nicht in den bayesianischen Auswertungsverfahren, sondern in unserer Vorher-Plausibilitätseinschätzung. Das könnte auch der klassische Statistiker aufgreifen und verlangen, dass wir für das vorläufige Akzeptieren besonders unplausibler Hypothesen besondere signifikante Ergebnisse mit sehr kleinen p-Werten benötigen.

Einen anderen möglichen Lösungsweg bieten *Replikationsstudien*. Die werden jedoch oft nicht durchgeführt, weil viele Zeitschriften sie nicht abdrucken, denn die hätten keinen hinreichenden Neuigkeitswert. Im Falle der Experimente von Bem sind die Experimente aber wiederholt und auch veröffentlicht worden. Galak et al. (2012) konnten Bems positive Ergebnisse für die Hellseherei nicht replizieren.

Aber auch Replikationsstudien sind natürlich kein Allheilmittel. Erstens, weil viele Studien nicht wiederholt werden oder auch nicht sinnvoll wiederholbar sind. Wer möchte etwa noch einmal ein Krebsmedikament im großen Stil vergleichsweise einsetzen und dabei gegen ein anderes testen, wenn sich das in einem früheren Test bereits als das schlechtere erwiesen hat? Zweitens handelt es sich bei den beiden Tests dann nur um ein größeres Gesamtexperiment und das kann natürlich wieder irreführende Ergebnisse haben. Die Quote für den Fehler erster Art bleibt mit 5% konstant und wird nicht etwa kleiner durch größere Fallzahlen. Wir werden außerdem noch sehen, dass größere Fallzahlen ihre eigenen Probleme haben.

In jedem Fall sind wir jedoch darauf angewiesen, die *Stärke der Bestätigung* durch unseren Test abschätzen zu können. Und natürlich sollten wir tatsächlich als klassische Statistiker für unplausible Forschungshypothesen stärkere Belege verlangen, bevor wir sie zumindest vorläufig akzeptieren. Wie sich die Stärke der Belege im Falle der klassischen Statistik bestimmen lässt, werden wir gleich diskutieren.

Davor möchte ich aber noch auf die bayesianischen Standardeinwände gegen die klassischen Signifikanztests eingehen.

6.4.3 Weitere problematische Aspekte des Signifikanztestens aus Sicht der Bayesianer

Zunächst wendet der Bayesianer gegen das Signifikanztesten ein, dass es auf einer probabilistischen Falsifikation beruht, die nicht wirklich komparativ ist. Daher findet sich im Nullhypothesentesten immer noch der oben angesprochene Fehlschluss wieder. Es gibt aber auch eine Reihe weiterer konkreter Einwände der Bayesianer, die von Wagenmakers (2007) besonders übersichtlich und verständlich zusammengefasst und in drei Gruppen eingeteilt werden. Darauf möchte ich mich im Folgenden stützen.

Bloß hypothetische Daten sollten nicht über die Bestätigung einer Hypothese entscheiden. Der erste Einwand bezieht sich darauf, dass der Signifikanztester nicht einfach die Likelihood der Daten heranzieht, um die Nullhypothese zu beurteilen, sondern erste eine Gruppierung vornimmt (die Abschwächung der Daten) und deren Likelihood entscheidet. Daher können dieselben Daten einmal signifikant und einmal nicht-signifikant sein, je nachdem, welche Daten für eine Gruppierung zur Verfügung stehen. So entscheiden hypothetische Daten, die nicht wirklich auftreten, über die Bewertung einer Hypothese. Dazu gibt es einfache Beispiele aus der Literatur (s. Wagenmakers 2007), die diese Problematik verdeutlichen sollen.

Nehmen wir etwa folgende einfache Situation an: Eine Variable x nehme einen der folgenden Werte an: {1,...,6}. Unsere Teststatistik sei t(x) = x. Dazu betrachten wir zwei Nullhypothesen, die die folgenden Wahrscheinlichkeiten f(x) bzw. g(x) für die 6 möglichen Werte vergeben (s. Tabelle 6.4). Der Test sei einseitig angelegt hin zu den größeren Werten. Unser Beobachtungsergebnis sei 5.

Verteilung	x=1	x=2	x=3	x=4	x=5	x=6
f(x)	0,04	0,30	0,31	0,31	0,03	0,01
g(x)	0,04	0,30	0,30	0,30	0,03	0,03

Tabelle 6.4: Zwei Wahrscheinlichkeitsverteilungen aufgrund zweier Nullhypothesen für unsere 6 möglichen Ergebnisse.

In beiden Situationen erhält unser Testergebnis (x=5) gerade die Wahrscheinlichkeit 3%. Doch nur im ersten Fall ist es signifikant, weil es nach der Gruppierung mit den weiter außenliegenden unwahrscheinlicheren Daten auf einen p-Wert von 4% kommt, während wir im zweiten Fall bei der Gruppierung auf einen p-Wert von 6% kommen. Die erste Nullhypothese wird also probabilistisch falsifiziert, die zweite dagegen nicht.

Das erscheint den Bayesianern als fehlerhaft, denn das Testergebnis erhält in beiden Fällen dieselbe Wahrscheinlichkeit und die Hypothesen unterscheiden sich i.w. nur darin, welche Wahrscheinlichkeit ein weiteres hypothetisches Datum – nämlich x=6 – von den Nullhypothesen zugewiesen wird. Da das Datum aber de facto nicht auftritt, bleibt es unverständlich, wieso das tatsächliche Datum so unterschiedlich beurteilt wird. Warum sollte die unterschiedliche Beurteilung von x=6, wenn das Ergebnis überhaupt nicht auftritt, zu unterschiedlichen Einschätzung des tatsächlich auftretenden Datums im Hinblick auf die beiden Hypothesen führen? Wie hoch die Wahrscheinlichkeit von x=6 ist, scheint doch für unseren Fall eigentlich irrelevant zu sein. Wenn unser Datum also im ersten Fall gegen die Hypothese spricht, sollte uns das auch im zweiten Fall überzeugen, dass das Datum so schlecht zur Hypothese passt, dass sie aufzugeben ist.

Das sieht der klassische Statistiker anders. Für ihn entscheidet tatsächlich die Wahrscheinlichkeit des nicht aufgetretenen Ergebnisses x=6 den Tag. Er hat uns allerdings zu erklären, wieso wir hier nicht einfach die tatsächlichen Daten mit der Hypothese vergleichen, sondern dabei auch solche hypothetischen Daten einzubeziehen sind. Das erscheint uns zunächst nicht wirklich sinnvoll. Die Herausforderung an den klassischen Statistiker ist nun jedenfalls klar erkennbar.

Bayesianer stützen sich nicht nur auf diese einfachen intuitiven Zusammenhänge, sondern bemühen hier auch noch das sogenannte

Likelihoodprinzip. Das Prinzip besagt, dass alles, was uns ein Datum E über eine Hypothese H im Hinblick auf deren epistemische Stützung sagen kann, in der Likelihood P(E|H) enthalten ist. Das ist gerade der Update-Faktor, den wir in Kapitel 5 für das bayesianische Updaten kennengelernt hatten. Auch für den Schluss auf die beste Erklärung spielte er eine wichtige Rolle. Wenn wir an probabilistische Hypothesen denken, scheint die Likelihood intuitiv ein gutes Maß für die Erklärungsstärke abzugeben.

Aber wir hatten in Kapitel 4.4.5 schon ein problematisches Beispiel von Fitelson (2007) gegen das Likelihoodprinzip kennengelernt. Die Situation war die folgende: Wir ziehen eine beliebige Karte aus einem gut gemischten Kartenspiel und haben dazu zwei Hypothesen: (H_1) *Die Karte ist ein Pikass* und (H2) *Die Karte ist schwarz* und unser Datum (E) besagt: *Die Karte ist ein Pik*. Dann gilt: $P(E|H_1)=1$ und $P(E|H2)=1/2$. Demnach würde H_1 durch E stärker gestützt als H2, aber andererseits folgt H2 sogar logisch aus E. Das passt nicht zusammen. Das Likelihoodprinzip kann also auch nicht als unbestreitbares Prinzip unseren Fall entscheiden.

Im Normalfall und in den meisten Anwendungen hat es jedoch eine große Plausibilität auf seiner Seite und sollte daher nicht vorschnell beiseite gelegt werden. Außerdem ist der Einwand der Bayesianer auch ohne das Likelihoodprinzip nachvollziehbar, und Wagenmakers nennt noch weitere instruktive Beispiele aus der Debatte der Bayesianer, die den Punkt der hypothetischen Daten weiter illustrieren.

Allerdings kann der klassische Statistiker darauf verweisen, dass wir die Daten immer gruppieren und zu bestimmten Typen von Daten zusammenfassen, wenn wir die Bestätigung einer Hypothese ermitteln möchten. Die Nullhypothese gibt für unsere Stichproben bestimmte Verteilungen vor, in welcher Gruppe sich unser Ergebnis vermutlich befinden wird. Es handelt sich in unserem Beispiel immerhin um zwei unterschiedliche Nullhypothesen, auch wenn diese einander recht ähnlich sind. Eine vergleichbare Gruppierung kann in solchen Fällen eben zu unterschiedlichen Bewertungen führen. Im ersten Fall fällt die Wahrscheinlichkeitsfunktion bei unserem Datum schon stark ab, im zweiten noch nicht so stark. Daher deutet das Datum auf zwei unterschiedliche Situationen hin. Die verschiedenen p-Werte müssen wir keineswegs als irgendwie inkonsistent betrachten. Schön wäre es

natürlich, wenn wir das an unseren Beispielen möglichst verständlich nachvollziehen könnten, aber unsere Intuitionen sind dazu nicht so klar. Hier sind also sicher weitere Debatten erforderlich.

Die Absichten des Experimentators sollten nicht über die Bestätigung einer Hypothese entscheiden. Der p-Wert eines Datums hängt aber nicht einfach nur von der Hypothese und dem Datum (und womöglich weiterem Hintergrundwissen) ab, sondern auch noch von dem *Stichprobenplan* bzw. der *Stoppregel*, die der Experimentator gerade verfolgt. So kann es passieren, dass zwei Wissenschaftler dieselbe Hypothese H_0 untersuchen und dasselbe Datum vorliegt, aber H_0 für den einen Wissenschaftler als Nullhypothese widerlegt ist, während das für den anderen nicht der Fall ist, weil sie beide nach unterschiedlichen Plänen vorgegangen sind. Hier erkennt der Bayesianer auf Seiten der klassischen Statistiker die Subjektivität, die der sonst ihm vorwirft. Es entscheiden nicht die Daten über die Hypothesen, sondern die Absichten des Experimentators, also das, was sich in seinem Kopf dazu noch abspielt. Das sieht überhaupt nicht mehr nach objektiver Wissenschaft aus.

Auch zu dem Problem der Stoppregel haben Bayesianer eine Reihe von Beispielen diskutiert, und wir finden einige bei Wagenmakers (2007). Eines beschreibt einen Fragebogen (über die Bedeutung von p-Werten) mit 12 Ja-nein-Fragen. Alle Fragen seien dabei etwa gleich schwer zu beantworten. Wenn wir hier als Nullhypothese annehmen, unsere Trefferquote bei der Beantwortung sei gerade $\theta = 0.5$, also eine reine Zufallsquote, dann ist die Teststatistik t = »Summe der Treffer« binomialverteilt. Wir erhalten nun für 9 richtige Antworten als p-Werte:

$$p_{\text{einseitig}} = P(t \geq 9 | \theta = 0{,}5) = 0{,}073 \text{ und}$$
$$p_{\text{zweiseitig}} = P(t \leq 3 \text{ oder } t \geq 9 | \theta = 0{,}5) = 0{,}146$$

Die Ergebnisse sind also noch nicht signifikant zum Niveau 5%, aber der einseitige Test ist signifikant etwa zum Niveau 10%. Doch nun zum Problem: Nehmen wir an, der Experimentator sagt nun, dass er nicht beabsichtigte, 12 Fragen zu stellen, sondern solange weiter zu fragen, bis man 3 Fragen falsch beantwortet hat und das war eben zufälligerweise nach 12 Fragen der Fall. Dann müssen wir allerdings die

Wahrscheinlichkeiten für die Teststatistik t anders berechnen, nämlich mit Hilfe der sogenannten *negativen Binomialverteilung* (vgl. Diez et al. 2012 Kap. 3.5.1). Sie gibt uns die Wahrscheinlichkeit dafür an, dass man beim n-ten Versuch gerade den f-sten Treffer erzielt:

$$P(\text{n ist f-ster Treffer beim n-ten Versuch}|\theta) = \binom{n-1}{f-1} \theta^{n-f} (1-\theta)^{f}$$

Der Unterschied zur Binomialverteilung liegt vor allem darin, dass bei der negativen Binomialverteilung der letzte Versuch in jedem Fall ein Treffer sein muss (in unserem Beispiel ist der »Treffer« allerdings eine falsche Antwort). Dadurch sind weniger Permutationen zu berücksichtigen als bei einer entsprechenden Binomialverteilung. Dadurch erhalten wir andere p-Werte für unsere Ergebnisse, allein aufgrund der anderen Absichten der Experimentatoren.

Dabei ist aber unsicher, wieviele Fragen zu stellen sind. Wir müssen auch noch damit rechnen, dass die dritte falsche Antwort erst sehr spät auftritt. Die unwahrscheinlicheren und extremeren Ergebnisse bei diesem Stichprobenplan sind also die, bei denen die dritte falsche Antwort erst oberhalb von 12 auftritt. Damit ergibt sich nun ein p-Wert von 0,33. Das kann man aber den Daten (drei falsche Antworten bei 12 Fragen) noch nicht ansehen. Welcher Stichprobenplan tatsächlich verfolgt wurde, kann uns nur der Experimentator sagen. Damit haben wir die oben geschilderte Situation, dass zwei Wissenschaftler dieselben Hypothesen und dieselben Daten haben können (3 Fehlversuche bei 12 Fragen, wobei gerade die letzte Frage falsch beantwortet wurde) und sie anhand desselben Auswertungsverfahrens trotzdem unterschiedliche Konsequenzen ziehen müssen, weil sie dabei unterschiedliche Absichten verfolgt haben.

Das wirkt unplausibel und birgt weitere Risiken für die Praxis. Wir bewegen uns schnell in einem Graubereich. Denken wir an ein Experiment aus der Medizin, bei dem wir im Laufe mehrerer Jahre immer wieder Patienten mit Krankheit X mit einem neuen Medikament behandeln und gegen eine Placebo-Gruppe testen. Selbst wenn man mit der Absicht gestartet ist, in beiden Gruppen jeweils 100 Patienten zu behandeln, kann das auf viele Schwierigkeiten stoßen. Es gibt etwa etliche Patienten, von denen wir irgendwann merken, dass sie längst ausgestiegen sind

(mangelnde »compliance«), oder bei einer selteneren Krankheit lassen sich eine Zeit lang keine neuen Probanden finden und wir müssen zum Ende kommen. Wir müssen also womöglich mit weniger Teilnehmern auskommen. Wie lassen sich die Ergebnisse dann auswerten? Oder sind sie im Sinne des Signifikanztestens nicht auszuwerten? Werden sich die Forscher auch daran halten und wie kann man das überprüfen?

Was ist, wenn sich schon erste Indizien für den Erfolg des neuen Medikaments zeigen, sie aber noch nicht signifikant sind? Der Forscher entschließt sich vielleicht nun, weitere 200 Patienten einzubeziehen. Kann er dann ein neues Experiment auswerten, in dem er die alten Daten nun mit den neuen kombiniert? Das scheint doch eine sinnvolle Strategie zu sein, einfach mehr Daten zu erheben und sie zusammen mit den bisherigen Daten auszuwerten. Aber wenn das erlaubt ist, können wir natürlich immer weiter tricksen. Das entspricht zunehmend dem Vorgehen, weitere Daten zu erheben, bis wir irgendwann doch auf ein signifikantes Ergebnis stoßen.

Für den Extremfall, dass das unser Stichprobenplan ist: »Nimm immer neue Fälle hinzu, bis das Ergebnis signifikant ist, stoppe dann und publiziere das Resultat«, erhalten wir mit Wahrscheinlichkeit 1 ein signifikantes Resultat wie schon William Feller (1940) zeigen konnte, wenn wir nur genügend Zeit haben. Tatsächlich kann mit dieser Strategie jede wahre Nullhypothese auf jedem Signifikanzniveau mit Wahrscheinlichkeit 1 probabilistisch falsifiziert werden.

Natürlich darf das im Rahmen des Signifikanztestens nicht erlaubt werden, auch wenn es für die Praxis nicht sehr relevant ist, da die Stichprobenserien sehr lang werden können, bevor signifikante Ergebnisse auftreten. Das Problem führt schnell zu einer Forderung nach einfachen starren Stichproblenplänen für das Signifikanztesten. Das passt aber eigentlich nicht zu den Anforderungen an die Flexibilität des Experimentierens, mit denen wir in der Praxis konfrontiert werden. Dort werden Gelder bewilligt oder gestoppt, die Zeit wird knapp, Mitarbeiter oder Versuchspersonen steigen aus, moralische Grenzen verlangen ein Ende der Experimente und vieles mehr, was unsere ursprünglichen Absichten durchkreuzen kann. Wie oft tatsächlich in der Praxis gegen die Anforderung verstoßen wurde, die ursprüngliche einfache Stoppregel auch einzuhalten, werden wir wohl auch nicht erfahren. Das lässt sich

aus den Ergebnissen nicht mehr ablesen. Der Bayesianer kennt keine Stoppregeln und scheint in diesem Punkt klar vorne zu liegen. Er kann hier dem klassischen Statistiker vorwerfen, dass dessen Ergebnis von subjektiven Faktoren abhängt.

Der p-Wert der Daten gibt nicht zuverlässig Auskunft über den Bestätigungsgrad der Forschungshypothese. Wir haben schon über das p-Wert Postulat gesprochen, wonach der p-Wert angibt, dass die Nullhypothese umso überzeugender falsifiziert wurde, umso kleiner der p-Wert der Daten ist. Das würden vermutlich die meisten klassischen Statistiker unterschreiben (Wagenmakers 2007, 787 gibt einige entsprechende Einschätzungen an), und es ist auch nicht ganz falsch. Aber leider auch nicht ganz richtig. Der klassische Statistiker hat dann aber kein Maß mehr für die Stärke der Bestätigung einer Forschungshypothese, wenn wir das p-Wert Postulat aufgeben. Intuitiv bestätigen die Daten eine Hypothese mehr oder weniger und wir hatten für den Bayes-Faktor eine entsprechende Interpretation angegeben, die Einschätzungen zum Bestätigungsgrad sichtbar macht.

Wagenmakers (2007) zeigt an einem Beispiel, dass derselbe p-Wert mit unterschiedlich starken Bestätigungen der Forschungshypothese im Sinne der Bayes-Faktoren einhergehen kann, je nachdem, wie groß die Anzahl der Versuchsobjekte ist und wie die Vorher-Dichten für die Forschungshypothese gestaltet sind. Den klassischen Statistiker muss das natürlich nicht beeindrucken, denn er kann sich darauf zurückziehen, dass der die in der bayesianischen Analyse eingesetzten epistemischen Wahrscheinlichkeiten als unseriös ablehnt. Ähnliche Ergebnisse wird allerdings auch eine entsprechende frequentistische Analyse im nächsten Kapitel erbringen. Spätestestens diese sollte ihm zu denken geben. Es ist jedenfalls offensichtlich, dass der p-Wert nicht die Vorher-Plausibilität unserer Forschungshypothese berücksichtigt und daher der Basisratenfehlschluss droht. Wir werden aber sehen, dass wir außerdem auch dem Fehlschluss der probabilistischen Falsifikation erliegen können.

6.5 Ein Maß für die Bestätigungsstärke von signifikanten Daten

6.5.1 Das Problem der fehlenden Likelihoods

Ein Problem der klassischen Statistik ist schon, dass wir nicht immer direkte Likelihoods angeben können, die von den Theorien vorgegeben werden. Insbesondere war in den bisherigen Beispielen die Forschungshypothese meist so beschaffen, dass sie keine definitiven Likelihoods mehr vorgibt. Die Hypothese, dass unsere Münze unfair ist, wäre ein solches Beispiel. Sie liefert definitiv keine Likelihoods für irgendwelche Ergebnisse mehr und erlaubt daher auch keine probabilistische Falsifizierung. Sie wäre somit als Nullhypothese ungeeignet. In der bayesianischen Statistik können wir – wie oben schon gesehen – mit einer Vorher-Dichte $p(\theta)$ für den Parameter θ (für die jeweilige Wahrscheinlichkeit für Kopf) an die Sache herangehen und erhalten durch Integration über θ dann wieder eine Likelihood für die verschiedenen Daten.

Der klassische Statistiker lehnt diesen Weg aber als unseriös ab, da er sich auf die *epistemische* Dichtefunktion p stützen muss, die sich nicht als relative Häufigkeit interpretieren lässt, die wir dann entsprechend messen könnten. Sie ist eben nur eine (subjektive) Schätzung dafür, für wie wahrscheinlich wir die verschiedenen Werte für θ halten. Im Falle der Münze besagt unser Hintergrundwissen, dass zumindest die Extremwerte nahe bei 0 und nahe bei 1 sehr unwahrscheinlich sind und wir daher vermutlich eine Dichtefunktion mit einem Bauch in der Mitte wählen sollten. Liegen spezielle Informationen über die Beschaffenheit der Münze vor, können wir die Funktion auch noch asymmetrisch gestalten. In Abwesenheit solcher Informationen, würden wir sie vermutlich symmetrisch um den Wert 0,5 annehmen. Bayesianer wählen gerne bestimmte Beta-Verteilungen, die gut geeignet erscheinen, unser Hintergrundwissen einzubringen und beim Updaten mit einfachen Daten wiederum Beta-Verteilungen ergeben.

Wir haben im Beispiel von Bems Hellseherexperimenten aber auch schon gesehen, dass diese Vorher-Dichte gerade ein entscheidender Streitpunkt sein kann und eine unterschiedliche Wahl zu anderen Ergebnissen führen kann. Daher sollte der Bayesianer nach Möglichkeit

auch an der Likelihoodanbindung festhalten, für die ich im vorigen Kapitel argumentiert habe, und dann nur auf objektive Likelihoods vertrauen. Das zeigt, dass hier auch für den Bayesianer erhebliche Probleme liegen.

Da der klassische Statistiker diesen Weg über eine Vorher-Dichte jedenfalls nicht mitgeht, kann er auch keine komparativen Werte des Typs $P(E|H_0)/P(E|H_1)$ bestimmen, weil die Likelihood im Nenner oft nicht bestimmt ist. Man muss die Falsifikation der Nullhypothese also ganz auf den ersten Wert $P(E|H_0)$ stützen. Wir hatten aber schon gesehen, dass das leider keine zuverlässigen Falsifikationen liefert, weil wir den dafür erforderlichen Vergleich der Likelihoods nicht mehr gewährleisten können.

Trotzdem wird der klassische Statistiker an dieser Stelle natürlich nicht gleich aufgeben, sondern argumentieren, dass die Kenntnis von $P(E|H_1)$ zwar schön wäre, aber eben leider nicht zu haben ist und wir deshalb den einfacheren Weg der Signifikanztests beschreiten müssen. Von den bayesianischen Kritiken dürfe man sich auch dabei nicht irremachen lassen, denn schließlich beruhen die jeweils auf solchen unseriösen Vorher-Wahrscheinlichkeiten und sind damit abzulehnen.

Wie können wir aber dann zu einer Bewertung der Signifikanztests gelangen? Frequentistisch sind nur die Verfahren selbst und nicht die einzelnen Schlüsse zu bewerten. Daher werde ich die *Filteranalogie* einführen, die genau dazu dient, eine rein frequentistische Analyse der Signifikanztests durchzuführen.

Doch das Problem fehlender objektiver Likelihoods kann natürlich sogar schon für die Nullhypothese auftreten. Für einige Fälle solcher disjunktiven Hypothesen bietet aber manchmal auch die klassische Statistik Lösungen an (vgl. u.a. Sober 1991). Nehmen wir etwa die Hypothese H, *dass die Wahrscheinlichkeit p für Kopf größer als 0,6 ist.* Wie hoch ist dann die Wahrscheinlichkeit für höchstens 40-mal Kopf bei 100 Würfen? Wir können zumindest sagen, dass diese Wahrscheinlichkeit im Lichte unserer Hypothese gerade für den Grenzfall, dass p=0,6 ist, am größten wird. Wenn unsere Ergebnisvariable E wieder die Anzahl der Köpfe angibt, können wir so die Wahrscheinlichkeit berechnen:

$$P(E \leq 40|p=0{,}6) \approx 0{,}00004$$

Also ist jedenfalls klar, dass bei jeder Wahrscheinlichkeit für p, die oberhalb von 0,6 liegt, ein Ergebnis von nur 40-mal Kopf zu einer probabilistischen Falsifikation führen sollte. Der klassische Statistiker kann so einen sinnvollen einseitigen Ablehnungsbereich berechnen:

Ablehnungsbereich
Bei 51-mal Kopf oder weniger ist die Hypothese H (p>0,6) abzulehnen.

Dieses Verfahren lässt sich aber nicht so ohne Weiteres auf alle Fälle disjunktiver Hypothesen anwenden und kann dem klassischen Statistiker Grenzen aufzeigen, die er nur mit Tricks aus dem Bereich des Bayesianismus überwinden kann.

6.5.2 Das Problem der mehrfachen Experimente

Ein weiteres Problem für die Anwendung von Hypothesentests im Rahmen der klassischen Statistik ist das Vorliegen mehrerer Experimente und Testergebnisse. Kommen wiederholt neue Daten herein, wird der Bayesianer die Hypothesen damit immer wieder updaten und auch die Reihenfolge des Dateneingangs spielt dabei bekanntlich keine Rolle. Doch für die klassische Statistik gibt es leider keine einfachen Verfahren, um die Ergebnisse mehrerer Signifikanztests miteinander zu verrechnen. Am ehesten kann man noch sagen, dass zwei signifikante Ergebnisse einander verstärken, aber es ist auch unklar, in welchem Ausmaß das der Fall ist.

Wir können die Daten zweier unabhängiger Experimente in der Regel nicht einfach zusammenwerfen und als neuen Test betrachten, da die genauen Experimentbedingungen jeweils für unsere Testverfahren relevant sind und wir mit einem Zusammennehmen der Daten kein entsprechendes neues Gesamtexperiment mehr erhalten. Gibt es also z.B. ein signifikantes und ein insignifikantes Ergebnis, besagt das nicht sehr viel, da das insignifikante Ergebnis normalerweise auch keine wirkliche Stärkung der Nullhypothese bedeutet. Das führt sogar dazu, dass mehrere insignifikante Ergebnisse keineswegs besonders stark gegen unsere Forschungshypothese H sprechen müssen. Die wird also letztlich nie falsifiziert, sondern eher aufgegeben im Sinne Kuhns, weil

sich mit ihrer Hilfe eben bestimmte Probleme nicht lösen lassen bzw. de facto keine signifikanten Resultate erzielt werden.

Das Zusammenrechnen insignifikanter Ergebnisse bietet sogar noch spezielle Schwierigkeiten: Nehmen wir wieder unsere einfache Hypothese H, eine bestimmte Münze sei unfair, und die entsprechende Nullhypothese H_0, wonach die Münze fair ist (θ=0,5). Betrachten wir dazu 10 einfache Experimente mit unserer Münze mit jeweils 10 Würfen. Als Ablehnungsbereich zum 5% Niveau erhalten wir etwa Z = {0,1} ∪ {9,10}. Nun werfen wir in allen 10 Versuchen gerade 7-mal Kopf. Dann sind alle Einzelergebnisse insignifikant, aber ihre Summe ist bereits signifikant, wenn wir die Experimente vereinigen, denn wir hatten als Ablehnungsbereich für 100 Würfe oben schon den Bereich Z = {0,1,...,39} ∪ {60,61,...,100} festgelegt. Statt unsere eigentliche Hypothese zu schwächen, haben diese vielen insignifikanten Ergebnisse unsere Hypothese sogar bestätigt, weil sie zusammen klar gegen die Nullhypothese aussagen.

Es bleibt uns bei mehreren Resultaten daher nur eine *Metaanalyse* im Sinne einer informellen Abwägung, die auch die Qualität der einzelnen Studien berücksichtigen wird, zumal diese in der Praxis oft nicht sehr hoch ist (vgl. Beck-Bornholdt & Dubben 1998). Man darf dabei außerdem nicht vergessen, dass die substantiellen Konkurrenzhypothesen in spannenderen Fällen durch die spezielle Gestaltung des Experiments ausgeschlossen werden müssen. Ob das etwa durch das Verfahren der Randomisierung immer zuverlässig gelingt, ist auch nicht ganz so eindeutig, wie wir es gern hätten (vgl. Kap 7).

Zusätzlich können Wiederholungen von Experimenten und ein Zusammenrechnen der Daten auch zu Konflikten mit der Stoppregel führen. Ähnlich wie innerhalb eines Experiments müssen wir auch hier wieder fragen, nach welchen Regeln wir weitere Experimente auswerten wollen. Die Strategie, solange weitere Experimente durchzuführen und einzubinden, bis wir etwa einen bestimmten Gesamt-p-Wert erreicht haben, entspricht der entsprechenden Stoppregel für Einzelexperimente, die wir schon als nichtzulässig erkannt haben. Hier bleiben also Defizite der klassischen Signifikanztests, die für die Bayesianer so nicht bestehen.

6.5.3 Die Filteranalogie als Bewertungsmaßstab

Uns wird aber vor allem die Frage beschäftigen, was wir mit einer *probabilistischen Falsifikation* im Rahmen eines Signifikanztests genau gewonnen haben. Bei einer strikten Falsifikation ist jedenfalls geklärt, dass die falsifizierte Hypothese definitiv falsch sein muss (auch wenn sich das in der Praxis nicht so leicht nachweisen lässt, weil es eben noch das Duhem-Quine-Problem der Hilfshypothesen gibt). Für eine probabilistische Falsifikation lässt sich aber nicht behaupten, dass die so falsifizierte Nullhypothese tatsächlich falsch ist. Letztlich wird es um die Frage gehen, was wir damit über unsere Forschungshypothesen aussagen können.

Im Rahmen der klassischen Statistik können wir jedoch nicht die einzelnen Hypothesen (etwa durch epistemische Wahrscheinlichkeiten) bewerten, sondern nur das Verfahren des Signifikanztests als Ganzes. Wir dürfen uns dazu nur auf die relativen Häufigkeiten beziehen, mit denen das Verfahren gute Ergebnisse liefert. Anschließend müssen wir allerdings noch weiter darüber diskutieren, was derartige relative Häufigkeiten uns auch über den Einzelfall sagen können, doch davon sind wir vorerst noch weit entfernt.

Einige Autoren haben dazu schon entsprechende bayesianische Maße entwickelt, die angeben sollen, wie hoch die (epistemische) Wahrscheinlichkeit der Forschungshypothese nach einem bestandenen Signifikanztest ist, wahr zu sein. Wacholder et al. (2004) führten die FPRP (»the false positive report probability«) ein und Ioannidis (2005) entwickelte die PPV (»the positive predictive value«), die die Nachher-Wahrscheinlichkeit der Forschungshypothese angeben soll. Bundschuh et al. (2013) setzten diese Konzepte für die Auswertung ihrer experimentellen Studien ein. Diesen Autoren ging es vor allem darum, den Einfluss der Vorher-Wahrscheinlichkeiten auf die Nachher-Wahrscheinlichkeiten aufzuzeigen. Dazu untersuchten sie eine Reihe von Faktoren, die Einfluss auf die Vorher-Wahrscheinlichkeit der Forschungshypothese haben. Sind die getesteten Hypothesen zu Beginn recht unplausibel – wie im Fall der Hellseherhypothese – droht wieder der Basisratenfehlschluss.

Mir geht es jedoch darum, eine rein frequentistische Analyse von Signifikanztests durchzuführen (die allerdings zu ähnlichen Größen

führt), da der klassische Statistiker die Analyse sonst ablehnen müsste, weil sie auf – seiner Meinung nach – zu subjektiven Größen beruht. Um eine solche Häufigkeitenanalyse möglichst intuitiv und durchschaubar zu gestalten, habe ich die *Filteranalogie* in Bartelborth (2016) entwickelt. Signifikanztests wirken wie ein Filter, der bestimmte Hypothesen (etwa als signifikant bestätigt) durchlässt und andere zurückhält. Ein Signifikanztest (zum Niveau 95%) wirkt dabei folgendermaßen: Von 100 *wahren Nullhypothesen*, die in den Filter hineingehen, passieren ihn im Durchschnitt 95, während 5 wahre Nullhypothesen im Filter hängen bleiben, d.h. probabilistisch falsifiziert werden. Das heißt im Umkehrschluss für die Forschungshypothesen, die wir tatsächlich untersuchen und möglichst begründen wollen, dass von 100 *falschen Forschungshypothesen*, die ich in den Filter des Signifikanztests gebe, ca. 5 Hypothesen als signifikant bestätigt durchschlüpfen, während die anderen 95 zuverlässig zurückgehalten werden.

Das klingt erst einmal nach einer guten Quote. Darf ich das so deuten, dass etwa 95% aller Forschungshypothesen, die den Filter passieren (d.h. als signifikant bestätigt werden), dann auch wahr sind? Genau das darf ich natürlich nicht tun und hier setzt u.a. die Kritik der Bayesianer an. Schicke ich etwa nur falsche Forschungshypothesen in den Filter, werden einige durchkommen, aber diese sind natürlich weiterhin alle falsch. Das entspricht dem Basisratenfehler: Wenn eine Hypothese eine geringe Vorher-Wahrscheinlichkeit hat, wird sie durch bestätigende Daten zwar in ihrer Wahrscheinlichkeit steigen, aber möglicherweise immer noch einen sehr niedrigen Wert aufweisen.

Das heißt für unseren Filter: Schicke ich überwiegend falsche Forschungshypothesen in den Filter, wird die Quote der falschen Hypothesen unter den Hypothesen, die den Filter passieren, weiterhin hoch sein. Die allgemeine Plausibilität einer Hypothese – bzw. die Quote der wahren Hypothesen, die in den Filter geschickt werden – wird im Signifikanztest allerdings nicht weiter berücksichtigt. Wenn wir also eine Forschungshypothese nach bestandenem Signifikanztest vorläufig akzeptieren, begehen wir damit aus Sicht der Bayesianer oft den Basisratenfehlschluss. Sie weist womöglich trotz der Daten, die tatsächlich (ein wenig) für die Hypothese sprechen, immer noch eine sehr geringe Wahrscheinlichkeit auf und sollte daher noch nicht einmal

vorläufig akzeptiert werden. Leider beantworten uns Signifikanztests eben nicht die Frage, ob die Forschungshypothese nun insgesamt plausibel oder zumindest plausibler als ihre Negation ist.

Die *Filteranalogie* kann das Problem gut verdeutlichen. Nehmen wir an, unsere Wissenschaftler haben Pech und testen 1000 *falsche Forschungshypothesen* bzw. schicken sie in den Filter (also 1000 wahre Nullhypothesen), dann werden trotzdem 50 unserer Theorien als signifikant bestätigt wieder herauskommen, aber alle davon sind nach Voraussetzung falsch. Die Wahrscheinlichkeit für solche Hypothesen wahr zu sein wäre also genau genommen Null trotz ihrer signifikanten Bestätigung. Solange wir nichts über die Quote wissen, mit der wahre und falsche Theorien in den Filter geschickt werden (Bayesianer würden das die Vorher-Wahrscheinlichkeit der Hypothesen nennen), können wir nur wenig über die epistemische Qualität der Theorien sagen, die den Filter passiert haben.

Und das ist leider nicht nur ein theoretisches Problem. Die technischen Möglichkeiten der Datenauswertung im Hinblick auf irgendwelche signifikanten Zusammenhänge sind inzwischen so gut, dass die Gefahr groß ist, dass u.a. sehr viele falsche Hypothesen getestet werden. So wird etwa für eine größere Zahl an Faktoren (Ernährungsgewohnheiten, Umwelteinflüsse etc.) untersucht, ob sie einen Einfluss auf das Entstehen verschiedener Krankheiten zeigen. Dabei werden also ziemlich viele Hypothesen (vom Typ »X hat Einfluss auf Y«) gebildet, von denen die meisten vermutlich falsch sein werden. Doch trotzdem passieren von den vielen falschen Hypothesen ca. 5% unseren Filter. Sie verderben den Brei an Hypothesen, die wir schließlich übrig behalten, die den Filter passiert haben.

Diese Menge an passierenden Hypothesen setzt sich in unbekannter Weise aus wahren und falschen Hypothesen zusammen und wir verfügen über keine Abschätzungen, wie hoch der Anteil falscher Hypothesen in dieser Menge von signifikant bestätigten Hypothesen zum Schluss ist. Genau für diese Herangehensweise mit Massentests ist der Signifikanztest also eigentlich besonders ungeeignet. Er dient vielmehr nur dazu, für an sich schon plausible Hypothesen zu zeigen, dass man die dazu alternative Nullhypothese zurückweisen kann, wonach die Daten

genauso gut als bloßes Zufallsresultat gelten dürfen und kein richtiger Effekt vorliegt.

Natürlich lassen sich die Ergebnisse wieder verbessern, indem wir das Signifikanzniveau senken, wenn wir schon viele Hypothesen gleichzeitig in *einem* Test beurteilen (s. Bortz & Döring 2006). Aber wann *ein* Test und wann *mehrere* vorliegen, ist sicher eine konventionelle Angelegenheit und ein Forscher muss uns noch nicht einmal mitteilen, bei welcher Gelegenheit er denn auf eine bestimmte signifikante Korrelation gestoßen ist. Seine anderen vielen Fehlschläge muss er uns nicht unbedingt nennen. Die Schlussfolgerung kann daher nur sein: Wenn wir Signifikanztestergebnisse ernst nehmen wollen, muss vorher schon belegt werden, dass die so getesteten Theorien bereits eine gewisse Plausibilität für sich haben, so dass zu vermuten ist, dass die Quote der dann im Signifikanztest untersuchten wahren Theorien hoch ist. Also ist auch der klassische Statistiker auf eine *intuitive Beurteilung der Anfangsplausibilität* seiner Theorien angewiesen genau wie der Bayesianer. Er ist nur nicht gezwungen, sie zu quantifizieren.

Schauen wir uns ein weiteres Beispiel an: Es werden nun z.B. 1000 falsche und 100 wahre Theorien in den Filter geschickt. Je enger unser Filter gestellt wird (umso niedriger also das Signifikanzniveau α angesetzt wird), umso mehr wahre Theorien werden übrigens ebenso ausgeschieden, wie das für einen Filter typisch ist. Das können z.B. gut 20% der wahren Theorien sein. Das nennt man den Fehler 2. Art oder β-Fehler: Hier passieren wahre Theorien den Filter nicht, bzw. es gelingt uns nicht, die an sich falschen Nullhypothesen zu unseren wahren Theorien probabilistisch zu falsifizieren. Dann passieren also 50 falsche und 80 wahre Theorien den Filter. Das sieht immer noch nicht sehr gut aus, wenn es sich um Theorien handelt, auf die wir uns etwa in der Behandlung schwerer Krankheiten verlassen wollen. Immerhin ist dann noch mehr als jede dritte Theorie falsch, die unseren Filter passiert hat.

Diese *Erfolgsquote* unserer Signifikanztests möchten wir nun etwas systematischer untersuchen. Dazu werden wir insbesondere den *Fehler zweiter Art* berücksichtigen müssen. *Fehler erster Art* liegen vor, wenn wir eine falsche Theorie nach einem signifikanten Ergebnis akzeptieren. Diesen Fehlertyp versuchen wir mit dem Signifikanzniveau klein zu halten. (Hier wurde dann eine wahre Nullhypothese falsifiziert.) Fehler der

zweiten Art liegen vor, wenn wir eine wahre Theorie nicht akzeptieren. Hier wurde eine falsche Nullhypothese einfach nicht falsifiziert. Die Nullhypothese wurde dann intuitiv sehr schwach bestätigt und unsere eigentlich wahre Forschungshypothese passiert leider den Filter nicht.

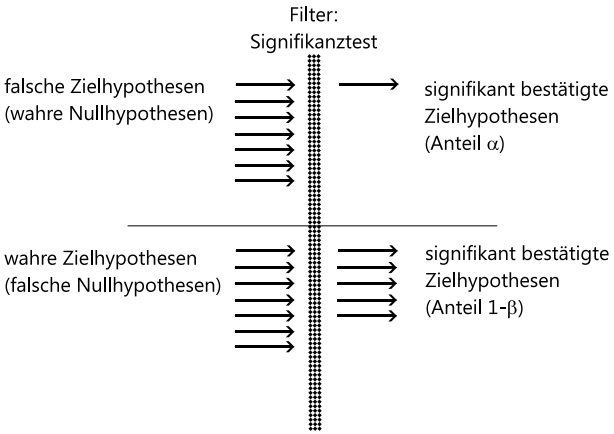
Der mögliche Fehler 2. Art wird meist etwas stiefmütterlich behandelt. Man ist aber natürlich durchaus daran interessiert, den Fehler 2. Art genauer abzuschätzen und ebenfalls möglichst klein zu halten, obwohl der Fehler 1. Art als schlimmer erscheint, da wir dort eine falsche Theorie akzeptieren, während wir beim Fehler 2. Art uns nur eine wahre Theorie durch die Lappen gehenlassen. Doch wie können wir die Wahrscheinlichkeit für einen Fehler 2. Art überhaupt bestimmen? Das ist oft nicht so einfach. In unserem Münzwurfbeispiel war die Nullhypothese H_0 gerade, dass die Münze fair ist (θ=0,5) und unsere Hypothese H_1 gerade, dass die Münze unfair ist ($\theta \neq 0{,}5$). Mit den üblichen Buchstaben für die Daten E und die Zurückweisungsmenge Z und die Akzeptanzmenge A erhalten wir:

Fehler 1. und 2. Art
Wahrscheinlichkeit für Fehler 1. Art: $P(E \in Z|H_0)$
Wahrscheinlichkeit für Fehler 2. Art: $P(E \in A|H_1)$.

Man beachte hier, dass mit Z der Zurückweisungsbereich für die Nullhypothese und mit A einfach das Komplement im Wertebereich unserer Teststatistik (oder der Gegenbereich), also der Akzeptanzbereich für die Nullhypothese gemeint ist. Das Problem ist nun $P(E \in A|H_1)$ zu bestimmen. Da unsere Hypothese H keine definitiven Aussagen über bestimmte Likelihoods trifft, ist das so in unserem Fall überhaupt nicht möglich. Wir stoßen wieder auf das Problem der fehlenden Likelihoods.

Um dieses Problem im Folgenden zu überwinden, werde ich immer eine ganze Reihe von konkreten Szenarien betrachten. In unserem Münzwurfbeispiel gibt es eine konkrete Punktwahrscheinlichkeit θ dafür, dass beim Werfen Kopf kommt – nur dass wir die leider nicht kennen. Eine der folgenden Aussagen ist allerdings in jedem Fall wahr: »$H_0 \vee H_\theta$«, für mindestens ein $\theta \in \Theta$. Deshalb untersuche ich für diese Szenarien (wobei ich mich auf einige typische Repräsentanten beschränken werde)

jeweils, wie sich der Signifikanztest als Hypothesenfilter in diesen Fällen bewährt. Wir vergleichen also unsere Nullhypothese jeweils mit verschiedenen Punkthypothesen als Forschungshypothesen, weil wir in diesen Fällen auch den Fehler der zweiten Art bestimmen können. Dazu können wir uns nun an einem einfachen Bild überlegen, wie der Filter insgesamt wirkt, wenn wir die beiden Fehleranteile α und β berücksichtigen:



Grafik 6.2: Der Signifikanztest als Filter:
Wieviele wahre und wieviele falsche Hypothesen passieren den Filter?

Außerdem benötigen wir noch die relative Häufigkeit, mit der unser Forscher wahre und falsch Forschungshypothesen in den Filter steckt. Nehmen wir q als Anteil der wahren an der gesamten Menge an Zielhypothesen, die wir in den Filter stecken. Nennen wir nun q die *Vorher-Wahrheitsquote* in Anlehnung an entsprechende bayesianische Konzepte. (Allerdings bleibt q einfach eine relative Häufigkeit, die wir nur nicht kennen.) Für die Qualität des Filters ist jeweils entscheidend, was nun »hinten herauskommt«, d.h., wieviele wahre Hypothesen unter den Hypothesen sind, die den Filter passieren. Wir nennen diesen Anteil w und bezeichnen ihn als die *Nachher-Wahrheitsquote* für unseren Filter bei einem gegebenem Szenario. Zusammen mit unseren beiden Fehleranteilen $\alpha = P(E \in Z|H_0)$ für den Fehler der zweiten Art: $\beta = P(E \in A|H_1)$ ergibt sich dann die folgende offensichtliche Formel (das sieht man an der Grafik) für die Nachher-Wahrheitsquote w:

Die Nachher-Wahrheitsquote: $$w = \frac{q(1-\beta)}{\alpha(1-q)+q(1-\beta)}$$

Dabei ist w einfach das Verhältnis der wahren Forschungshypothesen, die den Filter passieren, zu der Gesamtzahl der Hypothesen, die den Filter passieren. Der Bayesianer kann solche relativen Häufigkeiten natürlich immer auch als epistemische Wahrscheinlichkeiten deuten, aber wir bleiben bei unserer Deutung als bloße relative Häufigkeiten. Trotzdem sollte uns in irgendeiner Weise die Nachher-Wahrheitsquote etwas über die Qualität der Signifikanztests sagen. Welche genaue Bedeutung sie hat, werden wir später noch zu diskutieren haben (vgl. a. Bartelborth 2016).

Auch eine rein frequentistische Ansicht wird solche Häufigkeitsanalysen jedenfalls als relevant ansehen müssen, denn sonst könnten wir auch nichts Positives über die signifikant getesteten Hypothesen sagen. Sie sollen aber doch durch den Test positiv bestätigt werden. Zunächst einmal sollte jedenfalls eine hohe Nachher-Wahrheitsquote (ich denke dabei meistens an mindestens 90% s.u.) eine positive Auszeichnung für das Verfahren darstellen, ergeben sich aber niedrige Nachher-Wahrheitsquoten, ist das Verfahren weniger erfolgreich. Doch dazu später noch mehr.

Zunächst wollen wir uns einige spezielle Situationen ansehen. Dazu betrachten wir noch einmal verschiedene Fälle mit unterschiedlichen Vorher-Wahrheitsquoten und unterschiedlichen Signifikanzniveaus:

Vorher-Wahrheitsquote q	0	0,001	0,01	0,1	0,3	0,5	0,6
$\alpha = 0{,}05$ ($\beta = 0{,}2$)	0	1,6	14	64	87	**94**	96
$\alpha = 0{,}01$ ($\beta = 0{,}5$)	0	5	34	85	**95**	97	98
$\alpha = 0{,}001$ ($\beta = 0{,}8$)	0	17	67	**96**	99	99,5	99,7

Tabelle 6.5: Die Nachher-Wahrheitsquoten in Prozent

Dabei gehen wir schon davon aus, dass die β-Fehler (hier fiktiv angesetzt) anwachsen, wenn die α-Fehler kleiner werden. Wenn wir jeweils Experimente mit denselben Fallzahlen betrachten, ist das plausibel. Machen wir die Filter enger, passieren zwar weniger falsche Hypothesen (fälschlicherweise) den Filter, aber es bleiben auch immer mehr wahre Hypothesen in unserem Filter hängen. Außerdem möchte ich

ab jetzt den Filter immer als brauchbar betrachten, wenn die Nachher-Wahrheitsquote 90% oder mehr beträgt.

Jede einzelne Tabellenzelle im Inneren der Tabelle entspricht einer bestimmten Situation. Die fettgedruckte 95 in der Mitte steht etwa dafür, dass bei mindestens 30% wahren Hypothesen, die wir in den Filter schicken und einem Signifikanzniveau von 1% wir unter 100 Forschungshypothesen, die den Filter passieren, immerhin im Durchschnitt mindestens 95 sich als wahr erweisen. Das sieht nach einer Situation aus, mit der wir leben können. Man sieht aber auch, dass ein Signifikanzniveau von nur 5% noch keine zufriedenstellenden Ergebnisse liefern würde. Insgesamt sieht man hier, dass das p-Postulat nicht ganz falsch ist. Ceteris paribus gilt, dass die Nachher-Wahrheitsquoten größer werden, wenn die Irrtumswahrscheinlichkeit erster Art sinkt.

Allerdings hängt die Nachher-Wahrheitsquote auch von anderen Faktoren ab. Insbesondere von der Vorher-Wahrheitsquote q. Ist die niedrig, liefert der Filter keine guten Ergebnisse mehr. Den Extremfall zeigt die zweite Spalte der Tabelle. Der Bayesianer hat also natürlich Recht, dass auch der klassische Statistiker sich zunächst überlegen muss, ob seine Forschungshypothesen plausibel sind. Nur wenn das der Fall ist, liefert der Signifikanzfilter uns ein brauchbares Verfahren. Das ist die erste Anforderung an den klassischen Statistiker: *Schicke nur plausible Hypothesen in den Filter*!

Natürlich wird der Forscher auch persönlich »erfolgreich« sein, wenn er nur falsche Hypothesen auswählt. Schickt er 1000 falsche Hypothesen in den Filter mit Signifikanzniveau 5%, erhält er schließlich etwa 50 signifikante Resultate, die er veröffentlichen kann. Nur der Wissenschaft ist leider nicht gedient, denn alle seine Resultate sind leider falsch. Daher ist die erste Anforderung »nur plausible Hypothesen in den Filter zu schicken« so immens wichtig.

An dieser Stelle finden wir einen Konflikt zwischen den Zielen der Wissenschaft und den Zielen des einzelnen Forschers. Der Forscher möchte vor allem viele signifikante Ergebnisse veröffentlichen. Dazu bietet sich dann allerdings der Weg an, möglichst viele Hypothesen in den Signifikanztest-Filter (ST-Filter) zu schicken, denn das erhöht die Quote der signifikanten Ergebnisse und damit die Anzahl seiner Veröffentlichungen. Der Wissenschaft ist damit jedoch nur gedient, wenn

dann auch die meisten der signifikant bestätigten Hypothesen wahr sind. Das ist aber nur der Fall, wenn der Forscher nur die ganz plausiblen Hypothesen in den Filter schickt, oder ansonsten höhere Anforderungen stellt, bzw. den Filter enger stellt. Sonst ist der Weg frei für Hypothesen wie der Hellseherhypothese.

Wir gehen jedenfalls im Folgenden immer schon davon aus, dass die Forschungshypothesen, die wir in den Filter schicken, zumindest in 50% der Fälle wahr sind. Das ist schon eine recht anspruchsvolle Annahme. Das würde bedeuten, dass die Forscher beim Einsetzen des Signifikanztests höchstens in jedem zweiten Fall eine falsche Forschungshypothese testen. Ist dann alles in trockenen Tüchern? Leider nein. Wir rechnen nun für konkrete Szenarien die Nachher-Wahrheitsquoten aus. Denken wir dabei an unsere Münzexperimente und nehmen wir an, wir werfen die Münze 60-mal. Außerdem betrachten wir nun immer einseitige Forschungshypothesen, die behaupten, dass θ einen bestimmten Wert echt größer als null aufweist. Dann nehmen wir für einseitige Hypothesentests jeweils an, dass wir ein gerade so signifikantes Ergebnis erhalten, mit dem die Nullhypothese $\theta = 0{,}5$ probabilistisch falsifiziert werden kann. Was wird dann für verschiedene Signifikanzniveaus und verschiedene Werte von θ unserer Forschungshypothese H_θ aus der Nachher-Wahrheitsquote?

H_1: $100{\cdot}\theta$ =	70	65	60	59	58	57	55	52	51
β (für α=0,05)	6	25	55	61	67	72	82	92	94
w (für α=0,05)	95	94	**90**	89	87	85	78	63	56
$P(H_1\|k=37)$	67	82	83	83	81	80	75	63	57
β (α=0,01)	24	55	82	86	89	92	96	98	99
w (α=0,01)	98,7	98	95	93	**92**	89	82	60	51
$P(H_1\|k=40)$	96	97	95	93	92	**90**	85	68	60
β (α=0,001)	55	83	96	97	98	98,6	99,4	99,8	99,9
w (α=0,001)	99,8	99,4	98	97	95	**93**	86	60	48
$P(H_1\|k=43)$	99,7	99,5	98	97,7	97	95,6	**91**	73	62

Tabelle 6.6: Nachher-Wahrheitsquote w (und β-Fehler) für $\alpha = 0{,}05$ und $\alpha = 0{,}01$ und $\alpha = 0{,}001$ mit $n = 60$ und unterschiedlichen Zielhypothesen H_1 (Werte jeweils in Prozent)

Für die drei Signifikanzniveaus (5%, 1% und 1‰) werden die jeweils gerade so signifikanten Resultate (37/40/43-mal Kopf) betrachtet. Dazu werden in den unterschiedlichen Szenarien jeweils die β-Fehler, die Nachher-Wahrheitsquote und die bayesianische Nachher-Wahrscheinlichkeit der Forschungshypothese angegeben.

In dieser Tabelle sehen wir vor allem ein Problem unseres Filters: Werden die Effekte klein – testen wir die Nullhypothese also gegen Forschungshypothesen, die der Nullhypothese immer näher kommen –, so wächst der β-Fehler jeweils schnell an. Den können wir nun bestimmen, weil wir als Forschungshypothesen uns in jedem konkreten Szenario auf eine Punkthypothese beschränken, bei der θ einen festen Wert annimmt. Der β-Fehler wächst dabei (wenn wir nach rechts hin zu Situationen übergehen, in denen der Abstand der Hypothesen kleiner wird) gegen $1-\alpha$ und wird damit so groß, dass er in unserer Bestimmung der Nachher-Wahrheitsquote plötzlich doch relevant wird. Dadurch sinkt die Nachher-Wahrheitsquote auf Werte um 50% ab, die also nicht mehr besser sind, als unsere Plausibilitätseinschätzung der Hypothesen vor dem Test. Ähnliche Probleme zeigen uns auch die bayesianische Nachher-Wahrscheinlichkeit. Das NHST-Verfahren lässt sich dann auch nicht durch ein anspruchsvolleres Signifikanzniveau retten, wie die Tabelle zeigt.

Warum steigt der β-Fehler so stark an? Wenn die Forschungshypothese H_1 wahr ist, werden die meisten Daten E zu H_1 passen. Da H_1 und H_0 benachbart sind, wird E jedoch meistens ebenso zu H_0 passen. Dann fällt E aber in den Annahmebereich A der Nullhypothese und diese wird nicht verworfen, d.h. H_1 wird nicht akzeptiert. Also gilt für die meisten Daten E (bei Wahrheit von H_1), dass sie *nicht* zum Akzeptieren von H_1 führen, obwohl H_1 wahr ist. Es liegt also in diesen vielen Fällen ein Fehler 2. Art vor. Je enger H_0 und H_1 benachbart sind, umso größer wird auch dieser Fehler.

Die Bayesianer kennen das Phänomen schon, und es ist auch intuitiv plausibel: Wenn ich anhand von Daten zwischen zwei Hypothesen entscheiden muss und die Hypothesen sind eng benachbart, so benötige ich zunehmend mehr und möglichst eindeutige Daten, um noch entscheiden zu können, welche Hypothese die richtige ist. Da die Hypothesen ähnliche Behauptungen über die Daten aufstellen (also

ähnliche Likelihoods aufweisen), geben sie auch nur wenige Hinweise darauf, welches die richtige Hypothese ist. Jedenfalls liefern wenige Daten auch nur kleine Unterschiede in den Likelihoods bzw. im Bayes-Faktor und können daher höchstens schwach für die eine oder andere Hypothese sprechen.

Beim Nullhypothesen-Signifikanztest wird das Problem aber nicht sichtbar, weil er nicht wirklich *komparativ* vorgeht. Es wird nur ausgewertet, ob die Daten zur Nullhypothese passen. Ist das nicht der Fall, wird das Datum als Beleg gegen die Nullhypothese und damit als Beleg für die Forschungshypothese interpretiert. Es liegt also wiederum der *Fehlschluss der probabilistischen Falsifikation* vor, wenn das Ergebnis auch nicht zur Forschungshypothese passt. Das ist für benachbarte Hypothesen jedoch meistens der Fall. Wir schauen nur auf die kleinen Werte $P(E|H_0)$ und halten dann H_0 für widerlegt. Da jedoch H_1 ähnliche Werte für die Daten liefert ist $P(E|H_0)$ nicht viel kleiner als $P(E|H_1)$ und somit liegt keine begründete Falsifikation von H_0 vor. Dieses Problem wird jedoch erst beim Vergleich der Likelihoods sichtbar, der beim NHST nicht vorgesehen ist. Wir können daher davon sprechen, dass bei *kleinen Effekten der Signifikanztest die Daten überbewertet.* Dass der Signifikanztest nicht komparativ verfährt, führt immer wieder zu diesem Fehlschluss.

Das kann auch das Problem für die Experimente von Bem sein. Er selbst besteht ja darauf, dass wir berücksichtigen müssen, dass die auftretenden Effekte nur sehr klein sind. Das lässt sich auch in den Daten erkennen. Wir lagen da etwa bei geschätzten Trefferquoten von 53% statt der 50% der Nullhypothese. Dann erhalten wir zwar schon schnell signifikante Daten, aber die stellen höchstens eine sehr schwache Bestätigung der Hellseherhypothese dar, weil für diese Situationen der Signifikanztest nur niedrige Wahrheitsquoten liefert. Das bedeutet: In solchen Situationen (mit kleinen Effekten und kleinen Versuchszahlen) sind die signifikant bestätigten Forschungshypothesen nur selten auch wahre Hypothesen. Wenn wir dann noch berücksichtigen, dass die Vorher-Wahrheitsquote für solche Psi-Hypothesen auch noch klein sein dürfte, so stellen die Resultate keine guten Gründe dar, nun die Hellseherhypothese zu akzeptieren.

Die Frage ist aber, ob wir das nicht auch auf andere Fälle übertragen müssen, in denen in der Medizin oder den Sozialwissenschaften ebenfalls sehr kleine Effekte getestet werden? Außerdem sollten wir untersuchen, was wir dann unternehmen können, um unsere epistemische Situation zu verbessern.

Man beachte aber, dass das Problem der *Überinterpretation der Daten bei kleinen Effekten* nicht dem entspricht, was häufiger gegen Signifikanztests eingewandt wird, dass ein signifikantes Ergebnis nichts darüber aussagt, ob der Effekt auch wirklich (praktisch) bedeutsam ist. Ein solcher Effekt kann natürlich sehr klein sein, auch wenn er hochsignifikant bestätigt wurde. Über die Größe des Effekts sagt der p-Wert nichts. Er sagt nur etwas darüber, wie gut die Bestätigung dafür ist, dass die Nullhypothese falsch ist. Wie groß der Effekt ist, ist sicher für die Praxis eine wichtige Frage. Lohnt sich das ständige Pausenbrote-Schmieren, wenn der Effekt auf die Noten nur minimal ist? Mit dieser Frage beschäftige ich mich jedoch nicht. Die Größe des Effekts ist in unserem Beispiel leicht einzuschätzen, als der Abstand von θ bzw. dem geschätzten θ und dem Wert 0,5. Für den ST-Filter spielt die Effektgröße hier aber nur eine Rolle im Hinblick auf die Nachher-Wahrheitsquoten. Ob der Effekt praktisch bedeutsam ist, lasse ich komplett außen vor. Es zeigt sich in der Tabelle 6.6 nur, dass für kleine Effekte die *erkenntnistheoretische Bestätigung*, die von einem signifikanten Ergebnis ausgeht, durch den ST-Filter deutlich überschätzt wird. Das ist ein rein epistemisches Problem.

Doch lassen wir erst noch einmal den klassischen Statistiker zu Wort kommen. Er wird womöglich einwenden, dass diese Wahrheitsquoten schließlich nur *Durchschnittswerte* darstellen und daher nicht viel über den Einzelfall sagen können. Das stimmt natürlich. Diese Werte geben uns zunächst Auskunft über die *Qualität der Methode* dieses speziellen Typs von Signifikanztest und sagen uns zunächst noch nichts über die epistemischen Wahrscheinlichkeiten einzelner Hypothesen, die den Test bestanden haben. Auf solche epistemischen Wahrscheinlichkeiten möchte sich der klassische Statistiker schließlich nicht gerne einlassen.

Aber der Vertreter der Signifikanztests muss dann seinerseits erklären, wieso wir aufgrund von signifikanten Daten, bestimmte Forschungshypothesen als (zumindest inkrementell) *bestätigt* betrachten sollten.

Inwiefern stellen signifikante Daten seiner Meinung nach einen Grund (also einen Wahrheitsindikator) für bestimmte Hypothesen dar? Der beste Weg für einen Frequentisten, diese Frage zu beantworten, scheint mir der über den durchschnittlichen Erfolg der von ihm eingesetzten Verfahren zu sein. Er muss zeigen, dass seine *Verfahren* erheblich häufiger zu Erfolgen als zu Misserfolgen führen. Das heißt, er sollte sich in irgendeiner Form auf die Wahrheitsquoten seiner Methoden berufen. Um das übersichtlich und intuitiv darstellen zu können, dazu habe ich die Filteranalogie eingebracht. Der klassische Statistiker muss dabei m.E. denselben Weg einschlagen, den ich nun vorschlagen möchte, um die Nachher-Wahrheitsquoten auszuwerten.

Was bedeuten dann schlechte Nachher-Wahrheitsquoten für unser Einzelergebnis? An dieser Stelle müssen wir uns wieder auf den *statistischen Syllogismus* stützen. Zumindest sollte die Nachher-Wahrheitsquote uns ein erstes Maß für die Plausibilität der Forschungshypothesen auch im Einzelfall geben. Die Argumente für den statistischen Syllogismus aus dem letzten Kapitel möchte ich nicht alle wiederholen. Aber man muss sich die Situation so vor Augen führen: Alle Informationen über die Wahrheit einer Forschungshypothese T sollten in die Vorher-Wahrheitsquote eingehen. Dann vollziehen wir den Signifikanztest mit einem signifikanten Resultat und schätzen dafür (anhand einer Schätzung der Größe der Effekte) die Nachher-Wahrheitsquote ein. Nehmen wir an, sie sei 80%. Das besagt, dass das angewandte Verfahren (unser Filter) in gleichen Situationen im Durchschnitt in 80% der Fälle gerade wahre Hypothesen durchlässt. Haben wir keine anderen Informationen über die Wahrheit von T, sollten wir damit T eine gewisse Plausibilität von »immerhin schwach begründet« zugestehen. Wollen wir das auch noch quantifizieren, bietet sich natürlich ein Glaubensgrad von 0,8 für T im Sinne des statistischen Syllogismus an.

Auch der klassische Statistiker wird uns erklären müssen, welche epistemische Bedeutung die 80% der Wahrheitsquote für ihn besitzen. Wenn er rationale Glaubensgrade (als epistemische Wahrscheinlichkeiten) unbedingt vermeiden möchte, muss er uns eine andere Erläuterung anbieten. Die Auswirkungen auf unser Handeln sollten allerdings ähnlich aussehen. Er kann dazu eine objektive Entscheidungstheorie entwickeln oder andere Interpretationen der relativen Häufigkeiten liefern, aber

irgendeine epistemische Interpretation wird er letztlich akzeptieren müssen, sonst bleiben die ganzen Überlegungen über Häufigkeiten nur unnütze Zahlenspiele.

Tatsächlich bieten die Wahrheitsquoten eine Rechtfertigung für den Einsatz der klassischen Signifikanztests in geeigneten Situationen mit größeren Effekten. Kommen wir noch einmal auf unser Beispiel von 60 Münzwürfen zurück und nehmen wir an, wir wüssten schon, dass nur die Nullhypothese oder eine Wahrscheinlichkeit von 0,7 für Kopf (also ein mittelgroßer Effekt) in Frage kämen. Dann liefert der p-Wert eine ganz ähnliche Einschätzung wie die bayesianischen Überlegungen, wie die folgende Tabelle zeigt:

k	31	33	35	36	37	38	39	40	41	42	43	45
p-Wert	45	26	12	8	5	2.6	1.4	0.7	0.3	0.13	0.05	0.007
$P(H_1\|k)$	1	6	27	46	67	82	**92**	96	98	99	99.7	99.94

Tabelle 6.7: P-Werte (in Prozent) für n = 60 und k-mal Kopf
(und die Werte für $P(H_1|k)$ für die Zielhypothese $\theta = 0{,}7$)

Für dieses Szenario liefern uns die p-Werte einen recht guten Maßstab für die Bestätigung der Forschungshypothese, wobei wir aber auch für einseitige Hypothese hier eher an ein Signifikanzniveau von 2% oder sogar 1% denken sollte als an die üblichen 5%. (Bei zweiseitigen Tests erhalten wir immerhin schon zu jeder Seite nur noch etwa 2,5%.)

Kommen wir nun wieder auf die Situationen mit kleinen Effekten zurück: Lassen sich vielleicht die Probleme aus Tabelle 6.6 für kleine Effekte durch etwas höhere Fallzahlen lösen? Das sieht leider nicht ganz so einfach aus und es kommt ein neues Problem dazu. Wir sehen in Tabelle 6.8 zunächst, dass die größeren Fallzahlen das Problem der kleinen Effekte etwas abmildern können, aber noch nicht lösen können. Es tritt zudem noch ein neues Problem auf.

Für größere Effekte und große Fallzahlen, erhalten wir signifikante Ergebnisse und auch hohe Nachher-Wahrheitsquoten, aber unsere bayesianische Analyse zeigt schon, dass die Daten die Forschungshypothese nicht wirklich stützen. Hier stoßen wir auf Lindleys Paradox (1957), das wir auch intuitiv gut nachvollziehen können.

Betrachten wir in Tabelle 6.8 etwa die letzten drei Zeilen und dort die zweite Spalte. Wir nehmen hier an, dass mit 538-mal Kopf ein gerade so

H_1: θ =	0.7	0.65	0.6	0.59	0.57	0.55	0.52	0.51
α=0.05, n=20, beibehalten der Nullhypothese bis inklusive k=14								
w	94	**92**	89	88	86	83	78	75
$P(H_1\|k=15)$	**92**	89.6	83	82	77	71	60	55
α=0.05, n=200, beibehalten der Nullhypothese bis inklusive k=112								
w	95	95	94.5	94	**92**	88	70	58
$P(H_1\|k=113)$	0.2	20	77	81	84	83	71	62
α=0.01, n=200, beibehalten der Nullhypothese bis inklusive k=115								
w	99	99	98.6	98	97	**95**	80	67
$P(H_1\|k=116)$	5	75	94	95	94	**92**	77	65
α=0.001, n=200, beibehalten der Nullhypothese bis inklusive k=122								
w	99.9	99.9	99.7	99.6	99	**97**	81	64
$P(H_1\|k=123)$	87	99	99.5	99.4	99	**97**	84	71
α=0.05, n=1000, beibehalten der Nullhypothese bis inklusive k=525								
w	95	95	95	95	95	**95**	87	75
$P(H_1\|k=526)$	0	0	0	0.1	9	60	80	70
α=0.01, n=1000, beibehalten der Nullhypothese bis inklusive k=537								
w	99	99	99	99	99	99	**93**	80
$P(H_1\|k=538)$	0	0	1	7	69	**93**	**90**	80

Tabelle 6.8: Nachher-Wahrheitsquoten und Nachher-Wahrscheinlichkeiten für unterschiedliche Werte von n, θ und α (alle Werte in Prozent)

signifikantes Ergebnis (E) vorliegt und wir schon wissen, dass entweder θ = 0,5 oder θ = 0,7 ist. Im ersten Fall würden wir idealerweise 500-mal Kopf erwarten und im zweiten Fall 700-mal Kopf. Unser Ergebnis spricht dann intuitiv klar für die Nullhypothese, wonach Kopf die Wahrscheinlichkeit 0,5 aufweist, weil der Abstand von E zum idealen Ergebnis der zweiten Hypothese deutlich größer ist.

Das wird auch in der bayesianischen Analyse deutlich, aber nicht im Signifikanztest. Die Ursache dafür ist wiederum, dass der Signifikanztest *nicht komparativ* ist. Er vergleicht nicht, wie gut das Datum E zu den beiden Hypothesen passt, sondern bewertet nur, dass er nicht gut zur

Nullhypothese passt. Wenn die Nullhypothese wahr ist, ist E tatsächlich sehr unwahrscheinlich. Aber das falsifiziert in diesem Fall nicht die Nullhypothese, weil das Ergebnis noch unwahrscheinlicher wird, wenn die Nullhypothese falsch ist. Hier finden wir also wiederum den Fehlschluss der probabilistischen Falsifikation. Durch das NHST-Verfahren werden die Daten definitiv falsch interpretiert. Sie werden als Bestätigung der Forschungshypothese gedeutet, obwohl sie intuitiv die Nullhypothese stützen. Das nenne ich das Problem der *Fehlinterpretation der Daten bei großen Anzahlen und großen Effekten*.

Warum findet sich das Problem nicht in der Wahrheitsquote wieder? Hier erkennt man sehr gut den Unterschied zwischen allgemeinen einer frequentistischen Beurteilung des Signifikanztests als eines (Filter-) Verfahren und einer direkten epistemischen Bewertung der konkreten Daten und Hypothesen. Die bayesianische Bewertung finden wir in den direkten Likelihoods: P(k=538|θ=0,5) = 0,0014 und P(k=538|θ=0,7) = $1,8 \cdot 10^{-27}$. Damit ist die Wahrscheinlichkeit für das Datum k=538 etwa $7,5 \cdot 10^{23}$ so wahrscheinlich bei Gültigkeit der Nullhypothese wie bei Gültigkeit der Forschungshypothese (mit $\theta = 0,7$). Der entsprechende Bayes-Faktor gibt also noch einmal das an, was wir auch schon intuitiv erkannt haben, nämlich dass das Resultat zwar bei Vorliegen der Nullhypothese recht unwahrscheinlich ist, aber es trotzdem in überwältigender Weise für die Nullhypothese spricht, weil es noch viel unwahrscheinlicher wäre, wenn die Forschungshypothese wahr wäre.

Die *Nachher-Wahrheitsquote* wird aber ganz anders bestimmt. Sie orientiert sich einfach am Szenario, wonach θ gerade 0,5 oder 0,7 ist, und am Signifikanzniveau. Der α-Fehler liegt aber bei 0,01 und der β-Fehler praktisch bei 0. Schicken wir also 100 wahre und 100 falsche Forschungshypothesen in den Filter kommt nur eine falsche und praktisch alle wahren Hypothesen durch den Filter, woraus die Nachher-Wahrheitsquote von etwa 99% resultiert. Das Verfahren ist also in Ordnung. Das stimmt auch. Ein Datum wie k=538 wird bei diesem Szenario eben praktisch nur sehr selten auftreten, nämlich weniger als einmal bei 100 Experimenten. Daher schlägt diese Fehlinterpretation auch nicht negativ für das Verfahren zu Buche. Nichtsdestotrotz wird das Datum k=538 vom Signifikanztest dann, wenn es ausnahmsweise doch einmal vorkommt, völlig falsch gedeutet. Es spricht in unserem Szenario

sehr stark für die Nullhypothese, wird aber vom NHST als Bestätigung der Forschungshypothese eingestuft.

Tatsächlich tritt dieser Fehler für hinreichend große Fallzahlen in unseren Experimenten für alle Szenarien zwangsläufig auf. Das liegt einfach daran, dass die relative Häufigkeit, bei der wir ein signifikantes Ergebnis zu einem festen Niveau α erzielen, mit wachsendem n gegen 0,5 strebt. Wählen wir ein festes $\theta > 0$, so wird die relative Häufigkeit k/n für die k gerade so ein signifikantes Ergebnis darstellt für hinreichend großes n näher bei 0,5 als bei θ liegen und damit für die Nullhypothese sprechen, obwohl es im Sinne des Signifikanztests als Bestätigung der Forschungshypothese verstanden wird.

n =	50	100	200	500	1000	2000	5000	10000
k =	31	58	113	268	526	1037	2558	5082
k/n =	0,62	0,58	0,565	0,556	0,526	0,519	0,512	0,508

Tabelle 6.9: Die jeweils gerade signifikanten Werte k und die entsprechenden relativen Häufigkeiten für $\alpha = 0{,}05$.

In der Tabelle 6.9 kann man das eben beschriebene Phänomen gut erkennen. Wenn die Anzahl der Fallzahlen steigt, fällt die relative Häufigkeit der signifikanten Ergebnisse gegen 0,5. Also wächst dann irgendwann die Zahl der Daten, die vom Signifikanztest fehlinterpretiert werden, wenn wir eine Gegenhypothese mit festem θ festhalten. Das gilt dann mit hinreichend großem n sogar für kleine Effekte. Der entscheidende Fehler ist wiederum, dass die probabilistische Falsifikation nicht komparativ angelegt ist. Das Datum ist zwar sehr unwahrscheinlich aus Sicht der Nullhypothese, es ist aber eben noch unwahrscheinlicher im Lichte der Forschungshypothese. Wir haben somit zwei konkrete Probleme des Signifikanztestens identifiziert, die beide als Fehlschlüsse der probabilistischen Falsifikation gelten können.

Zwei Probleme von Signifikanztests

(1) Für kleine Effekte wird die Bestätigung durch die Daten überinterpretiert.

(2) Für große Anzahlen werden bestimmte Daten falsch interpretiert.

Das zweite Problem wird allerdings dadurch entschärft, dass die Daten, die offensichtlich falsch gedeutet werden, zumindest nur selten auftreten werden. Das erste Problem ist für die Praxis des Signifikanztestens dagegen das gewichtigere Problem. Sobald wir es mit kleineren Effekten zu tun haben – und das dürfte uns in einige Disziplinen begegnen –, sind die Nachher-Wahrheitsquoten so niedrig, dass wir ein signifikantes Ergebnis allein keineswegs schon als guten Grund für ein erstes Akzeptieren der Forschungshypothese ansehen dürfen.

Man kann das so formulieren, dass der Filter in diesen Fällen nicht mehr seine eigentliche Aufgabe erfüllt. Er sollte die falschen Forschungshypothesen zuverlässig zurückhalten – was weiterhin gelingt –, aber sollte die wahren Hypothesen genauso überwiegend durchlassen, denn nur dann erhalten wir hohe Wahrheitsquoten unter den durchgelassenen Hypothesen. Wenn der Fehler zweiter Art hingegen so groß wird, wie der erster Art, dann erfüllt das Verfahren seinen Zweck aber offensichtlich nicht mehr. Um abschätzen zu können, wie gut das Verfahren des NHST für eine bestimmte Situation funktioniert, können wir die Größe der Effekte abschätzen und dann anhand der Filteranalogie ermitteln, wie hoch die Nachher-Wahrheitsquoten jeweils sind.

Eine wesentliche Voraussetzung für die Anwendung von Signifikanztests ist dabei aber immer schon, dass die Vorher-Wahrheitsquote relativ hoch ist. Anderenfalls dürfen wir nicht davon ausgehen, dass uns ein signifikantes Ergebnis auch gute Gründe dafür liefern würde, die Forschungshypothese zu akzeptieren, weil in dem Fall zusätzlich der Basisratenfehlschluss drohen würde. Damit verliert der klassische Statistiker bereits einen wichtigen Einwand gegen den Bayesianer. Er hatte immer darauf verwiesen, dass eine erste Plausibilitätseinschätzung, bei der ein Bayesianer zunächst einschätzen muss, wie hoch die Vorher-Wahrscheinlichkeit der Forschungshypothese im Lichte unseres bisherigen Hintergrundwissens ist, zu subjektiv sei und damit in der Wissenschaft nichts verloren habe. Wir haben aber nun gesehen, dass auch der klassische Signifikanztester auf eine ähnliche anfängliche Plausibilitätseinschätzung angewiesen ist.

Es zeigt sich also wieder einmal, dass es keinen einfachen, objektiven Algorithmus gibt, der die Daten daraufhin auswertet, welche Hypothesen sie stützen. Das passt zu dem Motto im Buch von Beck-Bornholdt

& Dubben (2003, 177 ff.), das in einem Zitat besteht, das zwei Väter des Hypothesentesten Jerzy Neyman und Egon Pearson bereits 1933 formulierten:

> No test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis.

Auch wenn wir natürlich immer versuchen werden, einem solchen Algorithmus möglichst nahe zu kommen, sollten wir uns dabei an die obige Warnung erinnern. Vor allem dürfen wir nicht einfach alle möglichen Hypothesen wahllos in einen Signifikanztest schicken, weil wir dann niedrige Nachher-Wahrheitsquoten und damit selbst im Falle signifikanter Daten zu schwache Gründe für die Annahme der Forschungshypothese haben.

Ein anderes wichtiges Resultat der Debatte
Auch der klassische Statistiker ist im Rahmen eines Signifikanztests auf eine anfängliche *intuitive Plausibilitätseinschätzung* der Forschungshypothesen im Lichte unseres Hintergrundwissens zwingend angewiesen.

6.5.4 Das Neyman-Pearson-Lemma

Es gibt noch ein weiteres Bewertungskriterium für Signifikanztests, das manchmal als Ausweg aus manchen Problemen von Hypothesentests genannt wird: seine Trennschärfe, die maximal sein sollte. Es lässt sich anhand der *Gütefunktion* von Tests und dem Neyman-Pearson-Lemma erläutern. Nehmen wir wieder unser Beispiel des Münzwurfs. Wir suchen nach dem wahren Parameter θ, der die Wahrscheinlichkeit dafür angibt, dass Kopf kommt. Dazu haben wir für unseren rechtsseitigen Test zunächst zwei Hypothesen gebildet, die wir auch wie folgt beschreiben können:

$H_0 : \theta \in \Theta_0 = [0;0,5]$ und
$H_1 : \theta \in \Theta_1 = (0,5;1]$ sowie $\Theta = \Theta_0 \cup \Theta_1 = [0;1]$

Wählen wir den Parameter θ aus dem Intervall von null bis eins werden dadurch also alle möglichen Situationen beschrieben. Ein Testverfahren V war bei uns gekennzeichnet durch eine Entscheidungsregel, wonach wir die Nullhypothese ablehnen, wenn eine bestimmte Testgröße T in die Zurückweisungsmenge Z fällt, die wir vorher bestimmt haben, also etwa wenn T größer wird als ein bestimmter Wert. Dazu können wir auf Θ eine Gütefunktion wie folgt definieren:

Die ***Gütefunktion*** unseres Tests: $G(\theta) = P(T \in Z|\theta)$

Für einen möglichst guten Test zum Signifikanzniveau α sollte zunächst gelten: $G(\theta) \leq \alpha$ für alle $\theta \in \Theta_0$, denn in diesem Bereich der Nullhypothese stellt eine Ablehnung der Nullhypothese gerade den Fehler erster Art dar. Da die Funktion monoton wachsend ist, bedeutet das vor allem, dass $G(0,5) \leq \alpha$ ist. Im Bereich der Forschungshypothese sollte die Funktion dagegen möglichst groß sein, weil $1 - G(\theta)$ dort den Fehler zweiter Art beschreibt, der möglichst klein werden sollte. Im Falle unseres rechtsseitigen Tests der Münze steigt die Gütefunktion bis ca. $G(0,5) = 0,05$ an und steigt dann möglichst steil und nähert sich der eins an. Das zeigte sich in unseren Tabellen auch darin, dass die Werte für $G(0,51)$ immer noch nahe bei 0,05 liegen und G erst für größere Werte »bessere« Resultate liefert. (Unsere Tests waren allerdings *konservativ* und $G(0,5)$ bleibt etwas unter 0,05 bzw. schöpft das Signifikanzniveau nicht ganz aus, weil die Ablehnungsregel nur diskrete Anzahlen von Köpfen nennen kann, ab denen wir die Nullhypothese ablehnen.)

Eine naheliegende Anforderung an ein Testverfahren V ist dann, dass es im Bereich der Forschungshypothese für alle Parameter $\theta \in \Theta_1$ besser oder gleich abschneidet wie ein alternatives Testverfahren V* zu demselben Signifikanzniveau:

Für alle $\theta \in \Theta_1$ gilt: $G_V(\theta) \geq G_{V^*}(\theta)$

Wenn das für alle alternativen Testverfahren V* gilt, sagen wir, dass V ein *gleichmäßig bester Test zum Niveau* α für unser Testproblem ist.

Solche gleichmäßig besten Tests existieren aber nicht immer – vor allem nicht für zweiseitige Testsituationen. Aber zumindest für einfache Fälle kommt uns das Neyman-Pearson-Lemma zu Hilfe. Es besagt, dass

wir so einen Test zumindest für den einfachen Fall, in dem $\Theta_0 = \{\theta_0\}$ und $\Theta_1 = \{\theta_1\}$ einelementig sind, gefunden haben, wenn sich der Test auf folgende Art beschreiben lässt: Einen optimalen Ablehnungsbereich für Daten x finden wir anhand des Likelihoodquotienten und einer Konstante k_α: Lehne die Nullhypothese ab, wenn gilt:

$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha,$$

wobei k_α so bestimmt werden sollte, dass $G(\theta_0) = \alpha$ ist, bzw. möglichst wenig darunter liegt, wenn der Wert nicht genau getroffen wird.

Für solche speziellen Situationen spricht man auch vom Neyman-Pearson-Test. Der setzt allerdings voraus, dass man die Forschungshypothese bereits als Punkthypothese formulieren kann, was nicht oft der Fall sein dürfte. Letztlich ist er aber nicht wirklich komparativ, denn es geht wieder nur darum, die Nullhypothese zu verwerfen, und das in einem möglichst trennscharfen Test also einem mit einem möglichst hohen Wert für $G(\theta_1)$ zu erreichen.

Es soll aber auch noch eine intuitive Verteidigung von Signifikanztests zu Worte kommen. Wir hatten uns im Falle der Popperschen Falsifikationen überlegt, dass Deborah Mayo (1996) eine gute Explikation von *ernsthaften* oder *strengen* Theorientests anzubieten hat. Danach ist ein Testverfahren V dann ein strenges Testverfahren bzw. ein ernsthafter Theorientest, wenn die Wahrscheinlichkeit klein ist, dass es positiv ausfällt (V^+), wenn die Theorie T falsch ist, also wenn gilt: $P(V^+|\text{non-T})$ ist klein. Diese Wahrscheinlichkeit entspricht in Signifikanztests aber gerade unserer Wahrscheinlichkeit für einen Irrtum 1. Art und wird daher etwa mit 5% relativ klein gehalten. Wir können somit sagen, dass die recht intuitive Forderung von Popper, dass wir eine Theorie möglichst strengen Tests unterwerfen sollten, im Falle probabilistischer Theorientests gut umgesetzt wird.

6.5.5 Eine Bewertung des Signifikanztestens

Wo stehen wir nun in der Bewertung von Signifikanztests? Lassen wir zunächst noch eine Verteidigerin zu Wort kommen: Deborah Mayo oder andere Verteidiger der klassischen Statistik argumentieren, dass wir

natürlich gerne noch mehr über unsere Theorien erfahren würden, z.B. ob sie mit einer bestimmten Wahrscheinlichkeit wahr sind, doch das ist eben leider nicht zu haben. Wir sehen nur, dass unsere besten Theorien bestimmte strenge Tests bestehen, was natürlich für sie spricht und eine weitere Verrechnung damit, wie plausibel sie ansonsten sind, ist kaum möglich. Jeder Bayesianer fügt dem höchstens noch seine persönliche Einschätzung der Theorie hinzu, die aber eigentlich nicht Bestandteil einer allgemeinen wissenschaftlichen Beurteilung der Theorie sein sollte. Dafür haben wir nur unseren Filter und nicht mehr.

Doch wir haben immer wieder gesehen, dass die popperschen Falsifikationen nicht alles sind, was uns zu Verfügung steht, um eine Forschungshypothese zu beurteilen. Die Auswahl von Hypothesen ist ein komplexer Vorgang, der mit Hilfe von Experimenten oder anderen Verfahren und Überlegungen versucht, substantielle Konkurrenzhypothesen einer Hypothese H auszuschalten und in einer probabilistischen Falsifikation einer Nullhypothese gipfeln, aber auch auf positive Gründe für die Hypothese angewiesen ist. So sollten wir immer versuchen, die konkurrierenden Hypothesen im Hinblick auf ihre Erklärungskraft miteinander zu vergleichen.

Dabei spielen Falsifikationen durchaus eine Rolle: Die Nullhypothese behauptet typischerweise, dass ein bestimmtes Ergebnis, das wir beobachten konnten, keiner Erklärung durch unsere Hypothese H bedarf, sondern durchaus noch als bloße Zufallsschwankung erklärbar ist, wenn wir annehmen, dass nur ein »normaler« Zusammenhang vorliegt – eine Art von Normalfall mit einem »Nulleffekt«. Gerechtfertigt wird dieses Bestätigungsverfahren dadurch, dass es sich als eliminative Induktion darstellt, aber das Problem liegt u.a. darin, dass die probabilistischen Falsifikationen keine echten Falsifikationen sind und dass sie insbesondere nicht wirklich komparativ gestaltet sind.

Für die klassischen Signifikanztests haben wir zwei grundlegende Probleme kennengelernt. Es droht zum einen der *Basisratenfehlschluss*, wenn wir etwa anfänglich unplausible Hypothesen nach einem positiven Signifikanztest akzeptieren. Das passiert etwa im Falle der Hellseherhypothese, aber natürlich auch in brisanteren Beispielen aus den Wissenschaften.

Daher sind wir für das NHST zwingend darauf angewiesen, die *Vorher-Plausibilität* unserer Forschungshypothesen einzuschätzen – ähnlich wie das der Bayesianer tun muss –, weshalb wir nach zusätzlichen Hinweisen für die Wahrheit unserer Forschungshypothese H suchen müssen. Die sollten insbesondere darin bestehen, dass wir die *Erklärungskraft* von H für bestimmte Phänomene aufzeigen und begründen, warum die größer ist als die der Konkurrenz. Erst mit solchen weitergehenden Indizien erfährt H eine wirkliche Bestätigung. Doch das wird in der Praxis gerne vergessen, und wir vertrauen schon auf signifikante Ergebnisse allein und fragen nicht mehr danach, wie gut H relativ zur Konkurrenz unsere Daten erklärt.

Zum anderen droht der *Fehlschluss der probabilistischen Falsifikation* selbst für die plausiblen Hypothesen. Darunter finden wir vor allem das Problem der *Überinterpretation der Daten bei kleinen Effekten*, das sich als schwere Belastung für die Praxis etwa in den Sozialwissenschaften und der Medizin erweist. Auch hier sollten wir uns fragen, ob die Daten intuitiv tatsächlich so stark für unsere Forschungshypothese sprechen. Hilfreich ist dafür sicherlich eine zusätzlich Auswertung der Daten anhand des *Bayesianismus* trotz des dadurch bedingten erhöhten formalen Aufwands und eine Bestimmung des Bayes-Faktors bzw. des Likelihoodquotienten, denen auch für den Theorienvergleich im Rahmen eines Schlusses auf die beste Erklärung eine besondere Bedeutung zukommt.

Insgesamt sollten also alle Verfahren zusammengenommen uns ein besseres Bild der Situation liefern. Wir können weiterhin die Signifikanztests einsetzen, sollten aber versuchen dazu die Wahrheitsquoten einzuschätzen. Wir sollten das mit einer bayesianischen Auswertung und den entsprechenden Nachher-Wahrscheinlichkeiten vergleichen, und für die erforderlichen Plausibilitätseinschätzungen sollten wir ebenfalls den Schluss auf die beste Erklärung einsetzen. Statt eines Algorithmus, der uns die zu akzeptierenden Hypothesen ausspuckt, kommen wir um eine komplexe Abwägung leider nicht herum. Wissenschaftlicher Fortschritt ist eben nicht auf einfache Weise (viele Daten erheben und einen Algorithmus darauf loslassen) zu erzielen, sondern eher mühsam zu erreichen. Das wird sich auch im siebten Kapitel bestätigen, in dem

es uns um das Aufdecken von Kausalzusammenhängen geht – was m.E. das Ziel praktisch aller empirischen Forschungen ist.

6.6 Schätzen

Neben den Hypothesentests beschäftigt sich die klassische Statistik noch mit den Schätzverfahren. Die Techniken des Schätzens von bestimmten Größen anhand der Daten erscheinen noch elementarer zu sein, als die des Hypothesentestens. Wir nehmen in einfachen Fällen praktisch nur die Daten und generieren bestimmte Hypothesen, die durch die Daten gestützt werden. Entdeckungs- und Rechtfertigungskontext scheinen zusammenzufallen. Man extrapoliert ganz im Sinne einer konservativen Induktion scheinbar voraussetzungslos auf bestimmte Hypothesen. Typischerweise versuchen wir anhand einer Stichprobe auf einen bestimmten Parameter einer Grundgesamtheit oder einen Wert eines Versuchsaufbaus zu schließen. Die Qualität der Schätzungen und den verbleibenden Spielraum geben wir dann gerne anhand eines Konfidenzintervalls an. Überhaupt werden die Konfidenzintervalle häufig als die besseren Auswertungsmethoden gegenüber den Hypothesentests genannt, zumal sie nicht erkennbar auf probabilistischen Falsifikationen beruhen.

6.6.1 Punktschätzungen

Wir möchten z.B. wissen, wie hoch der Anteil der CDU-Wähler in einer bestimmten Gruppe G (G für unsere Grundgesamtheit) ist. Bei G könnte es sich etwa um die Deutschen zwischen 50 und 70 Jahren handeln. Wenn wir nicht alle befragen können, ziehen wir eine möglichst repräsentative Stichprobe S aus G und befragen die. Nehmen wir einmal an, die würden alle ehrlich antworten, dann sollte die Quote der CDU-Wähler in der Stichprobe der Quote der CDU-Wähler in G entsprechen, damit die Stichprobe als repräsentativ angesehen werden kann. Oder wir möchten wissen, welche Wahrscheinlichkeit Kopf bei unserer Münze hat. Als naheliegenden Schätzwert nehmen wir in beiden Fällen einfach die relative Häufigkeit in der Stichprobe für die Wahrscheinlichkeit in G.

Beide Male setzen wir bestimmte Annahmen voraus für unsere Schätzungen. Wir nehmen etwa an, dass unsere n Daten durch Zufalls-

variablen $X_1, \ldots, X_n$ dargestellt werden können, die alle unabhängig voneinander sind und dieselbe Wahrscheinlichkeitsverteilung aufweisen. Im Falle der Münze erwarten wir etwa, dass es keine zeitlichen Entwicklungen gibt, wonach sich die Wahrscheinlichkeit von Kopf langsam verändert. Dann würden wir andere Schätzverfahren für die nächsten Würfe einsetzen. Für die Stichprobenziehung nehmen wir an, dass für jede Ziehung X_i dieselbe Wahrscheinlichkeit vorliegt, dass es sich um einen CDU-Wähler handelt, wie sie die relative Häufigkeit in G dafür vorgibt. (Frequentisten müssen die Wahrscheinlichkeitsaussagen dann noch in entsprechende relative Häufigkeiten oder Durchschnittswerte übersetzen.) Das erreicht man in der Praxis am ehesten durch ein geeignetes Stichprobenverfahren, bei dem etwa jedes Element aus G dieselbe Chance hat, gezogen zu werden. Außerdem muss die Anzahl N der Elemente der Menge G sehr viel größer sein als die Stichprobe, da sonst die gezogenen Elemente die Häufigkeit in G selbst schon erheblich beeinflussen könnten. Als Faustregel sollte gelten, dass N größer ist als $100 \cdot n$. Für andere Größen kommen oft noch weitere Annahmen hinzu, etwa dass die Größen normalverteilt sind oder Ähnliches.

Wir sehen also auch für die Punktschätzungen, dass sie nicht ohne gewisse theoretische Modellannahmen auskommen, die unser Hintergrundwissen bereitzustellen hat. Nehmen wir nun etwa an, der gesuchte Wert sei $\theta = P(\text{Kopf})$. Dann geben wir dazu eine Schätzfunktion $T(x_1, \ldots, x_n)$ an, die zu bestimmten Daten $x_1, \ldots, x_n$ einen Schätzwert bestimmt. In unserem einfachen Beispiel nehmen die Größen X_i etwa den Wert 1 für Kopf und den Wert 0 für Zahl an, weshalb wir dann ansetzen würden: $T(x_1, \ldots, x_n) = \frac{1}{n} \sum x_i$.

Die klassische Statistik nennt nun noch einige Anforderungen an Schätzfunktionen, die vernünftig erscheinen. Typischerweise verlangt man von einem Punktschätzer, dass er *erwartungstreu* ist. Damit ist gemeint, dass der Erwartungswert des Schätzers mit dem wahren Wert übereinstimmt. In unserem einfachen Beispiel ist das der Fall:

$$E(T(X_1, \ldots, X_n)) = E\left(\tfrac{1}{n} \sum X_i\right) = \tfrac{1}{n} \cdot n \cdot \theta = \theta$$

»Erwartungswert« klingt so nach epistemischen Wahrscheinlichkeiten, die vermutlich leichter zu interpretieren sind. Ein Frequentist meint

damit so etwas wie einen *Durchschnittswert auf lange Sicht* für die auftretenden relativen Häufigkeiten. (Die Probleme des Ansatzes hatten wir bereits in Kapitel 5 diskutiert.) Im Durchschnitt sollte also bei dem Schätzverfahren gerade der wahre Wert geschätzt werden. Das ist bei unseren Modellannahmen offensichtlich der Fall. Die Forderung klingt recht natürlich, wird m.E. jedoch erst dann für uns bedeutsam, wenn wir sie mit Hilfe epistemischer Wahrscheinlichkeiten interpretieren. Dann erst ließe ein erwartungstreuer Schätzer in unserem konkreten Fall tatsächlich *erwarten*, dass der Schätzer den richtigen Wert angibt. Eine Abweichung des Schätzers von der Erwartungstreue wird auch *Verzerrung* oder im englischen *Bias* genannt: $\text{Bias}(T) = E(T(X_1, \ldots, X_n)) - \theta$. Die strikte Erwartungstreue ist nicht immer zu gewährleisten. Dann verlangt man zumindest noch asymptotische Erwartungstreue oder Konsistenz, also eine Annäherung des Erwartungswertes an den wahren Wert zumindest für wachsendes n.

Es gibt noch ein anderes Kriterium für die Qualität eines Punktschätzers, nämlich die *Streuung der Stichprobenwerte*. Das ist vor allem spannend in den Fällen, in denen wir nicht nur binäre Variablen X_i haben, sondern auch quantitative Größen. Wenn wir etwa das Durchschnittseinkommen für das Anfangsgehalt der deutschen Ingenieure ermitteln möchten, so weisen die Gehälter bereits selbst einen Spielraum auf, der sich natürlich in der Stichprobe wiederfinden wird. Nehmen wir n=10 und in der Stichprobe einen Spielraum von 1000 bis 5000 Euro und einen Durchschnitt von 3000 Euro, dann liegen wir mit der Schätzung 3000 Euro vermutlich weniger nahe bei dem wahren Wert, als wenn alle Werte zwischen 2800 und 3200 Euro liegen. Zumindest spricht das zweite Ergebnis intuitiv viel eher dafür, dass wir mit unserer Schätzung nahe am korrekten Wert liegen, als das erste Ergebnis. Wenn die Streuung der Werte so groß ist, ist die Gefahr viel größer durch das Ziehen von ein paar Ausreißern, sich mit dem Mittelwert der Stichprobe vom Mittelwert der Grundgesamtheit zu entfernen. Also stellt eine möglichst kleine Streuung einen klaren Vorteil für unsere Schätzfunktion dar.

Wir kennen das aus dem Bereich der *Messung quantitativer Größen* in der Physik. Einige Physiker behaupten sogar, dass eine Messung ohne Fehlerrechnung wertlos sei. Wir messen dazu eine bestimmte Größe M mehrfach hintereinander und erhalten eine Reihe von Messwerten.

Deren Mittelwert stellt dann unseren Messwert dar und die Streuung liefert uns den Fehlerbalken. Die Streuung wird dabei anhand der Stichprobenstreuung geschätzt und man nimmt meist an, dass die Werte normalverteilt sind. Also im Prinzip bilden wir ein kleines Konfidenzintervall, auf das wir gleich noch zu sprechen kommen. Auf jeden Fall zeigt eine große Streuung an, dass unser Messverfahren nicht sehr zuverlässig ist (und etwa viele Störfaktoren vorliegen, die zu unserem starken Rauschen geführt haben) und wir den einzelnen Werten weniger vertrauen sollten als bei Messungen, die sich nur wenig unterscheiden. In diesem letzten Fall besteht eher die Gefahr einer Verzerrung. Sind aber beide Fehlertypen (Rauschen oder Verzerrung) gering, dürfen wir der Messung vertrauen.

Außerdem führen weitere Messungen dazu, dass sich die Streuung zumindest für unseren Mittelwert verringert. Für den Durchschnittswert $\bar{X} = \frac{1}{n}\sum X_i$ gilt, dass sich seine Streuung $\sigma(\bar{X})$ ergibt aus der Streuung in der Grundgesamtheit geteilt durch die Quadratwurzel der Anzahl der Versuche:

$$\sigma(\bar{X}) = \frac{\sigma(X)}{\sqrt{n}}$$

Das zeigt, weshalb sich das Wiederholen einer Messung lohnt. Die Streuung für die Durchschnittsvariable $\bar{X}$ (auch *Standardabweichung* genannt) nimmt dann nämlich langsam ab und wir nähern uns somit dem richtigen Wert, wenn wir unsere Messungen häufig wiederholen. Das gilt natürlich in entsprechender Weise für unsere Schätzgrößen T.

Zur Erinnerung sei noch einmal kurz an einige Grundbegriffe erinnert: Für eine Zufallsvariable X, die die Werte $x_1, \ldots, x_k$ annehmen kann, finden wir den Erwartungswert E(X), die Varianz Var(X) und die Streuung $\sigma(X) = \sqrt{\mathrm{Var}(X)}$ auf die folgende Weise:

Der **Erwartungswert**: $E(X) = \sum_{i=1}^{k} P(X = x_i) \cdot x_i$ und
Die **Varianz**: $\mathrm{Var}(X) = \sum_{i=1}^{k} P(X = x_i) \cdot (x_i - E(X))^2$
(für stetige Größen X_i sind die entsprechenden Integrale zu bilden)

Die genannten Zusammenhänge und die Rechenregeln für diese Operatoren finden sich praktisch in jedem Lehrbuch der Statistik und werden hier nicht vertieft. Für die Varianz summieren wir über die Quadrate

der Abstände vom Erwartungswert der Größe auf, das ist auch intuitiv gut verständlich als ein Maß für »Auseinanderfallen« der Werte, die X in seinen Anwendungen annimmt.

Was hat ein Frequentist aber gewonnen, wenn er z.B. zeigen kann (immer anhand seiner Modellannahmen!), dass ein Schätzverfahren T erwartungstreu ist? Wie schon erwähnt bedeutet das für ihn nur: Wenn er das Verfahren immer wieder (auf Versuchsreihen mit n Versuchen) anwendet, erhält er Werte, die im Durchschnitt mit dem wahren Wert p übereinstimmen. Die Frage bleibt, was wir im konkreten Einzelfall gewinnen bzw. erwarten dürfen, wenn wir das Schätzverfahren anwenden? Ohne eine Anwendung eines epistemischen Prinzips wie dem *statistischen Syllogismus*, können wir wieder einmal nichts über den Einzelfall sagen. Doch das ist es, was uns in den meisten Fällen interessieren dürfte. Spätestens an dieser Stelle sollte der Frequentist daher die rein frequentistischen Pfade verlassen und die relativen Häufigkeiten in epistemische Erwartungsgrößen oder ähnliche epistemische Konzepte übersetzen. Er kann dann zumindest darauf verweisen, dass diese epistemischen Wahrscheinlichkeiten auf objektiven frequentistischen Wahrscheinlichkeiten und Abschätzungen beruhen und nicht – wie bei den Bayesianern – selbst nur auf der Grundlage geschätzter subjektiver Wahrscheinlichkeiten entstanden sind.

Ein anderes Problem, das uns in ähnlicher Form schon mehrfach begegnet ist, findet sich an dieser Stelle wieder, nämlich dass es *zwei* Gütekriterien für die Schätzfunktionen gibt (Erwartungstreue und kleine Streuung), die nicht übereinstimmen müssen. Beide werden kombiniert in einem gemeinsamen Kriterium, das auch für sich intuitiv plausibel erscheint, nämlich dem *mittleren quadratischen Fehler* MSE (»mean-squared error«) der Schätzfunktion:

$$\mathrm{MSE}(T) = \mathrm{E}\left[(T - p)^2\right] = \mathrm{Var}(T) + \mathrm{Bias}(T)^2$$

Dabei wird wieder einmal der quadratische Abstand zwischen unserer Schätzung und dem wahren Wert als Maßstab für Wahrheitsnähe herangezogen. Auch hier könnte man natürlich die Debatte führen, ob es nicht andere Maße und eine möglicherweise andere Verrechnung denkbar sind. Aber das soll hier nicht weiter verfolgt werden. Wir hatten schon

in anderen Kontexten im Kapitel 5 gesehen, dass solche quadratischen Abstandsmaße sehr gebräuchlich und durchaus begründbar sind.

Stattdessen möchte ich noch eine andere Idee erwähnen, die besonders gut zur Idee des *abduktiven Schließens* passen. Das ist die *Maximum-Likelihood-Schätzung*. Dabei schätzen wir anhand der Daten, dass der gesuchte Parameter vorliegt, bei dem wir die maximale Likelihood für die Daten erhalten. Der ML-Schätzer muss nicht immer mit unserem bisherigen erwartungstreuen Mittelwertschätzer zusammenfallen. Schon für das Schätzen der Varianz einer Grundgesamtheit fallen sie auseinander. Der ML-Schätzer ist dann nicht mehr *strikt* erwartungstreu, aber zumindest noch *asymptotisch* erwartungstreu. Außerdem besitzt er andere bemerkenswerte Eigenschaften, wie z.B. eine Invarianz unter Transformationen der Zufallsvariablen, mit denen wir arbeiten. Außerdem kann man zeigen, dass die Varianz eines erwartungstreuen Schätzers nicht kleiner werden kann als die sogenannte Cramer-Rao-Schranke. Daher weiß man, dass erwartungstreue Schätzer, die diese Schranke erreichen, im Sinne unserer Gütekriterien *optimal* sind. Tatsächlich gilt beides für den ML-Schätzer für $n \to \infty$ (vgl. Held 2008, Kap. 3).

Betrachten wir kurz ein einfaches Beispiel: Nehmen wir an, wir hätten 53-mal Kopf bei 100 Würfen geworfen (E). Was sollte dann unser bester Tipp für $\theta = P(\text{Kopf})$ sein? Es ist naheliegend darauf mit 53/100 zu antworten, wie es auch die klassische Statistik vorschlägt. Aber was ist unsere *Rechtfertigung* für diese Vorgehensweise? Eine Form von abduktiver Begründung finden wir in dem Maximum-Likelihood-Verfahren. Wir betrachten die sich ergebenden Likelihoods für unterschiedliche Hypothesen H_θ und fragen uns, bei welchem Wert von θ wir die höchste Likelihood erhalten. Dazu bestimmen wir das Maximum der Likelihoodfunktion:

Likelihoodfunktion: $L(\theta) = P_\theta(E)$ $(=\text{Bin}(100;53;\theta))$

Hier kommen endlich einmal unsere Schulkenntnisse der Kurvendiskussion zum Tragen, um das Maximum auszurechnen. Tatsächlich ist für $\theta = 53/100$ das Maximum gegeben. Wir können das auch so beschreiben, dass die Hypothese $\theta = 53/100$ *ceteris paribus die beste Erklärung* für

unser Resultat von 53-mal Kopf bietet und daher gewählt werden sollte. Allerdings muss man natürlich zugeben, dass eng benachbarte Werte wie 52/100 fast genau so gute Erklärungen abliefern. Trotzdem gilt: der Maximalwert liegt gerade bei 53/100 und sollte daher gewählt werden, wenn sonst keine weiteren Hinweise vorliegen.

Punktschätzungen können wir also mit Hilfe des ML-Schätzers als Schlüsse auf die beste Erklärung verstehen. Es wird nach der Hypothese – in einer vorgegebenen Liste von Hypothesen, die etwa durch den Parameter θ charakterisiert werden – gesucht, die die beste Erklärung für unsere Daten darstellt. Schwieriger ist dann aber die Frage zu beantworten, wie wir dabei weiteres *Hintergrundwissen* einbringen können. Nehmen wir etwa an, dass unsere Münze gebogen aussieht und wir daher die Vermutung haben, dass Kopf häufiger erscheinen wird. Außerdem wissen wir natürlich, dass die Extremwerte θ=1 und θ=0 sehr unwahrscheinlich sind, wenn unsere Wurfexperimente nicht besonders präpariert wurden.

Bayesianische Punktschätzer. Solches Vorwissen kann erst der *Bayesianer* einbringen, der als Punktschätzer u.a. den Parameter mit der maximalen Likelihood der Nachher-Wahrscheinlichkeit wählen kann, was nur für eine nichtinformative Vorher-Wahrscheinlichkeit mit dem ML-Schätzer zusammenfällt. Ansonsten wird der Bayesianer das Schätzproblem als Entscheidungsproblem mit einer bestimmten *Verlustfunktion* l(a,θ) darstellen, für die es dann den zu erwartenden Verlust zu minimieren gilt. Dabei betrachten wir den Abstand unseres Schätzwerts a von dem wahren Wert p als einen entsprechenden Verlust, den es zu minimieren gilt. Allerdings sind nun wiederum unterschiedliche Abstandsfunktionen denkbar, die jeweils zu etwas anderen Schätzern führen (vgl. Held 2008, Kap. 5). Wählen wir etwa die quadratische Verlustfunktion $l(a,\theta) = (a-\theta)^2$, dann erhalten wir als Punktschätzer den Erwartungswert der Nachher-Wahrscheinlichkeitsverteilung. Zu minimieren ist der erwartete Verlust $E(l(a,\theta)|x) = \int_\Theta l(a,\theta) \cdot p(\theta|x)\, d\theta$, wobei x unser Beobachtungsergebnis darstellt und $p(\theta)|x)$ die Nachher-Wahrscheinlichkeitsdichte für die Hypothese H_θ. Dabei hatten wir schon gesehen, wie sich die Nachher-Dichte $p(\theta|x)$ anhand des Bayesschen Theorems aus der Vorher-Dichte $p(\theta)$ errechnen lässt:

Bayesianisches Updaten: $p^+(\theta) = p(\theta|x) = \dfrac{f(x|\theta) \cdot p(\theta)}{\int f(x|\theta) \cdot p(\theta)\, d\theta}$

Dann erhalten wir als Schätzer: $a = \int \theta \cdot p(\theta|x)\, d\theta$ (s. Held 2008, 157) den entsprechenden Nachher-Erwartungswert. Für andere Verlustfunktionen ergeben sich noch andere Punktschätzer:

Bayesianische Punktschätzer für bestimmte Verlustfunktionen

$l(a,\theta) = (a-\theta)^2$	der Erwartungswert der Nachher-Dichte $p(\theta\|x)$
$l(a,\theta) = \|a-\theta\|$	der Median der Nachher-Dichte $p(\theta\|x)$
$l(a,\theta) =$ 0-1-Funktion	der Modus der Nachher-Dichte $p(\theta\|x)$

Dabei soll die 0-1-Funktion in einer kleinen Umgebung des wahren Parameters 0 sein und außerhalb 1. Der Modus gibt gerade den Wert des Maximums der Nachher-Dichte an und ist für eine Gleichverteilung als Vorher-Dichte gleich dem ML-Schätzer (Genaueres s. Held 2008, Kap. 5). Interessant ist für uns u.a., dass es hier unterschiedliche Vorschläge gibt und die jeweils eine entsprechende epistemische Interpretation besitzen. Außerdem stellen sie die Grundlage für bayesianische Kredibilitätsintervalle dar, auf die wir noch zu sprechen kommen werden.

Unser einfaches Beispiel kann auch die unterschiedlichen Punktschätzer ein Stück weit erläutern. Nehmen wir an, dass wir wieder die Wahrscheinlichkeit θ für Kopf ermitteln möchten, anhand von n Würfen der Münze und x-mal Kopf. Wenn wir mit einer Vorher-Gleichverteilung starten, erhalten wir als den Nachher-Erwartungswert $E(\theta|x) = (x+1)/(x+2)$ gerade die Laplacesche Regel, die wir im ersten Kapitel und dann wieder als Schlussregel der induktiven Logik kennengelernt haben.

Als Nachher-Modus erhalten wir dagegen den ML-Schätzer x/n (vgl Held 2008, 147). Man sieht an dieser Stelle, dass der Erwartungswertschätzer immer noch ein wenig die Gleichverteilung berücksichtigt, die vor unserem Experiment angenommen wurde, aber sich mit wachsendem n dann doch dem einfachen ML-Schätzer annähert, der nur von den Daten bestimmt wird. Da stoßen zwei plausible Philosophien des induktiven Schließens aufeinander, die wir schon kennengelernt hatten, die aber hier noch einmal neu begründet werden, anhand unterschiedlicher Verlustfunktionen im Rahmen des bayesianischen Ansatzes.

6.6.2 Konfidenzintervalle

Spannend wird es aber vor allem, wenn wir nicht nur wie in den bisherigen Fällen eine Punktschätzung vornehmen, sondern stattdessen ein Intervall bestimmen möchten, in dem sich der gesuchte Parameter mit einer bestimmten Wahrscheinlichkeit befinden wird. Dadurch können wir zugleich etwas über die Genauigkeit der Schätzung aussagen. Im Falle der klassischen Statistik sind da vor allem die *Konfidenzintervalle* zu nennen, die allerdings anders konstruiert werden als die Kredibilitätsintervalle im Bayesianismus und auch nicht auf dieselbe Weise interpretiert werden können wie diese.

Dabei suchen wir nach einem Verfahren bzw. einer Funktion C, die zu jedem Datum x (und jedem n für die Anzahl der Einzeldaten) ein Interval C(x,n) liefert, in dem der gesuchte wahre Parameter θ mit der Wahrscheinlichkeit $1–\alpha$ enthalten ist. (Das »n« lassen wir im Folgenden weg, da wir alle Überlegungen zu einem festen n durchführen.) Dabei nennen wir $1–\alpha$ das *Konfidenzniveau* unseres Intervalls. Oft wählt man hier ein Niveau von 95% bzw. einer Irrtumswahrscheinlichkeit α von 5%. Da wir den wahren Parameter θ noch nicht kennen, muss die Funktion C so gestaltet sein, dass sie die Anforderung für jeden möglichen Wert $\theta \in \Theta$ erfüllt. Wir erhalten somit:

C(x) ist ein Konfidenzintervall zum Datum x und Niveau $1–\alpha$ gdw
$P_\theta(\theta \in C(x)) = P(\theta \in C(x)|\theta) \geq 1–\alpha$ für alle $\theta \in \Theta$

Dabei ist zu beachten, dass das Intervall bzw. die untere und die obere Intervallgrenze C(x) = [U(x),O(x)] die Zufallsvariable darstellen und nicht unser Parameter θ. Der kann nur im bayesianischen Ansatz mit einer epistemischen Wahrscheinlichkeitsdichte versehen werden. Die Behauptung ist also, dass wir mit unserem konkreten Verfahren zur Konstruktion des Konfidenzintervalls zu jedem x (etwa einer Folge von Köpfen und Zahlen in unserem Münzwurfbeispiel) ein Konfidenzinterval C(x) angeben können, das den wahren Parameter mit einer gewissen Vertrauenswahrscheinlichkeit $1–\alpha$ überdeckt. (Diese Redeweise ist allerdings noch nicht ganz genau, aber dazu später mehr.)

Das ist natürlich trivial erfüllbar, denn wir können einfach $C(x) = \Theta$ wählen. Diskutieren wir das Ganze nun etwa an unserem Münzwurfbeispiel, dann könnten wir also $C(x) = [0,1]$ wählen und wären natürlich auf der sicheren Seite. Der gesuchte Parameter liegt nach unseren Modellvorstellungen zu dieser Situation schließlich mit Wahrscheinlichkeit 1 in diesem Konfidenzintervall. Doch das Konfidenzintervall wäre offensichtlich nicht hilfreich, denn wir wüssten nicht mehr als vorher. Wir müssen hier wieder einen Deal eingehen. Wir verrechnen den Informationsgehalt mit einem gewissen Irrtumsrisiko. Je größer das Irrtumsrisiko, das wir einzugehen bereit sind, umso kleiner können wir unser Konfidenzintervall wählen, d.h. umso informativer sind unsere Resultate. Wir dürfen hier wieder nicht nur nach Sicherheit streben, sonst erhalten wir nur triviale Konfidenzintervalle

Wie kann das für nichttriviale Intervalle funktionieren? Die Idee ist m.E. gut nachvollziehbar. Wir haben für jedes θ eine konkrete Likelihood für ein bestimmtes festes Datum x: $P(x|\theta)$. Wenn diese Wahrscheinlichkeit für ein bestimmtes θ^* nun unter α fällt, so müssen wir für unser x diesen Modellparameter θ^* nicht in unser Konfidenzintervall $C(x)$ aufnehmen, da er nur in so wenigen Fällen zu x führt, dass wir das als Irrtumsfälle unseres Verfahrens erlauben. Betrachten wir dazu den Fall $\alpha = 0{,}05$ und 100 Fälle in denen wir unser Experiment (mit jeweils n Würfen) wiederholen. Wäre θ^* nämlich der wahre Parameter, dann würde (im Durchschnitt) in höchstens 5 Fällen das Resultat x auftreten. Wenn wir also θ^* aus $C(x)$ herauslassen, ergibt das (im Durchschnitt) auch höchstens 5-mal den Fall, dass unser Verfahren ein $C(x)$ liefert, $C(x)$ aber den wahren Wert nicht überdeckt. Das liegt dann im Rahmen unseres erlaubten Irrtumsrisikos und wird somit in Kauf genommen.

Machen wir es etwas konkreter: Nehmen wir an, dass wir eine relative Häufigkeit von 0,4 (für Kopf) in unserem Experiment (mit Resultat x) erzielt haben. Nehmen wir außerdem an, n wäre groß genug, dass unser Intervall gerade $C(x) = [0{,}3;0{,}5]$ sei. Dann behaupten wir damit, dass ein Parameterwert wie etwa $\theta = 0{,}1$ nicht in unser Intervall $C(x)$ aufgenommen werden muss, weil bei diesem Wert nur so selten, das Ergebnis x aufträte, dass der begangene Fehler verschmerzbar sei, bzw. eben unterhalb von unserem Irrtumsrisiko liegt. Wir würden diesen Fehler nämlich (im Durchschnitt) in höchstens 5 der 100 Fälle begehen.

Das gibt schon erste Hinweise darauf, wie wir Konfidenzintervalle konstruieren können. Allerdings gibt es unterschiedliche Verfahren, die auch zu unterschiedlichen Ergebnissen führen können (vgl. Held 2008, Kap. 3). Einfache Beispiele werden wir uns noch anschauen.

Die Idee ist jedenfalls simpel: Wir nehmen an, dass in den meisten Fällen, die relative Häufigkeit der Köpfe nahe bei den wahren Werten θ der Wahrscheinlichkeit für Kopf liegt. Wenn wir also ein Intervall um diese relative Häufigkeit herum konstruieren, wird dieses Intervall auch in den *meisten Fällen* den wahren Wert θ überdecken. Das müssen wir nur noch quantifizieren und erhalten die Konfidenzintervalle.

Nehmen wir nun an, dass wir ein konkretes Konfidenzintervall C(x) wie das oben genannte C(x) = [0,3;0,5] für ein bestimmtes Resultat x vorliegen haben. Was haben wir damit gewonnen? Dürfen wir also nun *annehmen* bzw. *erwarten*, dass mit einer Wahrscheinlichkeit von 95% die wahre Wahrscheinlichkeit von Kopf in diesem Intervall liegt? Das können wir offensichtlich nicht. Die genannte 95%-Wahrscheinlichkeit wäre in jedem Fall eine *epistemische* Wahrscheinlichkeit, denn der wahre Modellparameter liegt schließlich fest. Er ist mit der objektiven Wahrscheinlichkeit 1 oder 0 in C(x), wir wissen nur nicht, welche der Optionen die richtige ist. Die 95%-Wahrscheinlichkeit könnte also nur unsere Unkenntnis beschreiben, aber solche epistemischen Wahrscheinlichkeiten lässt der klassische Statistiker nicht zu. Er kann nur sagen, dass unser Verfahren zur Konstruktion von Konfidenzintervallen, das wir auch in dem konkreten Fall angewandt haben, *im Durchschnitt in 95% der Fälle den wahren Parameter überdecken wird*. Die 95% geben uns wieder nur eine Eigenschaft des Verfahrens an, wie das schon für die Filteranalogie für Signifikanztests der Fall war. Wir können uns nun wiederum überlegen, wie wir diese Informationen womöglich weiter auswerten können.

Sollten wir denn in einem konkreten Fall erwarten, dass der wahre Parameter zwischen 0,3 und 0,5 liegt? Der klassische Statistiker kann dazu keine weiteren Auskünfte geben. Im Kapitel 5 haben wir bereits darüber diskutiert, dass es diese Interpretationsprobleme für die Ergebnisse gibt. Es ist nicht zu sehen, wie uns der klassische Statistiker helfen kann, eine Information über unseren Einzelfall zu geben, wenn er nicht über seinen Schatten springt und die relativen Häufigkeiten,

die er uns angibt, in irgendeiner Form auch in Erwartungen ummünzt. Und natürlich haben wir den Eindruck, dass die relativen Häufigkeiten sehr wohl etwas über den Einzelfall aussagen. Doch der Schritt hin zu bestimmten Erwartungen verlangt, dass wir wiederum das Grundprinzip der induktiven Logik, nämlich den *statistischen Syllogismus*, mit heranziehen. Erst wenn wir den einsetzen, gelangen wir zu der erwünschten Interpretation, dass der gesuchte Parameter mit einer Wahrscheinlichkeit von 95% vom Konfidenzintervall überdeckt wird.

Das kann allerdings die Frage aufwerfen, warum wir dann nicht gleich zu bayesianischen Verfahren gegriffen haben. Ein klassische Statistiker könnte darauf z.B. erwidern, dass er nur an solchen Stellen bereit ist, zu epistemischen Wahrscheinlichkeiten zu greifen, an denen diese durch *bekannte relative Häufigkeiten begründet* sind. Daher kommen epistemische Wahrscheinlichkeiten für ihn jedenfalls als Dichten für den Modellparameter θ nicht in Frage. Es bleibt also noch ein wesentlicher Unterschied zwischen einem frequentistischen Vorgehen, das nur am Schluss die Häufigkeiten auch epistemisch interpretiert, und einem komplett bayesianischen Ansatz. Das hat mich auch motiviert, die Filteranalogie zu entwickeln, die zunächst rein frequentistisch vorgeht und nur die Ergebnisse in Form bestimmter relativer Häufigkeiten zum Schluss epistemisch interpretiert. Eine solche Hybridposition scheint mir durchaus begründbar und damit vertretbar zu sein. Ohne diesen letzten Schritt kann uns der Frequentist allerdings tatsächlich keine befriedigende Antwort geben, was wir aus den Ergebnissen seines Verfahren für unseren konkreten Anwendungsfall lernen können.

Wie können wir derartige Intervalle nun auffinden? Typischerweise beginnen wir mit einem Punktschätzer und konstruieren darum herum (oft symmetrisch) ein Konfidenzintervall. Dazu ist wiederum eine Testgröße hilfreich, für die wir eine sogenannte Pivot-Verteilung kennen, das ist eine Verteilung, die selbst nicht mehr vom gesuchten Parameter abhängt. Das ermöglicht es uns, ein Intervall zu finden, das die Anforderungen für alle Werte des gesuchten Modellparameters erfüllt. Das lässt sich am besten an einem Beispiel aufzeigen. Dazu knüpfen wir an eine Verteilung und eine Teststatistik an, die wir im Kapitel 6.3 im Rahmen des Beispiels für einen t-Test bereits kennengelernt haben.

Nehmen wir z.B. an, unsere grundlegenden Zufallsvariablen $X_1,\ldots,X_n$ (die unsere Messergebnisse für eine bestimmte Größe wie etwa die Schulleistung angeben) sind alle normalverteilt zu denselben Parametern μ (Mittelwert) und σ^2 (der Varianz), die uns aber beide unbekannt sind. Wir suchen μ und werden dazu ein Konfidenzintervall um den Mittelwert $\bar{X}$ unserer Stichprobe herum konstruieren. Zu diesen $N(\mu,\sigma^2)$ verteilten Zufallsvariablen bilden wir dann die neue Zufallsvariable

$$(1)\quad T = \sqrt{n}\,\frac{\bar{X}-\mu}{S} \sim t(n-1),$$

von der uns die Statistik lehrt, dass sie standard-t-verteilt (s.o. bei den Hypothesentests) mit n–1 Freiheitsgraden ist. Das ist also die gesuchte Pivot-Verteilung, die uns dann schnell zu dem entsprechenden Konfidenzintervall führt.

Wir verstehen nach der Debatte um Punktschätzungen die Größen in diesem Ausdruck auch schon besser: $\bar{X}$ ist ein unverfälschter Schätzer für μ und $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$ ein unverfälschter Schätzer für σ. Da wir die Varianz σ auch nicht kennen, sondern schätzen müssen, sind wir aber auf die t-Verteilung (statt der Normalverteilung) angewiesen. Nun erhalten wir jedenfalls für den Schätzer T die gesuchte Ungleichung:

$$(2)\quad P[t_{n-1;\alpha/2} \le T \le t_{n-1;1-\alpha/2}] = 1-\alpha$$

Dabei ist $t_{n-1;\alpha}$ das α-Quantil der Standard-t-Verteilung, d.h. der Wert für den in der t-Verteilung gerade gilt: $P(T \le t_{n-1;\alpha}) = \alpha$. Da die t-Verteilung symmetrisch zur Null ist, finden wir außerdem für den anderen Wert: $t^* := -t_{n-1;\alpha/2} = t_{n-1;1-\alpha/2}$, denn beide Quantilswerte sind gleich weit von der Null entfernt. Daraus gewinnen wir nun leicht unser Konfidenzintervall, indem wir einfach die Werte für T aus (1) in die Ungleichung einsetzen und dann (2) nach μ in der Mitte auflösen:

$$(3)\quad P[-t^* \le T \le t^*] = 1-\alpha = P[\bar{X}-t^*\cdot\tfrac{S}{\sqrt{n}} \le \mu \le \bar{X}+t^*\cdot\tfrac{S}{\sqrt{n}}]$$

also ist unser Konfidenzintervall C(x,n) auf dem Niveau $1-\alpha$ zum Datum x für μ demnach (vgl. dazu Held 2008, 58):

$$C(x,n) = [\bar{X}-t^*\cdot\tfrac{S}{\sqrt{n}},\ \bar{X}+t^*\cdot\tfrac{S}{\sqrt{n}}]$$

Da t* nur von n und α abhängt, liefert das ein allgemeines Verfahren zur Bildung eines Konfidenzintervalls für die angegebene Situation. Für größere n oder bei bekannter Streuung können wir statt der Quantile für die t-Verteilung auch noch auf die der Standardnormalverteilung zurückgreifen.

Für die Normalverteilung gibt es einfache Regeln, die uns sehr schnell eine erste Idee vermitteln, wie groß die Konfidenzintervalle sind, bzw. wie wir auf sehr einfache Weise solche Intervalle konstruieren können. Für die Normalverteilung gilt nämlich:

Die σ-Regeln für die Normalverteilung
$P(|Z-\mu| \leq \sigma) \approx 0{,}68$
$P(|Z-\mu| \leq 1{,}96 \cdot \sigma) \approx 0{,}95$
$P(|Z-\mu| \leq 2\sigma) \approx 0{,}955$
$P(|Z-\mu| \leq 3\sigma) \approx 0{,}997$

Bei einer Normalverteilung unserer Stichprobenwerte sind also im Abstand von einer Standardabweichung vom Mittelwert bereits 2/3 der Werte zu finden, für zwei Standardabweichungen bereits 95% der Werte und in einer Umgebung von drei Standardabweichungen werden fast alle Werte erfasst. Physiker nehmen als Fehlerbalken für ihre Messergebnisse oft eine Standardabweichung, die anhand der beobachteten Stichprobenstreuung geschätzt wird. Für ein Konfidenzniveau von 95% wäre unser Konfidenzintervall für die Normalverteilung also $C(x,n) = [\bar{X} - 1{,}96 \cdot \frac{S}{\sqrt{n}}, \bar{X} + 1{,}96 \cdot \frac{S}{\sqrt{n}}]$. Für die t-Verteilung ist der Wert für t* etwas größer in Abhängigkeit von n:

n	=	1	5	10	20	30	60	120	∞
$t_{n;1-\alpha/2}$	=	12,7	2,57	2,23	2,09	2,04	2,00	1,98	1,96

Als Daumenregel nimmt man gern schon die Normalverteilung ab n=30 als brauchbare Approximation. Jedenfalls, wenn man sich die σ-Regeln merkt, hat man eine intuitive Vorstellung davon, wie solche Verteilungen aussehen und wie groß die Konfidenzintervalle ungefähr werden.

Schauen wir uns noch ein weiteres Beispiel an, nämlich unsere Schätzung der Größe θ=P(Kopf) bzw. eines *Anteils* einer Grundgesamtheit aus der wir eine Stichprobe gezogen haben. Wählen wir als Testgröße X

die Anzahl der Köpfe in der Stichprobe, dann ist X Bin(n,θ) verteilt, nur dass wir θ nicht kennen. Als Schätzer wählen wir dann $\hat{\theta} = X/n$. Für den Schätzer ergibt sich so die Varianz, die wir aus der Binomialverteilung kennen, geteilt durch die Anzahl der Versuche:

$$\mathrm{Var}(\hat{\theta}) = \frac{\theta(1-\theta)}{n}$$

Wenn wir darin θ durch $\hat{\theta}$ schätzen und bedenken, dass sich die Binomialverteilung durch die Normalverteilung approximieren lässt, erhalten wir nach demselben Verfahren wie oben das sogenannte Wald-Intervall für θ (vgl. Held 2008, 60) als approximatives Konfidenzintervall für größere n:

$$C(x,n) = \left[\hat{\theta} - z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}},\ \hat{\theta} + z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right],$$

wobei $z_{1-\alpha/2}$ das entsprechende 1–α/2-Quantil der Standardnormalverteilung darstellt, das wir zum Konfidenzniveau 95% mit 1,96 bereits kennengelernt haben.

Held (2008, Kap. 3) betont aber auch, dass wir andere Verfahren zur Konstruktion von Konfidenzintervallen einsetzen können, insbesondere auch Verfahren, um exakte Konfidenzintervalle zu bilden. Die durchaus unterschiedlichen Ergebnisse werden in einer Tabelle (Held 2008, 95) miteinander verglichen. Analysen zeigen z.B., dass das gern genannte (approximative) Wald-Intervall oft eine etwas geringere Überdeckungswahrscheinlichkeit aufweist als die angestrebten 1–α.

Ein klassisches Verfahren um ein Konfidenzintervall für einen Anteil zu finden, soll noch kurz erwähnt werden, da es ohne entsprechende Approximationen auskommt und daher eher auf der sicheren Seite angesiedelt ist. Es ist nach Clopper und Pearson benannt und bestimmt die Intervallgrenzen praktisch »zu Fuß«. In unserem Beispiel sei unser Datum x=k (der Wert zur Zufallsvariable X) die Anzahl der Köpfe und es gelte: $0 < k < n$. Wir suchen nun nach einem Intervall $C(x) = [p_u, p_o]$ um k/n herum, das die Anforderungen an Konfidenzintervalle erfüllt und möglichst unseren gesuchten Modellparameter θ=P(Kopf) enthalten soll. Dazu verteilen wir die erlaubte Irrtumswahrscheinlichkeit auf beide Seiten von k/n und ermitteln dann die beiden Intervallgrenzen anhand der Formeln:

Intervallgrenzen: (1) $P(X \geq k|p_u) = \frac{\alpha}{2}$ und (2) $P(X \leq k|p_o) = \frac{\alpha}{2}$

Das lässt sich so verstehen: Wenn wir die untere Intervallgrenze p_u festlegen, dann dürfen wir damit nur die Werte für θ ausschließen, für die das Ergebnis k-mal Kopf oder ein noch ungünstigeres Ergebnis höchstens eine Wahrscheinlichkeit von $\alpha/2$ hat. Das ist der Irrtumsspielraum, den wir auf jeder Seite akzeptieren. Liegt also das tatsächliche θ noch unterhalb von p_u, so haben wir uns zwar geirrt, aber die Festlegung der Intervallgrenzen sorgt dafür, dass dieser Irrtum im Durchschnitt nur seltenen auftritt, weil p dann so klein ist, dass k oder größere Werte gemäß der Binomialverteilung dann nicht mehr zu erwarten sind.

Entsprechend wird die obere Grenze festgelegt. Dabei wird das Konfidenzintervall meistens nicht symmetrisch um k/n ausfallen, da wir nur die Irrtumswahrscheinlichkeit symmetrisch verteilt haben, aber nicht die Abstände im Intervall. Die Gleichungen lassen sich mit Hilfe der Betaverteilung auch analytisch auflösen, aber die technischen Details muss ich den Statistiklehrbüchern überlassen. Im Prinzip kann man das auch mit Hilfe einer Tabellenkalkulation durch Ausprobieren herausfinden. Wir sehen uns gleich ein Beispiel dafür an.

Wir könnten natürlich auch die Irrtumswahrscheinlichkeit α anders auf die beiden Seiten verteilen. Das kann sinnvoll sein, wenn eine Abweichung in der einen Richtung für uns aus praktischen Gründen bedeutsamer ist. Ein Hersteller von Airbags möchte etwa wissen, wie hoch die Ausfallrate θ seiner Airbags ist. Die soll z.B. 1 Prozent nicht überschreiten, wegen der Konventionalstrafen, die er dann etwa an einen Automobilhersteller zahlen muss und natürlich auch wegen der Menschen. Er testet eine Stichprobe seiner Produktion und erhält k von n fehlerhafte Airbags. Dann sollte er vermutlich nur einen kleinen Teil seiner Irrtumswahrscheinlichkeit hin zu höheren Ausfallszahlen zulassen, während die zu geringeren Ausfallzahlen beliebig groß sein kann. Überhaupt bietet sich dann ein einseitiges Konfidenzintervall an, für das wir außerdem das Konfidenzniveau entsprechend erhöhen. Derartige Probleme müssen in der Anwendungspraxis jeweils weiter diskutiert werden.

Schauen wir uns noch ein konkretes *Zahlenbeispiel* an: Als Datum haben wir nun 120-mal Kopf von 200 Würfen und suchen nun nach

einem passenden Intervall zum Konfidenzniveau 95% $C(x) = [p_u(x), p_o(x)]$ dazu (wobei x die verschiedenen Werte der Zufallsvariablen X darstellt, die uns die Anzahl der Köpfe angibt), mit den folgenden Eigenschaften:

Wir können so ein Konfidenzintervall nun relativ direkt berechnen, indem wir die Werte in unsere zwei Ungleichungen (vgl. a. Büchter & Henn 2007, 422ff.), für die wir die Irrtumswahrscheinlichkeit in zwei gleiche Bereiche aufteilen, wie wir das eben vorgeschlagen haben. Das ist also unser neuer Weg zu einem Konfidenzintervall:

(1) $P(X \geq 120|p_u) = 0,025$ und

(2) $P(X \leq 120|p_o) = 0,025$ und $p_u \leq 120/200 = 0,6 \leq p_o$

Für diese Ungleichungen suchen wir dann nach einem möglichst großen Wert für $p_u(x)$ und nach einem möglichst kleinen für $p_o(x)$, so dass die Ungleichungen noch erfüllt sind. Dabei verlangt (1), dass die Wahrscheinlichkeit für den Irrtum klein bleibt, der sich ergibt, wenn p unterhalb des Konfidenzintervalls liegt. Da unser Datum m aufgetreten ist, müssen wir hier also mit m oder noch entfernteren Daten als möglichen Irrtümern rechnen. Entsprechendes gilt für (2) nur in der anderen Richtung. Daher suchen wir jeweils nach entsprechenden Parametern p mit:

(1*) $\sum_{i=120}^{200} \binom{200}{i} p_u^i \cdot (1-p_u)^{200-i} \leq 0,025$ und

(2*) $\sum_{i=0}^{120} \binom{200}{i} p_o^i \cdot (1-p_o)^{200-i} \leq 0,025$

Wenn wir zwei Stellen hinter dem Komma betrachten, ergibt sich als größter unterer Wert 0,53 (in 1*) und als kleinster oberer Wert 0,67 (in 2*). Dabei wurde jeweils ein wenig verschenkt, d.h. die tatsächliche Irrtumswahrscheinlichkeit dieses Intervalls C(x) = [0,53;0,67] ist etwas kleiner als die zugestandenen 5%.

Auf diesem Weg können wir auch bestimmte Hypothesen testen. Sagt unsere Nullhypothese etwa θ habe einen bestimmten Wert r und $r \notin C(x)$, dann können wir die Nullhypothese als widerlegt betrachten. Natürlich finden sich für diese Interpretation der Konfidenzintervalle dieselben Probleme, wie wir sie für direkte Hypothesentests kennengelernt haben.

Somit haben wir ein weiteres induktives Schlussverfahren kennengelernt, das zumindest auf den Fall angewandt werden kann, in dem

man noch einige Parameter bestimmen muss, aber ansonsten schon über ein theoretisches Modell für eine bestimmte Situation verfügt. Dazu hatten wir gesehen, dass eine Interpretation, die auch etwas über den einzelnen Fall eines konkreten Konfidenzintervalls aussagen kann, bereits auf epistemische Wahrscheinlichkeiten angewiesen ist, die der klassische Statistiker eigentlich ablehnt. Die Frage bleibt also, ob dann nicht doch der Bayesianismus bzw. die bayesianische Statistik gleich die bessere Wahl darstellt, doch das müssen wir der weiteren Debatte überlassen. Zumindest hat sich auch gezeigt, dass wir immer auf eine Reihe von idealisierenden Annahmen angewiesen sind, um so ein statistisches Modell der Situation zu konstruieren, mit dessen Hilfe wir dann ein Konfidenzintervall bestimmen können. Selbst in diesen einfachen Situationen gibt es also keine automatischen Verfahren, die ohne weiteres Hintergrundwissen zu induktiven Erkenntnissen führen.

Bayesianische Kredibilitätsintervalle. Bayesianer haben einen relativ einfachen Weg zu den sogenannten *Kredibilitätsintervallen*, die das bayesianische Gegenstück zu den Konfidenzintervallen darstellen. Dazu legen sie zunächst für den gesuchten Parameter θ eine Dichtefunktion $p(\theta)$ fest, die angibt, in welchem Bereich gemäß unserem bisherigen Hintergrundwissen der Parameter θ vermutlich zu finden sein wird. Kommen bisher alle Werte gleichermaßen in Frage, wird man mit einer möglichst informationsarmen Verteilung also etwa einer Gleichverteilung starten. Gewinnen wir dann neue Daten E werden wir mit Hilfe des bayesschen Theorems mit diesen Daten Updaten und erhalten die neue Dichte:

$$\textbf{Nachher-Dichte}: \; p(\theta|x) = \frac{f(x|\theta) \cdot p(\theta)}{\int f(x|\theta) \cdot p(\theta)\, d\theta}$$

Diese neue Dichte setzen wir nun ein, um ein bayesianisches Kredibilitätsintervall $B(x) = [t_u, t_o]$ für den gesuchten Parameter θ zu finden, das die folgende Eigenschaft aufweist:

$$\textbf{Eigenschaft eines Kredibilitätsintervalls}: \; \int_{t_u}^{t_o} p(\theta|x)\, d\theta = 1 - \alpha$$

Das heißt, wir setzen das Intervall $B(x)$ so an, dass wir nun erwarten dürfen, dass sich der gesuchte Parameter mit der Wahrscheinlichkeit

$1-\alpha$ in dem Intervall befindet. Dafür kann es natürlich wiederum viele unterschiedliche Intervalle B(x) geben, die alle diese Eigenschaft aufweisen, ähnlich wie im klassischen Fall. Eine Idee, ein solches Intervall zu finden, ist etwa, auf beiden Seiten gerade so viel der Nachher-Dichte abzuschneiden, dass auf jeder Seite dieselbe Masse $\alpha/2$ außerhalb des Intervalls angesiedelt ist. Solche Intervalle nennt man *gleichendig* (vgl. Held 2008, 144). Der plausiblere Weg ist sicherlich, ein sogenanntes HPD-Intervall (mit den höchsten Nachher-Dichten) zu suchen, für das die Werte im Inneren eine höhere Dichte aufweisen, als die außerhalb:

$$\forall\theta \in B(x),\ \forall\theta^* \notin B(x):\ p(\theta|x) \geq p(\theta^*|x)$$

Dadurch werden tatsächlich die wahrscheinlichsten Modellparameter in unser Kredibilitätsintervall aufgenommen und, es wird oft sogar eindeutig bestimmt sein. Außerdem entspricht es dem Schließen auf die beste Erklärung. Zumindest suchen wir dann nach einer Menge von Hypothesen, die jeweils bessere oder gleichgute Erklärungen darstellen, wie alle Erklärungen außerhalb von B(x). Dabei haben wir wiederum die jeweiligen Likelihoods der Daten im Lichte der Hypothesen als Ceteriparibus-Maß für unsere Erklärungsstärke hergenommen. Natürlich kann es für das abduktive Schließen auch noch andere Faktoren geben, die dann ebenfalls zu berücksichtigen wären.

Ein besonderer Vorteil der bayesianischen Intervalle ist offensichtlich, dass ihre Interpretation schon vorgegeben ist: B(x) hat eine (epistemische) Wahrscheinlichkeit von 1–α, dass der gesuchte Modellparameter θ in B(x) enthalten ist. Außerdem können wir weiteres Hintergrundwissen in die Konstruktion einbringen.

Probleme von Konfidenz- und Kedibilitätsintervallen. Es gibt leider bisher nur wenige wissenschaftstheoretische Debatten in den Wissenschaften und in der Wissenschaftstheorie über die Qualitäten der Anwendung von Konfidenzintervallen. Erste Überlegungen haben wir nun angestellt.

Zunächst einmal können wir die Verfahren für Konfidenzintervalle nicht ohne gewisse *Voraussetzungen* oder *Annahmen* anwenden. Im Münzwurfbeispiel müssen wir etwa annehmen, dass die Münze eine feste Einzelfalltendenz p (im Sinne einer Propensität p) hat, dass Kopf

kommt, die sich nicht im Laufe der Zeit verändert und dass die Würfe auch unabhängig voneinander sind, so dass wir für jeden unserer Würfe annehmen dürfen, dass immer wieder dieselbe Tendenz p wirksam ist. Erst dadurch wird unser Messverfahren sinnvoll, die Messung von p zu verbessern, indem wir die Messung mehrfach wiederholen. Nur so erhalten wir ein statistisches Modell (etwa ein Urnenmodell für unsere Situation), mit dem wir dann rechnen und Schließen dürfen. Für das Stichprobenziehen sind natürlich ähnliche Annahmen erforderlich, wenn wir aus den Stichproben etwas über die Grundgesamtheit lernen möchten. In anderen Fällen sind noch weitere Annahmen einzubringen, etwa darüber, dass der gesuchte Wert normalverteilt ist.

Damit unsere Punktschätzung und die entsprechende Intervallschätzung dann auch etwas über den Einzelfall aussagen, sind wir letztlich immer wieder auf den *statistischen Syllogismus* angewiesen. Unser statistisches Modell sagt, dass in den meisten Fällen (hier geht es um relative Häufigkeiten) die zu beobachtenden Durchschnittswerte nahe bei den wahren Werten liegen. Das drehen wir zunächst um und schließen, dass in den meisten Fällen, die wahren Werte in der Nähe der von uns beobachteten relativen Häufigkeiten liegen werden. Wenn wir nun den statistischen Syllogismus anwenden, dürfen annehmen, dass der gesuchte wahre Wert mit hoher epistemischer Wahrscheinlichkeit (also ziemlich sicher) auch in unserem Einzelfall in einer Umgebung unserer Punktschätzung liegt. Erst dadurch erhalten wir eine relevante Information für den Einzelfall.

Howson & Urbach (1996, Kap. 10.c) überlegen an dieser Stelle, ob wir aus den klassischen Ergebnissen vielleicht auch anhand des *Hauptprinzips* schließen dürfen, das wir im 5. Kapitel kennengelernt haben. Sie deuten die Wahrscheinlichkeit $P(\theta \in C(x)) = 1–\alpha$ als objektive Einzelfallwahrscheinlichkeit (und nicht im Sinne einer relativen Häufigkeiteninterpretation) und wenden darauf dann das Hauptprinzip an, wonach unsere epistemische Wahrscheinlichkeit ebenfalls $1–\alpha$ sein sollte, dass unser Intervall den wahren Parameter überdeckt. Das Ziel dabei ist wiederum, eine relevante Auskunft für den Einzelfall zu gewinnen, die uns das Konfidenzintervall liefert.

Doch das führt in Probleme wie Howson und Urbach aufzeigen. Eine prinzipielle Schwierigkeit ist dabei, dass die Konfidenzintervalle nur

für Situationen ohne Hintergrundwissen zu sinnvollen epistemischen Wahrscheinlichkeiten führen. Das Hauptprinzip ist aber so stark, dass es für objektive Einzelfallwahrscheinlichkeiten Ableitungen epistemischer Wahrscheinlichkeiten sogar für den Fall weiteren Hintergrundwissens erlaubt. Die Autoren zeigen, dass das in konkreten Fällen zu unplausiblen Ergebnissen führt. Der statistische Syllogismus passt hier besser zur klassischen Interpretation von Wahrscheinlichkeiten als relativen Häufigkeiten und kann auch nur in den Fällen angewendet werden, in denen wir über kein weiteres relevantes Hintergrundwissen verfügen. Das vermeidet Fehler in der Interpretation der Intervalle.

Morey et al. (2015) konstruieren dazu einige Beispiele, in denen wir etwa nach einem Unterseeboot suchen und dafür ein Konfidenzintervall bilden, das aber deshalb irreführend ist, weil wir bestimmtes Hintergrundwissen haben, das in der Konstruktion der Konfidenzintervalle nicht eingeplant werden kann. Besser sieht das für die bayesianischen Kredibilitätsintervalle aus. Dort können wir solches Hintergrundwissen in die Vorher-Wahrscheinlichkeiten einbringen und damit für die Konstruktion der Intervalle berücksichtigen.

Die Frage ist dann, ob wir nicht sogar in vielen Fällen ein bestimmtes Vorwissen besitzen, dass wir für unsere Schätzungen beachten sollten. Für den Münzwurf wussten wir schon, dass die extremen Werte p=0 oder p=1 und benachbarte Werte sehr unwahrscheinlich sind. Vielleicht haben wir in anderen Fällen noch weiteres Wissen, das wir einbeziehen müssen. Dann scheinen Konfidenzintervalle nicht der richtige Weg zu sein, die gesuchten Größen zu schätzen. Wir sollten vielmehr solche Verfahren anwenden, die das vorliegenden Hintergrundwissen berücksichtigen können.

Wir können allerdings auch versuchen, die *Spielräume* für die Einbeziehung von weiterem Hintergrundwissen zu nutzen, die sich bei der Konstruktion von Konfidenzintervallen aufgezeigt haben. Wir müssen die Intervalle z.B. nicht symmetrisch um den Punktschätzer herum legen. Ist unsere relative Häufigkeit bei 10 Münzwürfen etwa 0,9, können wir das Konfidenzintervall nach unten hin größer gestalten, da wir schon wissen, dass extremere Werte als 0,9 a priori unwahrscheinlich sind. Doch damit würden wir natürlich bereits das Verfahren zur Konstruktion von Konfidenzintervallen ändern. Wir würden uns einen Schritt den

bayesianischen Verfahren annähern und die ganze Konstruktion würde aus Sicht der klassischen Statistik dadurch »subjektiver«.

Vielleicht sollten wir dann gleich auf die bayesianischen Verfahren setzen, da für die schon besser geklärt ist, wie das Vorwissen mit unseren Daten verrechnet werden kann und wie wir die Ergebnisse interpretieren dürfen. Allerdings sind wir dann zwingend darauf angewiesen, konkrete Vorher-Dichten anzugeben und das ist ebenfalls kein leichtes Unterfangen. Für welche genaue Vorher-Dichte spricht z.B. unser Vorwissen im Münzwurfbeispiel? Kaum jemand wird dafür eine präzise Funktion angeben können. Die Debatte um Konfindenzintervalle hat wohl gerade erst begonnen und wird sicher fortzusetzen sein.

7 Kausaltheorien und Kausalschlüsse

7.1 Einleitung: Zur Bedeutung kausalen Schließens

Ein zentrales Ziel der wissenschaftlichen Arbeit ist es, bestimmte Phänomene *erklären* oder *vorhersagen* zu können. Dazu sind die empirischen Disziplinen vor allem auf Wissen über kausale Zusammenhänge angewiesen. Nur *kausales Wissen* ist letztlich geeignet, Phänomene erklären zu können, Ereignisse vorherzusagen oder in das Geschehen eingreifen zu können. Es nützt dagegen noch nicht viel zu wissen, dass gelbe Finger und langjähriges Rauchen eng mit dem Lungenkrebs korreliert sind. Erklären kann ich einen auftretenden Lungenkrebs jedenfalls nur anhand des Rauchens und nicht anhand der gelben Finger.

Das Wissen über eine Korrelation ist nicht die für uns entscheidende Information, auch wenn sie manchmal einfach so behandelt wird. Dabei schließen wir meist vorschnell von einer Korreliertheit auf eine Kausalbeziehung. In den Medien finden wir immer wieder Berichte vom Typ: »Mediziner haben festgestellt, dass Gummibärchenesser bessere Mathematikleistungen aufweisen, als ihre Kommilitonen, die keine Gummibärchen essen.« Das suggeriert uns, wir könnten unsere Mathematikleistung durch Gummibärchen-Verzehr steigern, aber der berichtete Zusammenhang ist bloß eine Korrelation, die das keineswegs hergibt.

Noch naheliegender erscheint uns der Schluss bei ernsthafteren Meldungen wie: »Übergewichtige erkranken dreimal häufiger an Lungenkrebs als Normalgewichtige.« Wiederum wird schon durch die Wortwahl eine kausale Lesart suggeriert, sonst wäre die Meldung wohl auch relativ uninteressant. Sie würde zwar eine seltsame Korrelation aufzeigen, aber ohne eine Idee dafür anzubieten, wie diese Korrelation zustande kommt, können wir damit nicht viel anfangen.

Selbst Vorhersagen aufgrund der gelben Finger sind nur dann zulässig, wenn die gelben Finger ihrerseits durch das Rauchen hervorgerufen

wurden und nicht etwa durch das direkte Einfärben der Finger. Damit setzen wir selbst für einfache Prognosen schon bestimmte kausale Annahmen voraus, ohne die keine Vorhersage begründet erscheint. Das vergessen wir nur allzu leicht, weil wir solche kausalen Zusammenhänge schnell implizit voraussetzen. Sollten diese Annahmen aber bekanntermaßen nicht erfüllt sein, so würden die entsprechenden Vorhersagen dadurch unterminiert. Insbesondere Vorhersagen, die die Effekte unseres Eingreifens betreffen, und die daher besonders interessant für uns sind, zeigen diese Abhängigkeiten von kausalen Annahmen.

Wenn ich etwas zur Vorbeugung gegen den Lungenkrebs unternehmen möchte, ist es völlig sinnlos, die Verfärbung der Finger auf direktem Wege zu beseitigen. Nur die Variable Rauchen ist dafür entscheidend. Diese müssen wir verändern, um damit den Faktor Lungenkrebs zu verändern. Ansonsten können wir nur vorhersagen, dass sich vermutlich deshalb nichts ändern wird, weil die gelben Finger eben keine Ursache des Lungenkrebses sind.

Das zeigt noch einmal den wesentlichen Unterschied zwischen *bloßen Korrelationen* und *genuinen Kausalbeziehungen*. Die Letzteren gilt es letztlich zu ermitteln, selbst wenn für manche induktive Schlüsse auf den ersten Blick einfache Korrelationen genügen mögen. So können wir normalerweise durchaus von den gelben Fingern auf eine Lungenkrebsgefährdung schließen, aber nur, wenn wir annehmen dürfen, dass diese Gelbfärbung eben auf die ganz bestimmte Weise durch das Rauchen (wie es normalerweise der Fall sein dürfte) verursacht wurde. Normalerweise können wir uns bereits auf ein entsprechendes kausales Hintergrundwissen stützen, das unsere induktiven Schlüsse leitet. Dann können diese Schlüsse von der Gelbfärbung der Finger auf das Lungenkrebsrisiko dadurch gerechtfertigt sein, weil wir schon wissen, dass die gelben Finger vermutlich durchs Rauchen entstanden sind und dieses die Krebsgefahr deutlich erhöht.

Um diese Zusammenhänge weiter aufzuklären, könnten wir zunächst explizieren, was wir unter einer Kausalbeziehung verstehen wollen, bevor wir auf das kausale Schließen eingehen. Doch hier sollen die induktiven Schlüsse im Vordergrund stehen, und ich setze nur ein intuitives Verständnis von Kausalität voraus. Wir haben bereits erkannt, dass das induktive Schließen im Allgemeinen sich zumindest indirekt immer

schon auf kausale Annahmen stützen muss. Es ist schon aus diesem Grund nur natürlich, sich den *Kausalschlüssen* nun direkt zuzuwenden, um zu ermitteln, wie wir zu einfachen Kausalbehauptungen gelangen.

Mit »Kausalschlüssen« sind hier vor allem direkte Schlüsse von Daten auf einfache kausale Hypothesen gemeint. Diese Hypothesen müssen nicht gleich hochabstrakte Theorien sein, sondern sind oft relativ nahe an bestimmten empirischen Zusammenhängen liegende Hypothesen, die oft keine theoretischen Terme enthalten werden. Dazu gibt es einen großen Bereich an eigenständiger Literatur, den wir hier nicht umfassend zur Kenntnis nehmen können. Daher möchte ich vor allem auf zwei grundlegende und typische Verfahren und insbesondere ihre Grundideen und Rechtfertigungen eingehen.

Dabei geht es uns zunächst um generische Kausalbeziehungen. Wir möchten wissen, welche Faktoren i.A. welche anderen Faktoren beeinflussen. Das ist der typische Weg, um zu ermitteln, ob auch im konkreten Fall eine Kausalbeziehung vorliegt. Unser wichtigster Anhaltspunkt sind dafür die *Regularitäten* oder *Korrelationen*, die wir eben noch gescholten haben. Natürlich erwarten wir, dass für eine Ursache A von W sich bestimmte Regularitäten zumindest unter bestimmten Rahmenbedingungen zeigen. Wenn Schokolade Kopfschmerzen verursacht, dann sollte sich das vor allem darin zeigen, dass wir nach Schokoladengenuss häufig (oder sogar immer) Kopfschmerzen aufweisen – zumindest wenn weitere Rahmenbedingungen erfüllt sind. Doch das Schließen von Regularitäten auf Kausalbeziehungen ist komplexer, als es die oben genannten Fälle darstellen. Wir werden dazu entsprechende Verfahren einmal für eine deterministische Welt und einmal für eine möglicherweise indeterministische Welt kennenlernen.

Das erste Verfahren dient der Ermittlung von Kausalbeziehungen in einem deterministischen Rahmen und wurde u.a. von John Stuart Mill, John Leslie Mackie, Michael Baumgartner und Gerd Grasshoff entwickelt bzw. weiterentwickelt. Es schließt aus bestimmten Mustern von Koinzidenzen und auftretenden Regularitäten anhand der verbesserten INUS-Bedingung auf Ursache-Wirkungs-Zusammenhänge. Solche Verfahren für *deterministische Kausalität* werden oft unterschätzt, weil man etwa annimmt, der Determinismus würde implizieren, dass, wenn eine bestimmte Ursache auftritt, immer auch die entsprechende Wirkung

eintreten müsste. Dann wäre es kaum anwendbar, weil wir solche strikten Regularitäten nur in wenigen Bereichen finden können. Doch das ist ein grundlegender Irrtum. Wir werden etwa sehen, dass auch im deterministischen Rahmen auf das Auftreten einer Ursache A hin womöglich nur in 5% aller Fälle die Wirkung W auftritt. Das liegt dann insbesondere daran, dass die erforderlichen Kofaktoren zu A etwa nur in 5% aller Fälle vorliegen.

Der zweite Typ von Verfahren nutzt dagegen sogleich probabilistische Hilfsmittel und ist u.a. von Judea Pearl, Richard Scheines, Clark Glymour und vielen anderen entwickelt worden. Damit schließt man von bestimmten statistischen Zusammenhängen wie Korrelationen von Zufallsvariablen, die jeweils kausale Faktoren repräsentieren, auf kausale Abhängigkeiten, und versucht so einen Kausalgraphen der betreffenden Zusammenhänge zu erstellen. Wir werden dabei sehen, dass beide Ansätze auf weiteres kausales Hintergrundwissen angewiesen sind, womit sie das Diktum von Nancy Cartwright bestätigen: »No causes in, no causes out.« Judea Pearl beschreibt das – gerichtet gegen zu hohe Erwartungen in der klassischen Statistik in (2009, 100) – so: »This would have violated our golden rule: behind any causal conclusion there must be some causal assumption, untested in observational studies.« In meinem Rahmen werde ich zu zeigen versuchen, dass es sich jeweils um Spezialfälle des Schlusses auf die beste Erklärung handelt, die allerdings so weit ausgearbeitet sind, dass sie schon eine Sonderstellung mit speziellen Regeln einnehmen.

7.2 Die minimale Theorie: eine Regularitätstheorie der Kausalität

7.2.1 Die Grundlagen der minimalen Theorie

Zunächst zum ersten Ansatz. Es handelt sich um einen Regularitätenansatz, bei dem eine geeignete Regularität darüber entscheidet, ob A Ursache von W ist oder nicht. Insbesondere erwarten wir, dass A normalerweise zu W führt und die Abwesenheit von A zumindest in bestimmten Situationen dann auch W vermissen lässt. Bekannt wurde

dieser Ansatz vor allem durch die sogenannte INUS-Theorie von John Leslie Mackie (1974).

> **INUS-Bedingung**: Danach ist A eine *Ursache* von W genau dann, wenn A ein allein nicht hinreichender, aber notwendiger Teil einer Bedingung ist, die selbst hinreichend aber nicht notwendig für W ist. (INUS von »***i***nsufficient but ***n***on-redundant part of an ***u***nnecessary but ***s***ufficient condition«)

Das klingt viel komplizierter als es ist. Denken wir an ein brennendes Streichholz (A), das in eine Scheune mit Heu geworfen wird, woraufhin diese abbrennt (W). Das geworfene, brennende Streichholz, ist allein noch nicht hinreichend dafür, dass die Scheune abbrennt. Es muss z.B. Sauerstoff in der Scheune sein (B) und das Heu muss trocken sein (C). Allerdings sind auch (B) und (C) ohne (A) zusammen noch nicht ausreichend dafür, dass die Scheune abbrennt (W). Also ist (A) notwendiger Teil der komplexen Bedingung (ABC), wobei diese dann hinreichend dafür ist, dass die Scheune abbrennt. Außerdem ist die komplexe Bedingung (ABC) selbst nicht notwendig für den Brand der Scheune, denn es könnte noch andere Ursachen dafür geben. Ein glühendes Stück Metall, das in die Scheune geworfen würde (D), würde ebenfalls für einen Brand ausreichen, allerdings wieder zusammen mit (B) und (C). Damit würde (DBC) eine ebenfalls hinreichende, aber nicht notwendige Bedingung sein, von der in diesem Fall D ein solch notwendiger Teil wäre. Dadurch würde auch das glühende Stück Metall korrekt als Ursache identifiziert werden. Insgesamt sind also sowohl das geworfene Streichholz (A) wie das geworfene glühende Metall (D) jeweils notwendige Teile einer hinreichenden Bedingung (ABC oder DBC), die jeweils selbst nicht als notwendig für (W) ansehen.

Allerdings zeigt sich hier schon ein kleines Problem, denn (B) und (C) werden ebenso als Ursachen des Brandes eingestuft. Das klingt zumindest umgangssprachlich seltsam, weil wir typischerweise anhand von Ursachen auch Erklärungen abgeben wollen. Natürlich können wir auf die Frage, warum die Scheune abgebrannt ist, antworten, dass das deshalb passiert sei, weil jemand ein brennendes Streichholz dahinein geworfen hat. Komisch klingt es aber, wenn wir antworten: »Weil

Sauerstoff in der Scheune war.« Das hat sicher etwas damit zu tun, dass diese Bedingung (B) typischerweise als selbstverständlich bzw. als *Normalbedingung* oder Standardbedingung vorausgesetzt werden kann. Für (C) ist das schon nicht so klar, und daher wäre eine Antwort, die nun (C) in den Mittelpunkt stellt, auch nicht ganz so seltsam.

Trotzdem werden wir uns an dieser Stelle nicht lange mit solchen vermutlich eher *pragmatischen Aspekten* von Kausalaussagen und Erklärungen aufhalten. Das gehört vielmehr in die Debatten darüber, was genau die Ursachen mit Erklärungskraft sind und welche Ursachen keine Erklärungskraft aufweisen, während hier schließlich die Frage im Vordergrund stehen soll, wann wir überhaupt auf bestimmte Kausalbehauptungen schließen dürfen. Dazu werde ich an dieser Stelle nur kurz auf die diffizileren Fragen nach dem Wesen der Kausalität eingehen, und halte mich meist einfach an möglichst klare Beispiele von Kausalzusammenhängen. Aber auch für die wird die Frage nur schwer zu beantworten sein, wann uns spezielle Daten dazu berechtigen anzunehmen, dass eine Kausalbeziehung vorliegt.

Zunächst möchte ich wenigstens kurz auf die in unseren Überlegungen zu Kausalbeziehungen meist schon mitgedachte Metaphysik eingehen. Wir gehen in unseren Vorstellungen von Kausalität davon aus, dass bestimmte Ereignisse andere *hervorbringen*. Das können sie, weil sie bestimmte *Eigenschaften* instantiieren bzw. exemplifizieren, die wir oft als kausale Faktoren bezeichnen werden. Diese Eigenschaften werden zu diesem Zweck als Eigenschaften mit *dispositionalem Charakter* gedacht bzw. stellen *Kräfte* oder *Vermögen* dar, andere Eigenschaften zu erzeugen, wenn die richtigen Rahmenbedingungen vorliegen (vgl. Esfeld 2007, 2008). Diese inzwischen vieldiskutierte Auffassung vom Wesen der Kausalbeziehung (als einer Theorie der dispositionalen Eigenschaften) entspricht unserer typischen Alltagsmetaphysik, passt aber auch gut zu unserer physikalischen Weltauffassung, was Esfeld und andere zeigen konnten. Sie bildet auch hier den metaphysischen Hintergrund für unsere Überlegungen, spielt dafür im Weiteren aber keine besondere Rolle.

Demgegenüber geht ein Empirist im Sinne von Lewis (1994) von der Annahme der sogenannten *Humeschen Supervenienz* aus. Danach gibt es nur *eine* Struktur in unserer Welt und das ist die Raum-Zeit-

Struktur. An den einzelnen Raum-Zeit-Punkten finden sich dann Instanzen kategorialer (bzw. nicht-dispositionaler) Eigenschaften, die aber untereinander keine weiteren Verbindungen aufweisen. Sie sind kausal inaktiv. Es gibt eben nur zufällige Regularitäten in der Welt, die uns helfen, die Erscheinungen in der Welt zu systematisieren. Die grundlegendsten davon nennen wir *Naturgesetze*, aber diese geben eigentlich keinen Anlass zu der Vermutung, dass wir mit ihrer Hilfe gezielt in die Welt eingreifen und etwas verändern könnten, denn es handelt sich bei Naturgesetzen wie gesagt nicht um die Beschreibung bestimmter Vermögen der Natur, sondern nur um eine Beschreibung zufälliger Zusammentreffen bestimmter Eigenschaften. Es fällt uns im empiristischen Weltbild schon schwer zu verstehen, wie wir diese kategorialen Eigenschaften überhaupt *wahrnehmen* können, denn das verlangt eigentlich bereits, dass sie gewisse Wirkungen auf uns bzw. unseren Wahrnehmungsapparat ausüben. Unsere normale Annahme, dass wir gezielt in die Kausalstruktur eingreifen können, bleibt dann erst recht ein Rätsel (vgl. dazu etwa Esfeld 2007, 2008).

Es lässt sich genau genommen noch nicht einmal sinnvoll begründen, dass wir einem an Lungenentzündung Erkrankten Antibiotika geben, weil wir annehmen, dass sie ihm helfen werden. Die Eigenschaften der Tabletten sind nicht mit anderen Eigenschaften wie der Heilung verknüpft. Selbst wenn also in der Vergangenheit auf die Gabe von Antibiotika regelmäßig eine Heilung erfolgte, dann war das nur ein zufälliges Zusammentreffen und gibt uns keinen Grund anzunehmen, dass das auch in der Zukunft so sein wird. Ob es sich dabei um ein Naturgesetz im Sinne von Lewis handelt, wissen wir nicht und können wir nicht entscheiden, denn darüber entscheidet erst die gesamte Weltgeschichte. Sollte diese zufällige Regularität auch in der Zukunft weiter vorliegen, dann wäre es vermutlich ein Gesetz im Sinne von Lewis (jedenfalls dann, wenn es zu der besten Systematisierung aller Ereignisse der Welt gehören würde), und dann könnte man es vielleicht rückblickend als hilfreich ansehen, Antibiotika eingenommen zu haben, aber eigentlich gibt es keinen inneren Zusammenhang zwischen der Antibiotikaeinnahme und der Heilung. Kausale Beziehungen sind demnach ebenfalls nur zufällige Zusammenhänge, denen aber in Form gesetzesartiger Beziehungen (im

Sinne von Lewis) ein besonderer Stellenwert in der möglichst einfachen Systematisierung der Ereignisse unserer Welt zukommt.

Die empiristische Konzeption von Kausalität und Gesetzesartigkeit lässt eigentlich auch kein induktives Schließen zu. Wir können genau genommen nichts aus der Vergangenheit lernen, weil jede zukünftige Entwicklung der Welt zu jeder vergangenen in gleicher Weise passt. Im Weltbild des Empiristen ist kein Platz für eine substantiellere Vorstellung von Kausalität nach der gleiche Ursachen immer wieder gleiche Wirkungen hervorbringen. Für mich steht eine solche dagegen immer im Hintergrund, weil es sonst auch sinnlos erschiene, sich so für Kausalbeziehungen zu interessieren. Die empiristische Kausalvorstellung gestattet weder begründete Vorhersagen noch ein gezieltes praktisches Eingreifen in die Welt anhand von kausalem Wissen.

Leider haben ähnliche empiristische Ideen und speziell ihre Ablehnung der Kausalbeziehung (als zu metaphysisch) über einen langen Zeitraum besonders die klassische Statistik beeinflusst, die sich statt mit Kausalbeziehungen dann nur noch mit Korrelationen beschäftigt hat. Die empiristische Metaphysik ist jedenfalls dringend ergänzungsbedürftig und ich gehe hier davon aus, dass wir es mit einer reichhaltigeren Ontologie zu tun haben, in der vor allem die Beziehungen zwischen Eigenschaften eine kausale Struktur für unsere Welt liefern. Davon müssen letztlich alle Ansätze in diesem Bereich ausgehen und tun es mehr oder weniger explizit auch.

Der Regularitätenansatz ist aber sicher ein Ansatz, der zumindest für das Aufspüren von Kausalbeziehungen einen wichtigen Weg weist, der noch mit möglichst wenigen metaphysischen Annahmen belastet ist. Auf den ersten Blick muss der Ansatz sich nur auf die klassische Logik und einige idealisierende Annahmen stützen. Zu diesem Zweck wurde von John Leslie Mackie die INUS-Konzeption entwickelt. Darüber hinaus können wir natürlich auch diesen Ansatz mit einer gehaltvolleren Vorstellung der zugrundeliegenden Kausalbeziehung kombinieren. Für den Fall einer deterministischen Welt liefert dieser Ansatz bereits wichtige Einsichten darüber, was Ursachen in besonderer Weise auszeichnet, und dient so als gute Grundlage für weitere Kausaltheorien auch im indeterministischen Fall.

Die INUS-Bedingung selbst scheitert aber leider an bestimmten Gegenbeispielen, die ich weiter unten erläutern werde. Daher musste sie weiterentwickelt werden zur sogenannten *minimalen Theorie* (MT), deren Faktoren dann die Ursachen von W darstellen. Zu diesem Zwecke haben die Autoren Baumgartner und Grasshoff (2004) ihr sogenanntes *Doppelkonditional* »⟺« eingeführt, das ich zunächst benutzen und anschließend erläutern möchte. Es sei z.B. die folgende Konstellation eine minimale Theorie für W:

(**MT1 für W**) $AX_1 \vee BX_2 \vee Y \iff W$,

dann ist mit dem Doppelkonditional zunächst gemeint, dass A, B und W einzelne kausale Faktoren seien und AX_1 eine Konjunktion der Faktoren A und X_1 (ein Faktorbündel), wobei X_1 und X_2 spezielle Variablen sind, die jeweils für eine Konjunktion weiterer Faktoren – also ein Faktorbündel – stehen (etwa X = CDE) und Y eine Variable ist, die für weitere Disjunktionen von weiteren Faktorbündeln steht, also: $Y = X_3 \vee X_4 \vee \ldots \vee X_n$. Außerdem ist das Doppelkonditional so zu verstehen, dass alle Redundanzen schon gestrichen wurden. Das heißt zum einen, dass die einzelnen Faktorbündel *minimal hinreichend* sind (also alle ihre Faktoren erforderlich dafür sind, dass das Bündel noch hinreichend ist), und zum anderen, dass die gesamte Disjunktion der Bündel *minimal notwendig* ist für die rechte Seite. Das werde ich gleich noch weiter erläutern.

Mit »Faktoren« sind hier *Typen von Ereignissen* gemeint, die man sich am anschaulichsten als *Eigenschaften* vorstellen kann, die Ereignissen zukommen und sie in Klassen einteilen. So werden die Ereignisse solcher Typen von Ereignissen durch »etwas in Brand setzen« oder »etwas einem Feuer aussetzen« zu beschreiben sein, und das würde dann einen derartigen Faktor darstellen. Oder auch die Einnahme einer bestimmten Medizin könnte mehrere Ereignisse zu einem Typ von Ereignis bzw. einem Faktor zusammenfassen.

Die auftretenden Faktorbündel sind so zu deuten, dass sie als Ganzes jeweils eine komplette Ursache von W abgeben. Tritt also z.B. eine Instanz von AX_1 auf, so wird immer auch eine Instanz von W auftreten, sagt unsere minimale Theorie. AX_1 ist logisch gesehen eine *hinreichende*

Bedingung für W, bzw., wenn eine Instanz der Faktoren AX_1 auftritt, so ist das *hinreichend* dafür, dass auch eine Instanz von W auftritt. Jedes solche Bündel auf der linken Seite von MT1 ist hinreichend für W. Dann nennen wir A eine Ursache von W, da sie als Teil eines solchen Bündels auftritt. Dabei werden im Folgenden »hinreichend« und »notwendig« (ganz empiristisch) im rein extensionalen Sinne verstanden und nicht etwa modal. »A ist notwendig für B« heißt demnach, dass de facto zu jeder Instanz von B eine entsprechende Instanz von A vorliegt. Ob sich diese rein extensionale Redeweise tatsächlich durchhalten lässt, werden wir allerdings weiter zu prüfen haben.

Immerhin besagt schon die INUS-Bedingung, dass wir nur die *notwendigen* Teile solcher Faktorbündel als Ursachen ansehen dürfen. Wir könnten AX_1 natürlich um weitere (an sich redundante) Faktoren C ergänzen zu ACX_1 und das wäre immer noch hinreichend für W, ohne dass C einen Beitrag für W liefern müsste. Das soll vermieden werden. Also verlangt man, dass alle Faktoren des Bündels tatsächlich dafür notwendig sind, damit das ganze Bündel hinreichend ist für W. Das ist der erste *Minimalisierungsschritt*, in dem die Bündel jeweils minimalisiert werden (vgl. auch Baumgartner 2008b). Dabei kann ein Bündel z.B. auch in mehrere minimale Bündel zerfallen. Alle Faktoren eines solchen minimalen, aber für W hinreichenden Bündels stellen dann Ursachen von W dar. Die Idee ist ähnlich wie in einer Relevanzlogik. Es sollen nur solche Teile des Bündels gezählt werden, die in dem Sinne *tatsächlich relevant* sind, dass das Bündel seine Arbeit ohne sie nicht mehr erfüllen würde, nämlich zu einer Instanz von W zu führen. Ein Faktor C ist demnach nicht notwendig im Bündel ACX für W, wenn es Vorkommnisse von $A(\neg C)X$ gibt, bei denen trotzdem eine Instanz von W auftritt, obwohl C nicht vorliegt und auch keine der alternativen Ursachen von W gegeben ist.

In unserem Fall (MT1) muss ein A-Ereignis damit aber nicht in jedem Fall zu einem W-Ereignis führen. Denken wir wieder an unser Beispiel oben: A könnte ein brennendes Streichholz sein und W der Brand einer Scheune. Sollte aber aus irgendwelchen Gründen kein Sauerstoff in der Scheune sein, führt A nicht zu W. Wir müssen in unserer Darstellung dafür Vorsorge treffen, dass zu A weitere *Randbedingungen* X_1 hinzutreten können, damit W eintritt. Da wir oft jedoch nicht alle erforderlichen

Randbedingungen kennen, unter denen A zu W führt, wird zu diesem Zweck eine Variable X_1 eingeführt, die für diese Randbedingungen (ich nenne sie auch die *Kofaktoren von A*) stehen soll.

Es spricht natürlich nichts dagegen, mehrere Faktoren explizit anzugeben, wenn sie denn bekannt sind, wie etwa: $AX_1 = ADEFX_1*$. Dann stellen die Faktoren DEF (wenn sie wiederum notwendig sind, damit das ganze Bündel hinreichend bleibt) weitere Ursachen von W dar, die ihrerseits auf den Kofaktor A angewiesen sind, um wirken zu können. Außerdem können möglicherweise weitere Kofaktoren X_1* hinzukommen, die uns noch nicht bekannt sind. Die Anforderung der *Notwendigkeit innerhalb eines Bündels* ist also die erste Minimalisierung von MT1. Sie besagt, dass immer nur eine minimale Anzahl von Kofaktoren in jedes Faktorbündel aufgenommen werden darf.

Die weiteren Bündel in der MT1 (für W) wie BX_2 stehen für *alternative Ursachen* von W. BX_2 ist somit ein *weiteres minimales Bündel* von Faktoren, das ebenfalls für das Auftreten einer Instanz von W hinreichend ist. Demnach wäre auch B eine Ursache von W. Da wir jedoch oft nicht alle alternativen Ursachen explizit angeben können, wird die Variable Y stellvertretend für weitere alternative Ursachen eingeführt. Sie ist durchaus so gemeint, dass nun alle alternativen Ursachen darunter subsumiert werden, und besteht somit in einer Disjunktion weiterer Faktorbündel. D.h. das Vorliegen zumindest eines der Faktorbündel AX_1 oder BX_2 oder eines der Bündel aus Y ist *notwendig* dafür, dass W auftritt, bzw. das Auftreten von W ist hinreichend dafür, dass eine Instanz zumindest von einem der drei Bündel vorliegt. Wie gesagt wird hier eine deterministische Kausalstruktur vorausgesetzt, bei der kein Ereignis ohne Ursache auftritt, d.h., wenn W auftritt, dann muss auch eine der Ursachen vorgelegen haben, deshalb dürfen wir aus einer Instanz von W darauf schließen, dass eine Instanz von AX_1 oder von BX_2 oder von einem der Bündel in Y vorliegt. Die gesamte Disjunktion ist somit notwendig für W.

Da es leider noch einige problematische Fälle geben kann, auf die ich unten kurz eingehen werde, ist an dieser Stelle noch eine zweite *Minimalisierungsanforderung* zu stellen (vgl. zum Folgenden: Baumgartner 2008a,b, 2009b, Baumgartner & Graßhoff 2004). Es sollen nämlich so wenige Disjunkte wie möglich aufgenommen werden, bzw. wir

wollen nur die potentiellen alternativen Ursachen aufnehmen, die auch tatsächlich Ursachen darstellen. Eine notwendige Bedingung bleibt natürlich notwendig, wenn ich sie durch weitere (eigentlich redundante) Disjunkte abschwäche, (anders ausgedrückt: wenn $AX_1 \vee BX_2$ aus W folgt, d.h notwendig ist für W, so folgt auch $AX_1 \vee BX_2 \vee CX_3$ aus W), aber wir wollen hier nur eine *minimale Anzahl an Disjunkten* für eine minimale Theorie zulassen, weil sonst zu viele Bündel als alternative Ursachen betrachtet würden.

Angewandt auf unsere minimale Theorie MT1 heißt das: Wäre AX_1 allein bereits notwendig für W, hätten wir auf das Bündel BX_2 und auf Y in der minimalen Theorie verzichtet, weil wir sonst überflüssigerweise auch B zu einer Ursache von W erklärt hätten. Unsere neue minimale Theorie hätte dann die Gestalt: $AX_1 \Longleftrightarrow W$. Da das in unserem Beispiel nicht der Fall ist, und es sich bei (MT1) um eine minimale Theorie handeln soll, ist BX_2 demnach ein unverzichtbarer Bestandteil für die Notwendigkeit für W und B damit eine Ursache von W. Die Aufnahme nur der minimal notwendigen Faktorbündel ist die *zweite Minimalisierung* innerhalb der minimalen Theorie. Sie soll dafür sorgen, dass nur die alternativen Ursachen, die tatsächlich in einer bestimmten Situation allein für das Auftreten von W sorgen, in unserer minimalen Theorie aufgelistet werden. Dieser Minimalisierungsschritt kann natürlich nur dann wirklich ausgeführt werden, wenn wir die Variable Y durch konkrete alternative Ursachen ersetzt haben. Bevor ich kurz darauf eingehe, warum diese zusätzliche Minimalisierung erforderlich ist, die eine Verbesserung gegenüber der INUS-Bedingung darstellt, möchte ich die logische Struktur der minimalen Theorie noch etwas ausführlicher beschreiben.

Die hier genannten *Faktoren* A, B,..., W kennzeichnen *Typen* von Ereignissen. Es handelt sich so um *Eigenschaften* von Ereignissen, die von bestimmten Ereignissen instantiiert werden können, und es ist deshalb sinnvoll von Ax zu sprechen. Dabei wird von bestimmten Ereignissen x ausgesagt, dass sie die Eigenschaft A aufweisen, bzw. zur Ereignisklasse A gehören. Das Doppelkonditional von MT1 ($AX_1 \vee BX_2 \vee Y \Longleftrightarrow W$) stellt dann eine Abkürzung für den folgenden komplexeren Sachverhalt plus den Minimalisierungsschritten dar:

(**MT1**) (1) $\forall x\,((Ax \,\&\, X_1x \vee Bx \,\&\, X_2x \vee Yx) \rightarrow \exists y\,(Wy \,\&\, x{\neq}y \,\&\, Kxy)) \,\&$
(2) $\forall y\,((Wy) \rightarrow \exists x\,((Ax \,\&\, X_1x \vee Bx \,\&\, X_2x \vee Yx) \,\&\, x{\neq}y \,\&\, Kxy))$

Hier wird über Ereignisse x und y quantifiziert und im ersten Teil (1) verlangt, dass es zu jedem Ereignis x, das zumindest eines der Faktorbündel AX_1 oder BX_2 oder eines der Bündel aus Y erfüllt, ein Ereignis y existiert, das W instantiiert (kurz: ein W-Ereignis). Da es sich bei Ursache und Wirkung um *verschiedene* Ereignisse handeln soll, wird das ausdrücklich noch verlangt ($x \neq y$). Andererseits sollen die Ereignisse aber auch in einem bestimmten Sinne raumzeitlich zusammengehören. Daher wird zusätzlich Kxy gefordert, was bedeuten soll, das x und y *koinzidierende Ereignisse* sind. Darunter verstehen wir, dass sie räumlich und zeitlich benachbart stattfinden und nach Möglichkeit an ein und demselben System auftreten. Das ist sicherlich eine etwas vage Beziehung, aber in der Praxis dürfte meistens klar sein, was damit gemeint ist.

Der erste Teil von (MT1) besagt somit, dass jedes der Faktorbündel hinreichend für Wy ist, und für eine minimale Theorie wird zusätzlich noch gefordert, dass die Bündel alle minimal hinreichend sind. Im zweiten Teil wird die Notwendigkeit ausgedrückt, wobei auch hier zusätzlich gefordert werden muss, dass nur die *minimal* notwendigen Disjunkte aufgeführt werden. Wenn also ein W-Ereignis y vorliegt, dann muss auch ein Ereignis x vorliegen, das zumindest eines der möglichen Ursachenbündel instantiiert. Die zweite Minimalisierung ist erforderlich, weil es sonst irrelevante Disjunkte geben kann, die durch die minimale Theorie fälschlicherweise als Ursachen gezählt werden könnten (s.u.).

Allgemein ergibt sich jedenfalls im Rahmen dieses Ansatzes für eine minimale Theorie $AX \vee BX' \vee Y \Longleftrightarrow W$, mit Variablen X und X' für Bündel von Kofaktoren für A und B und einer Variablen Y für eine Disjunktion solcher Faktorbündel, die folgende Charakterisierung von Kausalität:

Minimale Theorie (MT): Der Faktor A ist *kausal relevant* für W gdw. eine minimale Theorie $AX \vee BX' \vee Y \Longleftrightarrow W$ gilt, bei der die Faktorbündel AX, BX' und die Bündel in Y jeweils *minimal*

hinreichend für W sind und die Disjunktion AX ∨ BX' ∨ Y zusammen *minimal notwendig* für W ist.

Eine kleine mögliche Komplikation (die Baumgartner 2007, 101 ff. erwähnt) möchte ich noch kurz erwähnen, ohne sie im Folgenden explizit umzusetzen. Ein Bündel AD muss in unserer minimalen Theorie zwar koinzidierend realisiert sein, aber das verlangt nicht zwingend, dass die beiden Kofaktoren Eigenschaften ein und desselben Ereignisses sein müssen. Es könnte durchaus sein, dass es sich um zwei Ereignisse x_1 und x_2 handelt und dann $Ax_1\&Dx_2$ vorliegt. Allerdings wird man dafür vermutlich wieder fordern müssen, dass die Ereignisse x_1 und x_2 koinzidieren. Jedenfalls könnten in (MT1) so weitere Variablen x_i für weitere Ereignisse erforderlich werden. Für alle Ereignisse müsste dann jeweils gelten, dass sie koinzidieren und alle x_i sollten natürlich von y verschieden sein.

Da das die Darstellung hier weiter verkomplizieren würde, verzichte ich aber darauf, noch weitere Variablen einzuführen (zumal wir für jede Variable X jeweils eine unbekannte aber endliche Zahl von zusätzlichen Variablen einführen müssten) und spreche einfach davon, dass wir ein komplexeres Ereignis x haben, das alle Eigenschaften eines Faktorbündels aufweist. Es dürfte auch einigermaßen klar sein, wie sich die allgemeinere Formulierung umsetzen ließe (vgl. dazu Baumgartner 2008a,b). Hilfreich könnte auch die Redeweise von *Systemen* sein, so dass alle Ereignisse als Teile eines komplexen Systems aufgefasst werden können. In der Praxis gehen an dieser Stelle typischerweise schon allgemeine Rahmenannahmen über Kausalzusammenhänge ein, auf die kein Ansatz kausalen Schließens ganz verzichten kann. Welche Ereignisse dann überhaupt als Ursachen welcher anderen in Frage kommen und wann wir von koinzidierenden Ereignissen sprechen wollen, fällt in den Bereich dieser Rahmenannahmen.

Die minimale Theorie ist also insgesamt eine *spezielle Regularitätstheorie*, die behauptet, dass wir Kausalbeziehungen daran erkennen können (und dass sie womöglich auch nur daraus bestehen?), dass wir in unserer Welt eine spezielle Regularität vorfinden: »Eigenschaft A ist generische Ursache von Eigenschaft W« heißt demnach, dass W

immer auftritt, wenn die Eigenschaft A zusammen ihren Kofaktoren instantiiert wurde und dass umgekehrt ein Faktorbündel mit A oder eines der anderen minimalen Bündel instantiiert wurde, wenn W auftritt. Die erforderliche Regularität ist also nicht ganz einfach zu beschreiben und zu erkennen, aber es wird nicht mehr verlangt für eine Kausalbeziehung, als das Vorliegen einer komplexen Regularität.

7.2.2 Empirische Vollständigkeit als Ersatzmodalität

Bevor ich im nächsten Kapitel die minimale Theorie an einem problematischeren Beispiel erläutere, möchte ich noch auf eine wichtige zusätzliche Forderung der Autoren der minimalen Theorie eingehen. Diese Forderung können wir als *empirisches Vollständigkeitsprinzip* bezeichnen. Das verlangt intuitiv, dass alle Konstellationen, die von der kausalen Struktur her *möglich* sind, auch tatsächlich realisiert sind. In Baumgartner und Grasshoff (2004) wird noch das etwas schwächer erscheinende Relevanzprinzip genannt:

> **Empirisches Relevanzprinzip**: Ein kausal relevanter Faktor ist mindestens in einer Situation unverzichtbar für das Entstehen einer Wirkung.

In Baumgartner (2009b) wird das aber noch deutlicher formuliert zu einem richtigen *Prinzip der empirischen Vollständigkeit* und das scheint auch erforderlich, wollen wir zwischen akzidentellen Regularitäten und tatsächlich kausalen Zusammenhängen unterscheiden:

> **Principle of Empirical Exhaustiveness (PEX):** The collection of empirical data to be processed by CNA faces no practical limitations whatsoever. All coincidences of the analyzed factors that are compatible with the causal structure regulating the behavior of these factors are in fact observed.

Also kommen alle kausal möglichen Konstellationen der Faktoren, die zu einer Wirkung führen, in unserer Welt gemäß diesem Prinzip tatsächlich vor und werden auch von uns beobachtet. Das ersetzt praktisch die *modale Komponente*, die viele anderen Kausaltheorien aufweisen. Nehmen wir etwa an, A sei eine generische Ursache von W

und führt meist zusammen mit einem Faktor K (als möglichem Kofaktor) zu W. Wenn jedoch K eigentlich verzichtbar ist, d.h. A schon mit anderen Kofaktoren allein zu W führt, dann wird in (PEX) gefordert, dass es zumindest eine reale Situation gibt, in der die Wirkung W allein aufgrund von A zusammen mit den anderen Kofaktoren von A auftritt, ohne dass K instantiiert wurde. Nur dadurch können wir auch erkennen, dass K kein notwendiger Kofaktor von A für W ist.

Ohne (PEX) könnte per Zufall oder durch andere kausale Verknüpfungen (K könnte ebenfalls von A verursacht werden) immer K vorliegen, wenn A W verursacht. Dann würde unsere minimale Theorie K fälschlicherweise zur Ursache von W erklären. Als Verfahren zur Ermittlung von Ursachen hätte es dann versagt und als Verfahren zur Definition von Kausalität wäre es einfach falsch. Damit das nicht passiert, wird die zusätzliche, stark idealisierende Annahme (PEX) eingeführt.

Wir müssen natürlich zugleich definitiv wissen, dass keine der alternativen Ursachen von W vorliegen, sonst hilft uns das nicht weiter. Also sind zu jeder Wirkung W z.B. alle minimalen Faktorbündel F für W tatsächlich zumindest einmal allein (ohne alternative Ursachen) realisiert – das verlangt (PEX). In der Praxis können wir solche Zusammenhänge manchmal anhand kausaler Rahmenannahmen begründen, die uns sagen, was überhaupt für W als Ursache in Frage käme.

Wollen wir etwa die gelben Finger als Ursache für den Lungenkrebs ausschließen, können wir in modaler Redeweise sagen, der Lungenkrebs *wäre* auch ohne die gelben Finger aufgetreten bzw. wäre auch aufgetreten, *hätten* wir die gelben Finger rechtzeitig gesäubert. Das Vollständigkeitsprinzip verlangt für diesen Fall, dass tatsächlich zumindest einmal der Lungenkrebs auftritt und nur das Rauchen zuvor vorlag und alle möglichen alternativen Ursachen abwesend waren, aber auch die gelben Finger. Wir können dann auf unsere Modalbehauptung verzichten, weil (PEX) sicherstellt, dass alle kausal möglichen Situationen tatsächlich realisiert sind. Wenn wir diese Situation dann auch noch kennen, können wir herausfinden, dass die gelben Finger keine Ursache des Lungenkrebses darstellen.

Damit wird zugleich deutlich, welche Faktoren tatsächlich für das Auftreten der Wirkung im modalen Sinn des Wortes *notwendig* sind. Wenn wir in modaler Redeweise sagen möchten, K ist nicht notwendig

für W, obwohl es zufälligerweise immer kurz vor W auftritt, dann würden wir das wie folgt erklären: Es wäre durchaus *möglich*, dass W auch ohne K realisiert würde, etwa in einem etwas *anderen Weltverlauf* oder in einer benachbarten *möglichen Welt*. Im Rahmen der minimalen Theorie wird schlicht anhand von (PEX) verlangt, dass *alle solchen kausal möglichen Situationen* bereits in der aktualen Welt realisiert sind (und uns nach Möglichkeit auch bekannt sind). Damit entgehen wir zwar dem Problem, auf genuin modale Ausdrücke in unserer Theorie zurückgreifen zu müssen, haben aber den Preis zu bezahlen, dass wir die genannte sehr starke *Idealisierung* vornehmen müssen. Man könnte behaupten, der Ansatz würde dadurch zu einer *implizit modalen Konzeption*, denn die starke Idealisierung ersetzt gerade die modale Komponente, auf die wir in Kausalkonzeptionen sonst oft angewiesen sind. (PEX) lässt sich am ehesten so begründen, dass wir auf (PEX) mehr oder weniger angewiesen sind, wenn wir bestimmte Ursachen ermitteln wollen. Wird (PEX) verletzt, kann es immer zu kausalen Fehlschlüssen kommen, die praktisch unvermeidbar sind, wenn wir nicht schon weiteres kausales Hintergrundwissen heranziehen, das uns den Fehler vermeiden hilft.

So treten z.B. im Normalfall immer die gelben Finger vor dem Lungenkrebs auf, nur das Vollständigkeitsprinzip verlangt, dass es zumindest eine Situation gibt, in der der Lungenkrebs ohne die gelben Finger vorkommt. Die liefert uns dann den entscheidenden Hinweis darauf, dass der Lungenkrebs auch ohne die gelben Finger aufgetreten *wäre*. Statt somit zu verlangen, dass die Regularität, die (MT1) ausdrückt, in allen möglichen Situationen (Welten) Bestand hat (bzw. einen notwendigen Charakter hat), wird hier einfach verlangt, dass die *kausal möglichen Situationen* tatsächlich alle in unserer Welt realisiert sind und damit (MT1) in allen kausal möglichen Situationen gilt. Natürlich müssen wir dann zusätzlich noch den Fall desjenigen ohne gelbe Finger kennen, der dann einen Lungenkrebs entwickelte. Außerdem müssen wir noch wissen, dass keine der alternativen Ursachen vorlag, sondern dass es allein das vorgängige Rauchverhalten war, dass hier als mögliche Ursache in Frage kommt.

Das ist eine sehr starke Anforderung, um ohne weitere modale Annahmen auskommen zu können. Wenn die Vertreter der minimalen Theorie z.B. gegenüber ihren Konkurrenzansätzen darauf hinweisen, dass

etwa ein Zugang zu Kausalbeziehungen über Bayessche Netze auf solche idealisierenden Annahmen wie die Graphentreue der Kausalbeziehungen angewiesen ist, sollten sie zugleich zugeben, dass sie mit der Annahme des Determinismus, der Annahme, die relevanten Größen alle zu kennen, und der Annahme, dass alle kausal möglichen Konstellationen tatsächlich auftreten, ebenfalls sehr starke Idealisierungen voraussetzen.

Wenn wir ohne (PEX) auskommen wollten, müssten wir die Konditionale verstärken und zumindest mit *kontrafaktischen Konditionalen* arbeiten. Statt der minimalen Theorie $AX \vee Y \Longleftrightarrow W$ könnten wir dann etwa verlangen, dass zwei kontrafaktische Konditionale gelten sollen:

(1) Für alle Ereignisse x gilt: Wenn x die Eigenschaft $AX \vee Y$ aufwiese, dann gäbe es ein von x verschiedenes Ereignis y, das mit x koinzidieren würde und die Eigenschaft W besäße.

(2) Gäbe es kein Ereignis x, das die Eigenschaft $AX \vee Y$ aufwiese, dann gäbe es auch kein von x verschiedenes Ereignis y, das mit x koinzidieren würde und die Eigenschaft W besäße.

Man müsste sich überlegen, wie wir das mit Hilfe von möglichen Welten beschreiben könnten. Eine Formulierung könnte etwa sein:

(W1) In allen nächstgelegenen (und gleichweit von unserer Welt entfernten) Welten, in denen ein Ereignis x auftritt, das die Eigenschaften $AX \vee Y$ aufweist, gibt es ein Ereignis y, das mit x koinzidiert und die Eigenschaft W besitzt.

(W2) In allen nächstgelegenen (und gleichweit von unserer Welt entfernten) Welten, in denen kein Ereignis x auftritt, das die Eigenschaften $AX \vee Y$ aufweist, gibt es kein Ereignis y, das mit x koinzidiert und die Eigenschaft W besitzt.

Dazu kämen die Minimalisierungsforderungen. So ließen sich vielleicht kontrafaktische Anforderungen für Kausalbeziehungen auf der Ebene der Typen von Ereignissen formulieren. Kehren wir aber nun zur klassischen Darstellung im Rahmen unserer Idealisierung (PEX) zurück und überprüfen, welche Probleme dafür weiterhin auftreten können.

Schauen wir uns dazu noch einige Beispiele an. Nehmen wir an, jemand stellt die Hypothese H auf, dass das *Aufessen von Korken* aus einer Weinflasche zu einem Verkehrsunfall in der Nähe führt. Nehmen wir außerdem an, dass nur zwei Leute bisher einen Korken verspeist haben (etwa aufgrund einer verlorenen Wette) und dass sich in beiden Fällen tatsächlich ein Verkehrsunfall in der Nähe ereignete. Ähnliche Regularitäten finden wir leicht zwischen Eigenschaften mit sehr wenigen Instanzen und solchen mit recht vielen Instanzen. Was benötigt ein Regularitätstheoretiker dann an Daten, um zu erkennen, dass der Unfall nicht durch den Korken verursacht wurde? Es würde noch nicht einmal genügen, wenn viele andere Leute nun Korken äßen und dabei würden keine Unfälle auftreten (D).

In Baumgartner (2008b) wird diese »Lösung« etwa für die Hypothese suggeriert, dass das Rauchen einer Havanna durch einen Seemann zum Untergang der Titanic geführt hat. Doch die Vertreter unserer Hypothese H könnten behaupten, dass das Korkenessen nur in Verbindung mit bestimmten *Kofaktoren* zu Verkehrsunfällen führt. Die waren in diesen anderen Fällen eben nicht gegeben. Ohne weitere kausale Rahmenannahmen käme nun eine unüberschaubar große Anzahl an möglichen Faktoren als Kofaktoren in Frage, und wir hätten nur wenige Möglichkeiten, sie empirisch zu testen, weil es nur so wenige Korkenesser gibt. Es könnte sich darum handeln, dass der Korkenesser braune Schuhe trägt, ein Ei zum Frühstück hatte etc. Das Datum D hilft uns allein also noch nicht weiter. (Haben wir es nun auch noch bei der Wirkung mit einem Ereignistyp mit nur einer Instanz zu tun, finden wir eine Regularität, bei der noch nicht einmal alternative mögliche Ursachen für die Wirkung auftreten müssen.)

In jedem Fall müssen wir zusätzlich darüber Bescheid wissen, welche Faktoren als Kofaktoren überhaupt in Frage kommen (also ebenfalls als kausal relevante Faktoren für Unfälle sein können) und müssen dann noch wissen, dass die in bestimmten Fällen des Korkenessens alle instantiiert waren und trotzdem das Korkenessen keinen Unfall nach sich zog. Damit benötigen wir also bereits eine ganze Menge an *kausalem Wissen*, um solche Schlüsse ziehen zu können. Haben wir ein derartiges Wissen über mögliche Ursachen von Verkehrsunfällen, sagt es uns aber auch sogleich, dass das Korkenessen ansonsten unbeteiligter

Personen keinen Einfluss auf den Unfall hat. Die Regularitätsanalyse ist in diesen Fällen daher kaum von großem Nutzen, zumal es viele solcher zufälligen Regularitäten geben kann, die wir normalerweise nur links liegenlassen, weil wir natürlich schon gute Gründe für unsere Annahme haben, dass es sich nicht um Kausalbeziehungen handelt. Auf dieses *kausale Hintergrundwissen* sind wir immer angewiesen, wenn wir nach weiteren genauer bestimmten Kausalbeziehungen suchen. Nancy Cartwright fasst das sehr schön in dem Slogan zusammen: »*No causes in, no causes out.*«

Einen anderen Typ von Regularität, der keine Kausalbehauptung nach sich ziehen sollte, finden wir in den Fällen von *gemeinsamen Ursachen* zweier Eigenschaften. Dazu ein Beispiel: Die Vertreter der sogenannten *konstitutionellen Hypothese* behaupteten, dass es wohl bestimmte Gene g wären, die zum einen für unsere Suchtanfälligkeit gegenüber Nikotin verantwortlich wären und somit das entsprechende Rauchen auslösen, aber zugleich unsere Anfälligkeit für Lungenkrebs befördern. Hätten sie Recht, wäre es natürlich unsinnig, sich das Rauchen zu versagen, denn das wäre nicht die Ursache für unseren Lungenkrebs. Nehmen wir das einmal an und nehmen weiterhin an, es wäre eine klare Regularität zu beobachten, nach der auf das Rauchen (zumindest am selben System, wenn auch nicht in raumzeitlicher Nähe) dann später Lungenkrebs auftritt. Regularitätsvertreter würden also vorschnell und fälschlicherweise dazu neigen, das Rauchen als Ursache des Lungenkrebses zu identifizieren (unsere falsche Hypothese T).

Welche Daten benötigen sie, um den Fehler zu vermeiden? Dass auch Nichtraucher Krebs bekommen, ist natürlich kein Problem für T, denn es gibt immer alternative Ursachen. Das nehmen Baumgartner und Grasshoff sogar ganz allgemein an, um mit Hilfe dieser Annahme die *Asymmetrie der Kausalbeziehung* zu erhalten. Wenn hingegen einige Raucher keinen Krebs bekommen, so könnte das wieder an den *fehlenden Kofaktoren* liegen (selbst in einer deterministischen Welt). Also benötigen wir wiederum eine Liste aller möglichen Faktoren und müssen spezielle Instanzen finden, in denen die Gene g vorliegen und zu Lungenkrebs führen, aber nicht zum Rauchen führen, weil dafür bestimmte Kofaktoren fehlen.

In diesem Fall müssten wir also wohl auch noch alle genetischen Ausstattungen vieler Menschen kennen und um ihre potentielle Gefährlichkeit für Lungenkrebs wissen, um T widerlegen zu können. Das klingt recht utopisch. In jedem Fall müssen wir zunächst einmal Einiges an Hintergrundwissen einbringen, ehe wir sinnvoll eine Regularitätsanalyse anwenden können. Wir müssen schon wissen, dass unsere Daten vollständig sind. Wir benötigen eine Liste aller relevanten Faktoren und müssen Koinzidenztabellen von Daten haben, von denen wir wissen, dass alle kausal möglichen Konstellationen darin auch vorkommen und keine solche Konstellation nur zufälligerweise bisher noch nicht aufgetreten ist bzw. ihr Auftreten von uns schlicht nicht bemerkt wurde.

Neben der Gefahr der vorschnellen Annahme von Kausalbeziehungen gibt es natürlich auch das Risiko, solche zu übersehen. Nehmen wir an, A sei kausal relevant für W, aber die dafür erforderlichen Kofaktoren BCD lägen jeweils nur in wenigen Fällen vor. Ohne eine Vorauswahl zu treffen, stehen wir dann vor einer unüberschaubar großen Zahl von Faktoren und Koinzidenzen und der Zusammenhang zwischen A und W kann nur auffallen, wenn wir tatsächlich für alle kausal relevanten Faktoren alle Kombinationen kennen. Insbesondere sind es nur für unser Beispiel wenige Fälle, in denen A und W koinzidieren. Sollten wir realistischerweise davon ausgehen, bestimmte Faktoren nicht zu kennen bzw., dass uns bestimmte mögliche *Kombinationen* nicht bekannt sind, werden wir unter den möglicherweise vielen Fällen, in denen A ohne W und in denen W ohne A auftritt, den wenigen Fällen kaum Beachtung schenken, in denen A und W gemeinsam auftreten. Für riesige Koinzidenztabellen fehlen uns sowieso die Daten und die Möglichkeiten, sie auszuwerten. Für n Faktoren sind bis zu 2^n Kombinationen möglich. Ohne eine Vorauswahl der vermutlich relevanten Faktoren und der vermutlichen Wirkungen zu treffen, ist das also kaum ein realistisches Verfahren.

Doch im Normalfall können wir dieses Hintergrundwissen einsetzen und haben dann nur bestimmte Kombinationen zu überprüfen. Nur rein theoretisch bietet die recht idealisierte Forderung (PEX) den Hintergrund für unser kausales Schließen. In der Praxis geht es nicht ohne die entsprechenden kausalen Rahmenannahmen. Immerhin können wir wenigstens in manchen Fällen anhand gezielter Experimente heraus-

zufinden versuchen, ob bestimmte Faktorenkombinationen auftreten können oder nicht.

Man könnte für das Prinzip (PEX) auch so argumentieren, dass es für jedes Verfahren kausalen Schließens unmöglich wird, die korrekten Kausalzusammenhänge zu ermitteln, wenn die Natur sie vor uns verbirgt oder unsere Daten systematisch unvollständig sind. Allerdings ist das zumindest noch eine Frage für die Praxis, wie unvollständig die Daten sein dürfen, so dass wir immer noch bestimmte kausale Schlüsse ziehen können. (PEX) stellt dagegen schon die maximale Anforderung für das kausale Schließen dar (vgl. auch Kelle 2003). Insgesamt beweist diese kleine Debatte wieder meine These, dass es keine einfachen Algorithmen für das induktive (kausale) Schließen gibt, die uns von Daten direkt zu bestimmten Hypothesen führen. Wir können diesen Schritt nur vollziehen, wenn wir schon kausales Hintergrundwissen voraussetzen.

Durch (PEX) lassen sich also die naheliegenden Einwände zurückweisen, die darauf beruhen, es könnten in (MT1) vielleicht nur akzidentelle Regularitäten (zufällige Zusammentreffen) beschrieben werden, die womöglich nur aus wenigen Einzelfällen bestehen und andere Koinzidenzen sind bisher zufälligerweise eben noch nicht aufgetreten. Dann könnte eine Regularität im Sinne von (MT1) schon für kurzzeitig auftretende Muster vorliegen, und uns zu vorschnellen Kausalschlüssen verleiten, die durch neue Daten schnell widerlegt werden könnten. Das wird durch das empirische Vollständigkeitsprinzip ausgeschlossen. Da wir aber in der Praxis tatsächlich nie sicher wissen, ob alle kausal möglichen Situationen schon realisiert und von uns registriert wurden, sind wir dort immer auf bestimmte modale bzw. kausale Vermutungen darüber angewiesen, was sonst noch möglich oder unmöglich ist.

Das ist auch nicht verwunderlich, ist doch Kausalität auch auf der Ebene der Typen dem Wesen nach ein *modaler Begriff*. Das findet sich in unserem Verständnis entsprechender Kausalaussagen wieder: Man will mit einer Kausalbehauptung wie »Schokolade verursacht Kopfschmerzen« zum Ausdruck bringen, bzw. etwas darüber aussagen, was in den Fällen passieren *würde*, in denen jemand Schokolade essen würde oder in denen er sie weglassen würde. Damit sage ich normalerweise nicht nur, dass de facto immer nach dem Schokoladengenuss ein Kopfschmerz aufgetreten ist, sondern dass der Schmerz auch aufgetreten *wäre*, hätte

jemand bei einer anderen Gelegenheit ebenfalls Schokolade gegessen. So behaupte ich also gleich etwas für *kontrafaktische Situationen*. Das gilt auch in der anderen Richtung. Hätte ich bei einer bestimmten Gelegenheit die Schokolade weggelassen, so wären die Kopfschmerzen nicht aufgetreten. Ganz im Sinne unserer kontrafaktischen Deutung der minimalen Theorie. Es muss vor allem bestimmte Situationen geben, in denen der Schokoladengenuss einen Unterschied bezüglich der Kopfschmerzen machen würde.

Das war schon die eine von Humes Intuitionen zur Kausalität. Dahinter steht bei uns heute die Idee, dass bestimmte natürliche Eigenschaften der Ereignisse diesen Unterschied mit sich bringen und somit diese Wirkungen auch in kontrafaktischen Situationen hervorbringen würden (vgl. Esfeld 2008 und 2008a). Im Ansatz der minimalen Theorien wird diese modale Intuition also nur in dem Vollständigkeitsprinzip »versteckt«, das praktisch dafür sorgt, dass unsere Regularitäten statt nur einige wenige zufällig aufgetretene Situationen schon alle *kausal möglichen* Welten abdecken und so alle möglichen Wirkungen einer Eigenschaft tatsächlich zu beobachten sind. Dadurch finden sich alle möglichen Situationen (also unsere Modalität) in der aktualen Welt wieder.

Das ist jedoch wieder ein Fall einer kausalen Hintergrundannahme, die wir voraussetzen müssen, um kausal schließen zu können. In der Praxis werden wir überlegen müssen, welche Erklärung jeweils die bessere ist. Einerseits können wir schließen, dass A wohl eine Ursache für W sein muss, weil A bisher immer vor W aufgetreten ist, andererseits mag auch unsere *beste Erklärung* für diese Regularität die sein, dass es sich nur um einen Zufall handelt, dass wir also bisher noch nicht genügend Situationen beobachtet haben, um solche zu finden, in denen A auch ohne W vorkommt. Wir müssen uns wiederum zwischen *konkurrierenden Erklärungen* entscheiden. Damit wird zugleich deutlich, dass solche Unvollständigkeiten *typische Unterminierer* für unser kausales Schließen anhand der minimalen Theorie kennzeichnen.

Schließlich wirkt der Regularitätsansatz somit nicht mehr ganz so reduktiv, wie er auf den ersten Blick erschien. Nehmen wir die Minimalisierungsbedingungen und die Vollständigkeitsbedingung zusammen, können wir das auch so ausdrücken: Ein Faktor A ist dann Ursache von W, wenn A ein *kausal* notwendiger Teil einer *kausal* hinreichen-

den Bedingung für W ist, die *kausal* nicht redundant ist in einer Disjunktion von *kausal* notwendigen Bedingungen für W. Durch die Vollständigkeitsbedingung kommt diese Art von kausaler Modalität ins Spiel und wird in meiner Formulierung nur explizit gemacht. Erst diese Bedingung sorgt dafür, dass durch die materiale Implikation auch die kausal hinreichenden und tatsächlich erforderlichen Faktoren gekennzeichnet werden können.

Wie gesagt, das ist nicht verwunderlich und auch kein anderer Ansatz kann die Kausalbeziehung auf kausal unverdächtige Begriffe reduzieren. Die minimale Theorie gibt uns schon wesentliche Einsichten in die *Bedingungsstruktur* von deterministischen Kausalbeziehungen und die sind wiederum das Vorbild für die probabilistischen Kausalvorstellungen. Sie leiten weiterhin unsere Ermittlung von Kausalbeziehung etwa durch Experimente an, aber sie ist nicht wirklich reduktiv in einem starken Sinne.

Außerdem ist die minimale Theorie auch nicht reduktiv, weil sie uns nicht erklärt, was die Kausalbeziehung ist bzw., was dazu führt, dass eben ganz bestimmte Kombinationen von Faktoren auftreten können und andere nicht. Es wird einfach eine bestehende Kausalstruktur vorausgesetzt und die Forderung, ein notwendiger Bestandteil einer hinreichenden Bedingung zu sein, gibt uns nun nur an, wie wir diese kausalen Zusammenhänge aufspüren können. Sie erklärt aber nicht weiter, welche Elemente in unserer Welt wie beschaffen sein müssen, um solche Strukturen zu generieren.

Dabei denke ich – wie gesagt – an die Überlegungen, wie wir sie z.B. bei Esfeld (2007, 2008, 2008a oder 2010) finden (s.o.). Dort wird erläutert, inwiefern bestimmte *grundlegende Dispositionen* für das Auftreten einer zugrundeliegenden Kausalstruktur in unserer Welt verantwortlich sein können; doch das ist nicht das Thema der Vertreter der minimalen Theorie und muss es natürlich auch nicht sein. Man sollte dann nur nicht davon sprechen, dass man Kausalität auf logische Beziehungen reduziert hätte, denn die entscheidende Voraussetzung der minimalen Theorie ist schon, dass die kausale Struktur unserer Welt bereits gegeben ist und sich dann in all ihren Möglichkeiten in unseren Daten manifestiert. In solch einer stark idealisierten Welt können wir sie durch die logischen Zusammenhänge in den Daten womöglich wieder erschließen, erklären

dabei aber natürlich nicht, was die Grundlage für diese Kausalstruktur ist. Es handelt sich also um eine bloß *epistemische Theorie kausalen Schließens* und nicht um eine ontologische oder naturphilosophische Theorie der Kausalität. Weitere grundlegende Probleme für diesen Ansatz werden sich vor allem für den Übergang zu singulären Kausalbeziehungen zeigen.

7.2.3 Ein Problem für die INUS-Theorie

Betrachten wir nun – wie versprochen – einen Problemfall für jede Regularitätstheorie und auch für probabilistische Ansätze, nämlich den Fall von *Common-Cause-Strukturen*, bei denen wir Regularitäten bzw. Korrelationen finden, die jedoch keine direkten kausalen Beziehungen darstellen. Die stellen ein besonderes Problem für die INUS-Theorie dar, weshalb die INUS-Theorie auch nicht mehr weiter verfolgt wurde. Nehmen wir an, dass Rauchen (R) deterministisch zu Lungenkrebs (L) und gelben Fingern (G) führt (Kausalstruktur: G$\Leftarrow$R$\Rightarrow$L) und lassen alternative Ursachen für den Lungenkrebs einmal außen vor. Weiterhin können die gelben Finger auch durch Anmalen (A) entstehen (also ist die ausführlichere Kausalstruktur: A$\Rightarrow$G$\Leftarrow$R$\Rightarrow$L), dann könnte man auf die Idee kommen, dass die folgende Theorie minimal für L sei:

(T1) (non-A)G $\vee$ R $\Longleftrightarrow$ L,

denn sowohl R, als auch (non-A)G sind *minimal hinreichend* für L. Das können wir uns in einer Tabelle ansehen:

S	A	G	¬AG	R	L	Schlussfolgerungen
S1	1	1	0	1	1	zeigt, dass R allein hinreichend für L ist
S2	1	1	0	0	0	zeigt, dass R notwendig ist für L (mit S4)
S3	0	1	1	1	1	zeigt, dass (¬A)G hinreichend ist für L
S4	0	0	0	0	0	(¬A) ist notwendiger Teil von (¬A)G für L

Tabelle 7.1: Ein Problemfall für die INUS-Bedingung

Liegt *kein* Anmalen vor und treten trotzdem gelbe Finger auf, so folgt nämlich aus unserer Kausalstruktur, dass in jedem Fall R gilt und dann folgt aus dem angenommenen Determinismus und der vorliegenden

Kausalstruktur, dass auch L instantiiert sein muss. Also dürfen wir aus (non-A)G auf das Vorliegen von L schließen, denn in allen kausal möglichen Situationen wird mit einer Instanz von (non-A)G auch eine Instanz von L vorliegen. Damit liefert (T1) zumindest eine INUS-Bedingung für L und im Rahmen der INUS-Konzeption müssten wir also non-A fälschlicherweise als kausal relevant für L betrachten.

Aber (non-A)G und R sind in einer Disjunktion nicht zusammen *minimal* notwendig für L, d.h., (T1) ist *keine minimale Theorie*. Das erkennt man so: Wenn (non-A)G instantiiert ist, so muss in jedem Fall auch R instantiiert sein. Oder anders ausgedrückt: Wenn L instantiiert ist, so muss auf jeden Fall R vorliegen, d.h., R ist in jedem Fall notwendig für L und gehört daher in eine minimal notwendige Bedingung für L hinein. In unserem Beispiel muss jemand geraucht haben, um Lungenkrebs bekommen zu können. Das Rauchen ist also in jedem Fall notwendig für den Krebs. Ob aber A oder (non-A) vorliegen, ist keineswegs ausgemacht, wenn L der Fall ist, und damit muss auch nicht zwangsläufig (non-A)G vorliegen. Lungenkrebspatienten müssen nicht in jedem kausal möglichen Fall auf das Gelb-Anmalen ihrer Finger verzichtet haben. Die korrekte minimale Theorie wäre also:

(T2) $R \Longleftrightarrow L$,

Es zeigt sich so eine gewisse Asymmetrie zwischen (non-A)G und R. Es muss in jeder Situation, in der L auftritt, auch R gegeben sein, aber nicht unbedingt (non-A)G. Das Vollständigkeitsprinzip sorgt dafür, dass sich tatsächlich solche Situationen finden lassen, in denen zwar eine Instanz von R auftritt, aber keine von (non-A)G. Es findet hier zusätzlich ein Anmalen der Finger statt. Die Finger werden dann in überdeterminierter Weise gelb. Damit können wir auf den Faktor (non-A)G verzichten. Er muss nicht vorliegen, wenn L auftritt. Es bleibt als minimale Theorie nur (T2) übrig, die die Ursachen von L in der Tat korrekt benennt. Das ist der entscheidende Vorteil des zweiten Minimalisierungsschrittes, der es gestattet, Scheinursachen von L wie (non-A)G auszuschalten.

In unserer Tabelle fehlt eine Zeile, in der (non-A)G vorliegt, aber R nicht und trotzdem L auftritt. Das Vollständigkeitsprinzip besagt dann, dass diese Zeile nicht auftreten kann. Das zeigt für uns an, dass (non-A)G

allein keine alternative Ursache von L ist und entsprechend auch nicht zu den minimal notwendigen Bedingungen für L gehört.

Das entsprechende etwas komplexere »Manchester Hooters«-Beispiel wird bei Mackie (1974) vorgestellt und bei Baumgartner & Grasshoff (2004) ausführlich diskutiert. Es hat dazu geführt, dass die INUS-Theorie nicht mehr weiter verfolgt wurde. Doch durch den zweiten Minimalisierungsschritt hin zur minimalen Theorie lässt sich dieses Defizit der INUS-Theorie beheben.

Außerdem verlangen die Autoren noch, dass jede minimale Theorie *alternative Ursachen* aufweisen muss. Dann müssten wir (T2) noch um weitere Ursachen für L erweitern. Das hat vor allem die Aufgabe, eine *Asymmetrie* zwischen den beiden Seiten des Doppelkonditionals aufzuzeigen. (T2) stellt eine ganz symmetrische Beziehung dar und es wäre unklar, was dann als Ursache und was als Wirkung auszuzeichnen wäre. Wir verzichten in unserem Beispiel auf die Einführung alternativer Ursachen zugunsten der besseren Übersichtlichkeit, aber es ist natürlich bekannt, dass es neben dem Rauchen auch andere Ursachen für Lungenkrebs gibt, wie z.B. eine Asbestkontamination. Dann wird eine Asymmetrie in der minimalen Theorie deutlich (auf der einen Seite steh eine Disjunktion auf der anderen nicht), die nach Ansicht von Grasshoff und Baumgartner der Asymmetrie der Kausalbeziehung entspricht.

Allerdings könnte es passieren, dass es keine eindeutig bestimmte minimale Theorie mehr gibt. Nehmen wir an, dass das Rauchen R *zwingend* das Anmalen der Finger verhindern würde. Das ist in unserem Fall etwas unrealistisch, aber solche Kausalstrukturen sind natürlich möglich. Unsere Raucher haben die »ökonomische« Disposition sich in jedem Fall das Gelb-Anmalen der Finger zu ersparen, da sie schließlich schon durch das Rauchen gelb gefärbt werden. R verursacht also zusätzlich (non-A). Dann verschwindet die obige Asymmetrie zwischen R und (non-A)G und wir erhalten eine zweite minimale Theorie:

(T3) (non-A)G $\Longleftrightarrow$ L,

denn sobald jetzt L instantiiert wurde, so muss auch (non-A)G vorhanden sein. Damit liefert die Konzeption der minimalen Theorie für diesen Fall keine eindeutigen Ergebnisse mehr oder schlimmer noch, die minimale

Theorie (T3) stellt eine Scheinursache als Ursache dar. Allgemein können solche Zyklen in der Kausalstruktur wie in unserem Beispiel zwischen A, G und R gefährlich werden für die Analyse von Kausalbeziehungen. Vielleicht muss man hier weitere Anforderungen an die zu analysierenden Kausalbeziehungen stellen, um zu eindeutigen Ergebnissen zu kommen.

In dem obigen problematischen Beispiel könnte man auch sagen, dass es ein intuitives Problem von (T3) ist, dass hier ein *negativer Faktor* (non-A) – also die Abwesenheit einer Instanz von A – als Ursache zugelassen wird. Das ist sicher nicht unproblematisch, aber manchmal reden wir so. Wir sagen etwa, meine Blumen im Vorgarten sind vertrocknet, weil meine Nachbarin, die sich für meine Urlaubszeit verpflichtet hatte, sie zu gießen, es dann leider nicht getan hat. Die Abwesenheit des Gießens durch meine Nachbarin war also die Ursache für das Vertrocknen. Andererseits hätte auch Angela Merkel meine Blumen im Vorgarten gießen können. Trotzdem würden wir kaum sagen wollen, dass das Nichtgießen durch Angela Merkel die Ursache für das Eingehen meiner Blumen sei, selbst wenn die ebenfalls in meiner Nachbarschaft gewohnt hätte. Wenn ich ihr deshalb eine Rechnung über meine eingegangenen Blumen präsentieren würde, wäre sie vermutlich sehr irritiert. Meine Nachbarin steht da schon mehr in der Pflicht, die Sache wiedergutzumachen. Hier scheinen plötzlich recht seltsame und eher pragmatische Aspekte in die Frage nach einer Kausalbeziehung durch negative Faktoren hineinzuspielen, wie die besondere moralische Verpflichtung der Nachbarin. Den Debatten um negative Faktoren kann ich hier nicht nachgehen, da sie zu einer weitergehenden umfangreichen Diskussion gehören, was man unter Kausalität zu verstehen hat (vgl. die Diskussion negativer Faktoren in Baumgartner 2007, Kap. 3.6.4), für die unsere Vorstellungen von Kausalität nicht immer klar sind.

Die Konstrukteure der minimalen Theorie versuchen aus der recht intuitiven Konzeption, die Ursachen und ihre Kofaktoren so übersichtlich zusammenstellt, nun eine Theorie kausalen Schließens zu gewinnen. Aber auch hier sind wir wieder auf weiteres kausales Hintergrundwissen angewiesen. Bevor wir uns den Schlüssen auf Kausalbeziehungen auf der Ebene der Typen zuwenden, möchte ich zunächst auf die wichtige Problematik der Schlüsse auf singuläre Kausalbeziehungen eingehen.

7.2.4 Singuläre Verursachung

Ein weiterer kausaler Schluss, den wir nun untersuchen, geht von der generischen Ebene der Kausalität zur singulären. Wir möchten bestimmte konkrete Ereignisfolgen als kausale Instanzen unserer bekannten Kausalgesetze interpretieren und hoffen dabei, das aus der minimalen Theorie Gelernte hilfreich einsetzen zu können. Nehmen wir an, wir kennen schon die kausalen Beziehungen auf der Ebene der Typen von Ereignissen und ihre minimalen Theorien und möchten nun ermitteln, welche singulären konkreten Ereignisse jeweils in der Ursache-Wirkungs-Beziehung stehen. Das ist für jede Kausaltheorie ein entscheidender Schritt, denn dabei geht es um wichtige Anwendungen der Theorie. Letztlich möchten wir wissen, ob es sich für Fritz lohnt, das Rauchen aufzugeben, ob Franz der Verursacher eines konkreten Unfalls war etc.

Die Frage nach der singulären Kausalität scheint auf den ersten Blick mit den bisherigen Erkenntnissen leicht zu beantworten zu sein, erweist sich aber dann doch als durchaus problematischer, als die Vertreter des Ansatzes sich das denken. Sie hoffen (etwa in Baumgartner 2013), sie könnten die kausale Relevanz von der generischen Ebene relativ elementar auf die Instanzen der jeweiligen Faktoren in etwa mit Hilfe des folgenden Schemas übertragen (das dort verwendete Schema ist noch etwas komplexer, um dem Problem Rechnung zu tragen, dass Kausalität nicht unbedingt transitiv sein muss, womit wir uns später noch beschäftigen werden):

Singuläre Verursachung (**SV**)
Ein Ereignis a ist *Ursache* eines Ereignisses b gdw.
(1) a ist Instanz von A und b ist Instanz von B,
(2) A ist Ursache von B im Sinne der minimalen Theorie und die Kofaktoren von A für B werden ebenfalls von a instantiiert,
(3) a ist ungleich b und
(4) a und b koinzidieren, d.h., sie treten raumzeitlich benachbart auf.

Dabei stehen Kleinbuchstaben immer für konkrete (singuläre) Ereignisse und Großbuchstaben für Ereignistypen. Dann soll (SV) zusammen mit

dem angenommen Determinismus dafür sorgen, dass a Ursache von b sein muss, aber so leicht gelingt der Schritt auf die Ebene der konkreten Ereignisse leider nicht.

Nehmen wir einmal an, dass wir in unmittelbarer Nachbarschaft zwei Zündschnüre entzünden. Das seien die Ereignisse a und a*, die im übrigen dieselben Faktoren A (Zündschnüre entzünden, die zu einem Tischfeuerwerk führen) instantiieren. Die erste Zündschnur (durch a entzündet) entzündet dann Tischfeuerwerk 1 und die zweite (durch a* entzündet) das Tischfeuerwerk 2, die eng benachbart liegen wie auch die Zündschnüre. Diese letzteren Ereignisse nennen wir b und b* und auch sie sollen dieselben Faktoren insbesondere B (Tischfeuerwerk brennt) instantiieren. Außerdem seien alle erforderlichen Kofaktoren wie die Anwesenheit von Sauerstoff etc. instantiiert. (Womöglich kreuzen sich die Zündschnüre sogar noch – ohne sich gegenseitig zu entzünden –, so dass alle Ereignisse eng benachbart stattfinden.) Also verursacht a nun b und a* verursacht b*. Da alle Ereignisse raumzeitlich benachbart sind, müsste nach (SV) aber ebenfalls gelten: a verursacht b* und a* verursacht b, was jedoch tatsächlich nicht der Fall ist. (SV) liefert also falsche singuläre kausale Zusammenhänge für diesen Fall.

Die korrekten Zusammenhänge würden hier vermutlich besser durch die Prozesstheorie der Kausalität oder einen kontrafaktischen Ansatz beschrieben. Die Prozesstheorie könnte den Prozess korrekt nachzeichnen, der jeweils von a nach b und von a* nach b* geführt hat, und könnte so die richtigen Ursachen für die beiden Ereignisse b und b* finden. Im kontrafaktischen Ansatz könnte man darauf hinweisen, dass, wenn a nicht stattgefunden hätte, auch b so nicht stattgefunden hätte. Das Tischfeuerwerk 2 wäre vielleicht später durch die Nähe zu dem anderen Tischfeuerwerk 1 abgebrannt, aber das wäre nicht das Ereignis b gewesen, da das früher stattgefunden hätte. In der nächsten Welt, in der a nicht vorkommt, wäre also auch b nicht aufgetreten. Aber trotzdem hätte weiterhin b* stattgefunden, da b* schließlich durch a* verursacht wurde, das weiterhin stattgefunden hätte. Auch eine Interventionstheorie der Kausalität, könnte das entsprechend auflösen.

Um mit Hilfe der minimalen Theorie zu singulären Kausalbeziehungen zu gelangen, sind jedenfalls weitere Annahmen nötig. Wir müssen etwa ausschließen, dass es entsprechende kausale Paare in der Nachbarschaft

gibt, mit denen ein solcher Partnertausch möglich ist. Das ist ein typisches Merkmal der *eliminativen Induktion*. (SV) ist nur anwendbar, wenn wir solche benachbarten Ereignispaare als typische Unterminierer von (SV) bzw. die entsprechenden Kausalhypothesen ausschließen können.

Oder wir müssen zu noch feineren Faktoren übergehen wie dem: Entzünden einer Zündschnur, die an Tischfeuerwerk 1 hängt. Doch das sieht dann langsam schon nach den Epizyklen der minimalen Theorie aus, auf die jedenfalls der kontrafaktische Ansatz nicht angewiesen ist.

Übrigens finden wir an dieser Stelle auch die sogenannten *»Preemption«-Probleme* wieder. Das sind Fälle, in denen eine tatsächliche Ursache U nicht unbedingt notwendig für das Auftreten der Wirkung W ist, weil eine *Ersatzursache* E zur Stelle ist, die einspringt, falls U ausfällt und dann ihrerseits W hervorruft. Nehmen wir dafür ein einfaches Beispiel her: Schütze 1 schießt auf den Diktator (U) und wird ihn dadurch töten (W). Kurz danach schießt aber ebenfalls Schütze 2 mit derselben Präzision (E) und würde den Diktator töten, wenn er nicht schon tot wäre. Dann haben wir das spezielle Problem für alle Ansätze zur Kausalität, die behaupten, dass Ursachen vor allem ein *Unterschiedsmacher* für das Auftreten der Wirkung sind, dass U eigentlich für den Tod des Diktators keinen Unterschied bedeutet, denn schließlich sorgt auch E zuverlässig für W.

Was ein solcher Unterschiedsmacher ist, wird dabei in den verschiedenen Ansätzen zur Kausalität unterschiedlich beantwortet. Die kontrafaktischen Ansätze behaupten, dass U einen Unterschied für W macht, wenn in der nächsten möglichen Welt, in der U nicht vorliegt, dann auch W nicht vorliegt. In der minimalen Theorie wird dagegen für einen Unterschied verlangt, dass U ein notwendiger Teil einer hinreichenden Bedingung ist, d.h., dass wir aus der Bedingung nicht mehr auf das Vorliegen der Wirkung schließen dürfen, wenn dieser Teil nicht vorliegt.

Das »Preemption«-Problem überträgt sich letztlich auf alle Ansätze der Differenzmethode zur Ermittlung von Ursachen. Die »Ersatzursachen«, die in diesen Beispielen auftreten, sind typische Unterminierer für das Schließen auf Ursachen anhand der Differenzmethode, weil sie dafür sorgen, dass die Wirkung W auftritt, ganz gleich, ob U vorliegt oder nicht. Das gilt zumindest für die singuläre Kausalität, weil die

Randbedingungen (speziell die Ersatzursache) dort jeweils gegeben sind und das Prinzip (PEX) uns daher hier nicht direkt weiterhelfen kann. Ist U aber kein Unterschiedsmacher für W mehr, gilt er im Rahmen der Differenzmethode nicht mehr als Ursache von W. Damit werden solche Ursachen nicht mehr korrekt eingestuft.

Im Bereich der Ereignistypen ist das für die minimale Theorie nicht so bedeutsam, weil sie nur verlangt, dass U zumindest in bestimmten Fällen als Unterschiedsmacher für W erkennbar ist. Sind die kausal möglich, so sagt das Vollständigkeitsprinzip, dass die auch tatsächlich realisiert sind und uns bekannt sind. Es verbietet sozusagen, dass die Unterminierer *immer* vorliegen. Da die minimale Theorie für die Ebene der Ereignistypen geschaffen wurde, genügt das, um dem Preemption-Problem zu entkommen. Nur bei der Übertragung auf den Einzelfall kommen sie uns wieder in die Quere – genau wie in den kontrafaktischen Ansätzen, die normalerweise direkt für den Einzelfall konzipiert sind.

Nehmen wir an, Uwe hätte ein Pfund Arsen zu sich genommen und würde daran sicher sterben. Aber bevor die Wirkung des Arsens eintritt, wird er vom Bus überfahren. Ursache seines Todes ist dann der Busunfall und nicht die Arseneinnahme. Doch natürlich wollen wir der Arseneinnahme nicht ihre generische kausale Relevanz für das Ableben von Menschen absprechen und da sie im Beispiel instantiiert wurde (mit allen relevanten Kofaktoren) ist sie nach (SV) dann eine Ursache des Ablebens von Uwe. Doch das stimmt in unserem Beispiel nicht, denn sie ist hier nur eine Ersatzursache, die nicht wirklich zum Zuge kommt. Nur der Bus stellt die tatsächliche Ursache für Uwes Tod dar. Hier wird also ein Ereignis von (SV) fälschlicherweise als Ursache betrachtet.

Die problematischen Beispiele können noch komplizierter werden, wie Strevens (2007) aufzeigt. Nehmen wir an, Sabine wirft eine kleine Kanonenkugel auf einen Krug (S), und würde ihn auch perfekt treffen, wenn nicht Tom zugleich eine Kugel auf den Krug geworfen hätte (T), die allerdings deutlich daneben gezielt war. Die Kugeln treffen sich allerdings in der Luft und Sabines Kugel wird nun vom Krug abgelenkt, während Toms Kugel danach direkt auf den Krug fliegt und ihn zerstört (K). Toms »Danebenwerfen« (T) ist also nur im Verein mit Sabines Wurf (W) eine hinreichende Bedingung für das Auftreten von K. Das heißt, die Bedingung TS ist hinreichend, aber nicht die Bedingung T. Das

Problem ist nur, dass auch S allein hinreichend ist. T verliert damit seine Notwendigkeit für das Auftreten von K und S ist allein die minimal hinreichende Bedingung. Damit wird aber Toms Danebenwerfen nach der minimalen Theorie als Ursache ausgeschlossen, obwohl es in unserer Geschichte eindeutig eine Ursache des Krugzerbrechens darstellt. Das ist wiederum ein Spezialfall von »Preemption«, der der minimalen Theorie direkt Schwierigkeiten bereitet und vor allem Probleme verursacht, wenn wir auf die singuläre Ebene übergehen.

Eine generelle Problematik ist dabei, dass die minimale Theorie auf der Ebene der Typen von Ereignissen und der kausalen Relevanz solcher Typen untereinander recht *liberal* ist (wie die meisten Ansätze auf dieser Ebene). Ist etwa das *Husten einer Person* relevant für das Entflammen des danebenstehenden Hauses? Ja, natürlich, denn es ist folgende Konstellation kausal möglich und sollte daher gemäß dem Vollständigkeitsprinzip auch realisiert sein: Klaus ist so psychisch disponiert, dass er das Haus in Brand steckt, sobald Ute unter bestimmten Randbedingungen hustet. Dann ist das Husten in diesem speziellen Fall tatsächlich eine Ursache für das Entflammen des Hauses und damit liegt eine kausale Relevanz für die entsprechenden Ereignistypen bzw. Eigenschaften vor.

Ist Klaus nur einmal in dieser unguten Stimmung, so handelt es sich dabei um eine deterministische Regularität, wie sie von der minimalen Theorie gefordert wird. Zum Husten müssen hier noch bestimmte Kofaktoren hinzukommen, die etwa nur einmal realisiert sind. Wenn die entsprechenden Bereiche deterministisch sind, so sollte es klar sein, dass für jedes Eigenschaftsvorkommnis von A, das einmal zu einem Vorkommnis von B geführt hat, sogleich gilt, dass A kausal relevant für B ist. Auf der Ebene der Ereignistypen finden sich also sehr viele mögliche kausale Beeinflussungen. Spannend wird es dann oftmals erst, wenn wir uns auf der Ebene der konkreten Ereignisse fragen, ob eine dieser Beziehungen dort instantiiert ist. Der Schritt zur singulären Kausalbeziehung ist deshalb von besonderer Bedeutung und keineswegs trivial. Im Prinzip wird hier wieder ein Schluss auf die beste Erklärung bzw. eine eliminative Induktion erforderlich. Unter den vielen möglichen Kausalbeziehungen müssen wir etwa durch Ausschließen der Konkurrenten diejenigen aussondern, die tatsächlich instantiiert sind.

Die Vertreter der minimalen Theorie weisen selbst noch auf weitere Probleme mit *negativen Faktoren* hin, die wir auch schon angesprochen haben. Während sich die Instanzen positiver Faktoren entsprechend kausal deuten lassen, kann die Zugehörigkeit zu einem abwesenden Faktor gerade auf der singulären Ebene nur schwer als Ursache gedeutet werden. Auch die Einbettung eines singulären Ereignisses in einen Ereignistyp selbst ist ein induktiver Schluss, der wieder auf die Regeln normalen induktiven Schließens angewiesen ist. Insgesamt können wir einen induktiven Schluss erkennen, der danach fragt, wie die Abfolge zweier Ereignisse am besten zu erklären ist. Entweder als zufällige Koinzidenz oder als Instanz bestimmter Kausalgesetze, die wir schon kennen. Dazu müssen wir alle bekannten Umstände der jeweiligen Situation auswerten und (SV) ist dabei sicherlich ein wichtiges Schema, aber ebenso sicher keines, das automatisch oder deduktiv zu den richtigen Ergebnissen führt. Die Autoren (Baumgartner & Grasshoff 2004) sind in diesem Punkt deutlich zu optimistisch.

7.2.5 Kausales Schließen und Homogenisierung

Beim kausalen Schließen können wir vier unterschiedliche Typen von Schlüssen unterscheiden. Zum einen können wir aus bekannten Kausalstrukturen auf der Ebene der Typen von Ereignissen auf Zusammenhänge auf der *singulären Ebene* schließen. Das haben wir im letzten Kapitel besprochen und gesehen, dass diese Schlüsse keinesfalls so einfach sind, wie sie manchmal dargestellt werden. Zum anderen können wir bei bekannter Kausalstruktur und vorliegenden Wirkungen oft in Form *diagnostischer Schlüsse* auf die vermutlichen Ursachen schließen. Das hatten wir im Rahmen der bayesschen Netze schon kennengelernt und ebenfalls als typische Fälle von Faktenabduktionen besprochen. Des weiteren können wir aus dem Vorliegen von Ursachen und der Kenntnis der Kausalstruktur auf die wahrscheinlichen oder sogar sicheren Wirkungen schließen im Sinne *prognostischer Schlüsse*. Der für uns wichtigste Fall kausalen Schließens besteht aber in der *theoretischen Abduktion* bzw. dem Schluss auf die vermutliche Kausalstruktur, die zu bestimmten Korrelationen von Typen von Ereignissen geführt hat. Auf diesen letzten Fall werden wir uns hier konzentrieren.

Um kausal im Rahmen der minimalen Theorie auf die zugrundeliegende Kausalstruktur schließen zu können, benötigen wir als kausale Hintergrundannahmen die des *Determinismus*, die der *empirischen Vollständigkeit* und Annahmen darüber, welche singulären Ereignisse tatsächlich *koinzidieren* und für welche das nicht gilt, obwohl sie raumzeitlich benachbart auftreten.

Für die theoretischen Kausalschlüsse sind wir vor allem auf das *Vollständigkeitsprinzip* angewiesen. Mit seiner Hilfe wird angenommen, dass unsere Koinzidenztabellen vollständig sind, wonach alle kausal möglichen Koinzidenzen auch tatsächlich auftreten; daher lässt sich aus ihnen die dahinterliegende Kausalstruktur effektiv ermitteln. Zunächst beginnen wir wie die Autoren der minimalen Theorie mit sehr einfachen Fragestellungen: Wir wollen etwa wissen, *ob ein Faktor A für eine mögliche Wirkung W kausal relevant ist.* Die Grundidee der dazu passenden *Differenzmethode* findet sich schon bei John Stuart Mill und ist recht einfach:

Mills Differenzmethode: Wir wählen zwei Testsituationen T_1 und T_2 so, dass in T_1 gerade A und W instantiiert ist und in T_2 beide nicht vorliegen. Stimmen T_1 und T_2 ansonsten in allen Umständen überein, so muss A für das Auftreten von W verantwortlich sein.

Im Rahmen der Differenzmethode wird gezeigt, dass A der alleinige »*Unterschiedsmacher*« für das Auftreten von W ist. Da die Testsituationen in allen anderen Bedingungen übereinstimmen, können wir nur A für das Vorliegen von W verantwortlich machen bzw. nur A kann den Unterschied *erklären*. Unsere Grundannahmen über die kausale Struktur in unserer Welt lassen uns dann darauf schließen, dass A eine Ursache von W sein muss, wenn es z.B. ausgeschlossen ist, dass W Ursache von A ist. Diese Suche der Differenztests nach einem verantwortlichen Unterschiedsmacher ist hier unsere sehr plausible Leitidee, mit der wir uns immer wieder die Daten ansehen können.

Sie ist auch die Grundidee moderner *kontrollierter Experimente*, in denen wir durch Kontrolle der relevanten Faktoren genau so eine spezielle Differenz herbeiführen möchten. Allerdings hat das Ergebnis nur

dann Aussagekraft, wenn die genannten *Randbedingungen* tatsächlich erfüllt sind, die anderen Bedingungen in den beiden Situationen also gleich sind. Doch natürlich finden wir keine so umfassend gleichartigen Bedingungen, wie von Mill gefordert in der Praxis vor. Es genügt aber schon, wenn T_1 und T_2 in allen für das Auftreten von W *relevanten Bedingungen* übereinstimmen. Insbesondere müssen wir alle möglichen *Störfaktoren* (oder Konfundierer) für das Verhältnis (A,W) ausschalten, bzw. dafür sorgen, dass die Situationen *kausal homogen* im Hinblick auf W sind bis auf A, also nur A ein kausaler Unterschiedsmacher ist.

Welche Faktoren bzw. welche potentiellen Störfaktoren müssen wir für einen Differenztest homogenisieren? Als *Störfaktor* wollen wir solche Faktoren ansehen, die uns zu Fehlschlüssen beim Differenztest verleiten können, wenn wir sie nicht homogenisieren für unsere beiden Testsituationen. Ein solcher Störfaktor F für die Frage, ob A W verursacht, ist ein Faktor, der das Auftreten von W beeinflusst, dies aber kausal unabhängig von A tut. Das bedeutet also, dass F entweder eine direkte Ursache von W ist, die nicht von A verursacht wird, oder auch eine gemeinsame Ursache von A und W darstellt und natürlich wiederum nicht von A verursacht wird.

Nehmen wir an, wir hätten unsere Testergebnisse (T_1: A und W und in T_2: non-A und non-W). Dabei könnte A nur dann unwirksam bzgl. W sein, wenn es einen solchen Störfaktor F gäbe, der in T_2 vorliegt, aber nicht in T_1 (oder umgekehrt), denn der hätte den Unterschied zwischen T_1 und T_2 herbeiführen können, wenn er eine Ursache von W wäre. A wäre in diesem Fall nicht mehr der (alleinige) Unterschiedsmacher zwischen unseren beiden Testsituationen. So könnte z.B. das Auftreten von W in T_1 auf einen Faktor F zurückzuführen sein, der eine Ursache von W darstellt, der aber in T_2 fehlt. Dann muss eben nicht mehr A der verantwortliche Unterschiedsmacher für das Auftreten von W sein. Oder F liegt nur in T_2 vor und verhindert dort das Auftreten von W, während andere Faktoren ohne das Vorliegen von F wiederum zu W geführt hätten. Also müsste auch in diesem Fall A nicht mehr der entscheidende Unterschiedsmacher sein. Es sind vielleicht andere Faktoren B, die W hervorrufen und A ist kausal irrelevant für W. Damit das Ergebnis unseres Tests also kausal interpretierbar ist, muss jeder potentielle Störfaktor, der in T_1 auftritt, auch in T_2 vorliegen und umgekehrt. Wir sagen, dass die

beiden Situationen dann *homogen* bzgl. des Paars (A, W) sind. Dann ist A nachgewiesenermaßen der Unterschiedsmacher zwischen unseren beiden Testsituationen und lässt sich so als Ursache von W erkennen.

Welche Faktoren können demnach Störfaktoren darstellen? Ein Faktor F ist nach unserer Definition ein typischer Störfaktor, wenn er einen *Unterminierer der Differenzregel* darstellt. F ist dagegen kein Störfaktor, wenn er zwar kausal relevant für W ist, aber auf einem kausalen Pfad liegt, der über A nach W führt und nur auf diesem Wege kausal relevant für W ist. (Gemeinsame Ursachen von A und W sollten also im Normalfall homogenisiert werden.) Insbesondere *Zwischenfaktoren*, über die A seine Wirkung auf W ausübt (und auch *Vorfahren* von A, die ihre Wirkung auf W nur über A entfalten), dürfen natürlich nicht homogenisiert werden, da sonst der erhoffte Unterschied, den A in T_1 und T_2 bedeutet, nicht zum Tragen käme. Alle anderen Faktoren, die Ursachen von W sind, sind zu homogenisieren. Hier hilft uns wieder die Redeweise der gerichteten Graphen, um diese Unterscheidung zu treffen. Sind die kausalen Wege durch gerichtete Pfade in einem solchen Graphen charakterisiert, so sind die Faktoren F, die auf einem gerichteten Pfad über A nach W liegen, keine Störfaktoren, wenn sie dort zwischen A und W liegen, oder auf dem Pfad vor A liegen, aber kein anderer Pfad von F nach W führt, der nicht über A läuft. Die anderen kausal relevanten Faktoren für W sind Störfaktoren. Die Anwendung der Differenzregel verlangt hier also schon gewisse Kenntnisse von Kausalzusammenhängen.

Besonders störend ist ein »common cause« von A und W (also ein Faktor, der sowohl A wie auch W verursacht), aber ebenso problematisch ist ein von A ganz unabhängiger Faktor, der zu W führt. Ist F ein Störfaktor müssen die beiden Prüfsituationen im Hinblick auf diesen Faktor homogenisiert werden. Genau genommen können wir diese Bedingung im Hinblick auf MT noch etwas abschwächen (vgl. Baumgartner & Graßhoff 2004). Für jeden Störfaktor F gibt es ein Faktorbündel Φ, das ihn enthält und eine minimal hinreichende Bedingung für W ist. Die *Homogenisierung* bzgl. F verlangt nur, dass mindestens ein Faktor aus Φ genau dann in T_1 fehlt, wenn mindestens ein (möglicherweise anderer) Faktor aus Φ in T_2 fehlt, weil dann das Bündel Φ für W in beiden Situationen nicht wirksam wird oder, sollte kein Faktor in beiden Fällen fehlen, in beiden Fällen wirksam ist. Die Anwendung der

abgeschwächten Anforderung verlangt allerdings noch viel genauere Kenntnisse der kausalen Zusammenhänge (hier sogar der jeweiligen Kofaktoren) und ist daher in der Praxis wohl nur in besonderen Fällen einsetzbar.

Für die *Kofaktoren* von A (als möglicher Ursache von W) gilt dabei, dass sie in beiden Situationen alle vorliegen müssen, wenn wir A als Ursache von W identifizieren wollen. Fehlen bestimmte Kofaktoren von A in T_1, werden wir A normalerweise durch unseren Test nicht als Ursache von W identifizieren können. Negative Aussagen der Art, dass A keine Ursache von W ist, sind mit dem Differenztest jedoch nicht erzielbar. Das Fehlen von Kofaktoren führt also lediglich zu dem Fehlen von kausal interpretierbaren Daten und noch nicht gleich zu kausalen Fehlschlüssen. Trotzdem werden wir sie der Einfachheit halber weiter als Störfaktoren betrachten, weil sie unter unsere einfache Definition von Störfaktoren fallen:

> Ein ***Störfaktor*** für den Differenztest, ob A eine Ursache von W ist, ist jeder Faktor F, der eine Ursache von W darstellt, die auf einem kausalen Pfad liegt, der nicht über A zu W führt.

Die Bestimmung von *Störfaktoren* lässt sich übertragen auf den komplexeren Fall des Tests, welche der n Faktoren $A_1,\ldots,A_n$ Ursachen von W sind. Ein Faktor F ist dafür ein Störfaktor, wenn er nicht auf einem kausalen Pfad liegt, der über eines der A_i zu W führt, aber trotzdem eine Ursache von W darstellt. Für den einfachen Fall müssen wir nun nur noch ermitteln, welche der folgenden Situationen tatsächlich auftreten und können daraus unsere Schlüsse ziehen:

Tabelle 7.2: Mögliche Koinzidenzen von A und W in einem homogenen Test. Jede Spalte entspricht einer Situation (bzw. möglichen Welt).

A	S1	S2	S3	S4
1	1	1	0	0
0	1	0	1	0
Schluss	?	A$\Rightarrow$W	$\neg$A $\Rightarrow$W	?

Unter den Situationen ist jeweils vermerkt, ob W auftritt (1) oder nicht (0). Nur wenn wir die Situationen S2 und S3 beobachten können, können wir auch Schlüsse daraus ziehen, die anderen Situationen sagen uns nichts über die Frage, ob A Ursache von W ist. In S1 und S4 ist A kein

Unterschiedsmacher für W, was jedoch keine Schlüsse auf seine kausale Wirksamkeit zulässt, denn in S1 könnte es andere homogenisierte Faktoren B geben, die neben A ebenfalls zu W führen, und in S4 könnte einfach ein für A erforderlicher Kofaktor fehlen, weshalb A (obwohl eigentlich wirksam) in dieser speziellen Situation eben nicht zu W geführt hat.

Gehen wir nun einen Schritt weiter und nehmen an, wir möchte neben A noch einen zweiten Faktor B mit heranziehen, um das Zusammenspiel zweier möglicher Ursachen zu ermitteln. Dabei gehen wir hier davon aus, dass A jedenfalls kausal relevant für W ist (das hat der einfache Differenztest ergeben) und möchten nur noch ermitteln, ob B das ebenfalls ist, und in welchem Verhältnis B zu A steht, ob es etwa zu Interaktionen zwischen A und B kommt bzw. ob B ein Kofaktor von A ist oder ob B etwa eine von A unabhängige Ursache von W darstellt. Außerdem nehmen wir an, dass in unseren Situationen weder die A-Instanzen Ursache von B-Instanzen noch umgekehrt die B-Instanzen Ursachen der A-Instanzen sind. Auf diese spezielle Annahme der kausalen Unabhängigkeit der Testfaktoren sind wir offensichtlich angewiesen. Sie setzt allerdings wiederum darauf, dass wir bereits über ein gewisses kausales Vorwissen verfügen. Erst dann können wir den »Vierertest« im Sinne der Autoren Baumgartner & Graßhoff (2004) auswerten. Im Rahmen dieses Tests betrachten wir alle Konstellationen von A und B, die auftreten können, und nehmen an, alle Situation S_i seien jeweils für W homogenisiert bis auf A und B. Dann erhalten wir etwa die folgenden möglichen Situationen:

A	B	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16
1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	1	0
1	0	0	1	0	0	1	0	0	1	1	0	1	1	0	1	1	0
0	1	0	0	1	0	0	1	0	1	0	1	1	0	1	1	1	0
0	0	0	0	0	1	0	0	1	0	1	1	0	1	1	1	1	0

Tabelle 7.3: Alle möglichen Koinzidenzen der Faktoren A, B und W in einem homogenisierten Test. Jede Spalte entspricht einer Situation (bzw. möglichen Welt).

In dieser Tabelle stehen vorne die Kombinationen der Faktoren A und B und dann die 16 unterschiedlichen Situationen, in denen jeweils die

Wirkung W auftritt oder nicht. Eine Spalte entspricht so einer Versuchsserie mit identischen relevanten Faktoren, wobei nur die Prüffaktoren A und B und die Wirkung W variieren können. Einige der Situationen lassen bestimmte kausale Schlüsse zu. So tritt in S1 z.B. W nur dann auf, wenn A und B gemeinsam vorliegen. In derselben Situation (zumindest gleich bzgl. aller möglichen Störfaktoren), in der B ohne A vorliegt, tritt kein W mehr auf. Das spricht dafür, dass in einer solchen Welt B ein Kofaktor von A bei der Hervorbringung von W ist, jedenfalls in einer ganz bestimmten Konstellation anderer relevanter Faktoren. Man könnte auch sagen, dass hier A&B der Unterschiedsmacher ist.

Einige Situationen wie S15 oder S16 lassen keine weiteren Schlüsse zu und andere sind sogar mehrdeutig wie etwa S13 und S14 (vgl. Baumgartner & Graßhoff 2004). Die minimalen Theorien, die aus den Situationen jeweils folgen, stellen wir wiederum in einer Tabelle zusammen:

S1	$ABX \vee Y \Longleftrightarrow W$	S9	$AX_1 \vee (\neg B)X_2 \vee Y \Longleftrightarrow W$
S2	$A(\neg B)X \vee Y \Longleftrightarrow W$	S10	B ist kein Kofaktor von $(\neg)$A
S3	$AX_1 \vee (\neg A)BX_2 \vee Y \Longleftrightarrow W$	S11	$AX_1 \vee BX_2 \vee Y \Longleftrightarrow W$
S4	$AX_1 \vee (\neg A)(\neg B)X_2 \vee Y \Longleftrightarrow W$	S12	$AX_1 \vee (\neg B)X_2 \vee Y \Longleftrightarrow W$
S5	B ist kein Kofaktor von A	S13	$ABX_1 \vee (\neg A)X_2 \vee Y \Longleftrightarrow W$?
S6	$AX_1 \vee BX_2 \vee Y \Longleftrightarrow W$	S14	$A(\neg B)X_1 \vee (\neg A)X_2 \vee Y \Longleftrightarrow W$?
S7	$ABX_1 \vee (\neg A)(\neg B)X_2 \vee Y \Longleftrightarrow W$	S15	kein Schluss möglich
S8	$A(\neg B)X_1 \vee (\neg A)BX_2 \vee Y \Longleftrightarrow W$	S16	kein Schluss möglich

Tabelle 7.4: Schlussfolgerungen auf bestimmte minimale Theorien in den einzelnen Situationen.

Der Faktor AX_1 stammt jeweils aus unserem Vorwissen. In den Situationen S13 und S14 könnte B bzw. $(\neg B)$ ein Kofaktor von A sein, wie das hier notiert wurde, aber B bzw. $(\neg B)$ könnte auch ein eigenständiger Faktor für W sein und dann hätten wir in S13 etwa die minimale Theorie $ABX_1 \vee (\neg A)X_2 \vee BX_3 \vee Y \Longleftrightarrow W$.

Problematisch und nicht definitiv überprüfbar ist in diesem Testverfahren zunächst wieder die Homogenitätsanforderung, aber noch problematischer ist die Forderung, dass wir auch tatsächlich alle relevanten und möglichen Kombinationen von Faktoren bereits beobachten konnten. Es könnte z.B. sein, dass in unseren Experimenten bestimmte wesentliche

Kofaktoren von B nie vorlagen und daher die Wirksamkeit von B für W nie beobachtet wurde. Oder sie fanden irgendwo statt, aber wir haben es eben nicht gesehen. Jedenfalls können wir nur *induktiv* darauf schließen, dass wir schon die möglichen Fälle alle beobachtet haben und das ganze Verfahren hat damit eindeutig einen induktiven Charakter und nicht, wie Baumgartner & Graßhoff (2004) annehmen, die Form eines deduktiven Schlusses.

Wir erhalten jedenfalls immer nur einige Spalten aus dem Raum der möglichen Situationen und wissen nicht sicher, ob es noch andere Fälle von Randbedingungen gibt, die zu anderen Spalten führen würden, die wir bisher noch nicht beobachtet haben. Sind wir uns dessen sicher, dann können wir allerdings in manchen Fällen im Sinne der eliminativen Induktion durch unsere Beobachtungen viele kausale Zusammenhänge ausschließen und bleiben vielleicht nur noch mit einer Hypothese zurück, die dadurch sehr gut gestützt wird. Verletzungen der Homogenitätsforderung sind in unserem Testverfahren übrigens wieder typische Unterminierer, die den Schluss untergraben und so Wissensansprüche verhindern.

Einige Fälle sehen zudem recht seltsam aus. In S3 z.B. finden wir einen eigentümlichen Zusammenhang. Wir wissen schon, dass A kausal relevant für W ist, nur dass hier einer der Kofaktoren nicht vorlag. Aber gleichzeitig ist demnach auch non-A kausal relevant für W, wobei non-B einer der Kofaktoren ist. Daher schließen wir auf $AX_1 \vee$ (non-A) (non-B) $X_2 \vee Y \Longleftrightarrow W$. Das ist zwar rein theoretisch möglich, wäre allerdings schon seltsam, da non-A gerade bedeutet, dass keine Instanz von A vorliegt. Wenn nun eine Instanz zu W führen kann, aber auch das Fehlen einer Instanz von A, würden wir uns vermutlich stärker auf die anderen Faktoren beziehen und die jeweils als Ursachen von W betrachten und nicht so sehr das Auftreten oder Fehlen von A. Das hat auch etwas mit der Problematik zu tun, *negative Faktoren* (also die bloße Abwesenheit eines Faktors) als Ursachen begreifen zu können. Ähnlich seltsame Konstellationen finden wir daneben in anderen Fällen, die wohl mehr der theoretischen Vollständigkeit halber zu diskutieren sind, als dass man erwarten darf, in der Praxis auf solche Fälle zu stoßen.

Noch einmal zurück zur Homogenitätsforderung: Für bekannte Störfaktoren können wir in bestimmten Experimenten vielleicht dafür

sorgen, dass sie in allen Situationen abwesend sind, denn das ist die ergiebigste Konstellation für einen Kausaltest. Doch gerade in den Sozialwissenschaften kennen wir viele Faktoren nicht oder können nicht alle vorstellbaren Experimente durchführen, in denen wir bestimmte Faktoren einfach verändern können (vgl. Kelle 2003).

Möchten wir z.B. herausfinden, warum Kinder aus der sozialen Oberschicht erfolgreicher in der Schule sind als Kinder unterer Schichten (W) und vermuten, dass es (A) der Ausbildungsstand der Eltern sein könnte, die dann ihren Kindern schon in der vorschulischen Erziehung bessere sprachliche Fertigkeiten und ein bestimmtes Wissen mitgeben, oder (B) einfach doch nur die bessere finanzielle Ausstattung dieser Kinder, die es gestattet, mehr Bücher und andere Hilfsmittel zu kaufen oder auf anderen Wegen zu mehr Selbstvertrauen und einem höheren sozialen Ansehen in der Schule führt, so werden wir den zweiten Faktor B als möglichen Störfaktor für unseren ersten Test trotzdem nicht einfach ausschalten können. Weder die Kinder noch die Eltern aus der Oberschicht werden zustimmen, dass sie für alle Schuljahre komplett auf ihr Geld verzichten. (Die Kinder aus den unteren Schichten finanziell viel besser auszustatten, würde andererseits vermutlich die finanziellen Möglichkeiten des Experimentierens sprengen.)

Noch problematischer sind natürlich die Fälle, in denen *unbekannte Störfaktoren* vorliegen, die wir nie ganz ausschließen können. Speziell Fälle von unbekannten gemeinsamen Ursachen von A und W sind besonders störend. Sie können den Eindruck einer scheinbaren Wirkung von A auf W hervorrufen, obwohl keine direkte Kausalbeziehung vorliegt. Also ist die *Homogenitätsbedingung* wiederum die zentrale Annahme für das kausale Schließen im Rahmen der minimalen Theorie.

Doch wie soll man diese Bedingung für unbekannte Störfaktoren erfüllen oder ihr Zutreffen feststellen? Da es sich um nicht bekannte Faktoren aus einer Art von Ceteris-Paribus-Bedingung handelt, wird man sie nicht definitiv überprüfen können. Baumgartner & Graßhoff (2004, Kap. X) geben dazu an, wir könnten den Test einfach mehrfach wiederholen, und es wäre dann unwahrscheinlich, dass Prüffaktor und Störfaktor immer in derselben Weise variieren würden. Sie betrachten einen Fall, in dem bei 100 Wiederholungen nur eine Abweichung vom normalen Schema auftritt. So hat man einen Grund, das als Störfaktorszenario anzusehen

und die anderen 99 Fälle als homogen zu betrachten und damit als korrektes Indiz für eine bestimmte Kausalbeziehung.

Doch erstens ist es recht unrealistisch, dass wir in der Praxis ein Experiment so oft wiederholen können. Vor allem müssen die Autoren voraussetzen, dass es gelingt, genau dasselbe Experiment mit denselben Faktorkonstellationen selbst der unbekannten Faktoren hundertmal zu wiederholen. Dabei werden meist schon erste Replikationsstudien in vielen Journalen nicht veröffentlicht und lohnen daher für den Wissenschaftler den Aufwand nicht.

Zweitens ist es unrealistisch, dass wir so einseitige Ergebnisse erhalten, insbesondere dann, wenn wir an den Fall von gemeinsamen Ursachen denken. Solange wir sie nicht kennen, können wir sie nicht definitiv ausschalten und vielleicht ist gerade das, was die Autoren als *genuine Prüffaktor Ursache von A* bezeichnen (mit der wir etwa in der Versuchsgruppe gerade A herbeiführen) ebenso eine versteckte zusätzliche Ursache von W.

In jedem Fall sind wir hier wieder in klassischen Induktionsverfahren gelandet. Wir versuchen anhand der Häufigkeit, mit der etwas auftritt, zu schließen, dass vermutlich eine Homogenisierung vorliegt. Die Idee der Autoren, eine Art von deduktivem Schlussverfahren etablieren zu können, stößt damit eindeutig an ihre Grenzen (vgl. Baumgartner & Graßhoff 2004, Kap. VIII und IX). Für die wichtigste Voraussetzung des kausalen Schließens sind wir wieder auf einfache induktive Schlüsse angewiesen.

Letztlich müssen wir eine Abwägung im Sinne des *abduktiven Schließens* bzw. von Kohärenzüberlegungen vornehmen. Wenn ein Unterschied zwischen Versuchs- und Kontrollgruppe vorliegt, welche der beiden Hypothesen erklärt den dann am besten: 1. *Der Unterschied geht auf den Unterschied im Vorliegen von A zurück* oder 2. *der Unterschied wird von einem Störfaktor hervorgerufen, der in der einen Gruppe vorliegt und in der anderen nicht.* Der Differenztest weist uns also zwar den richtigen Weg zur Ursachenermittlung, aber er ist immer auf weitere kausale Annahmen und Abwägungen angewiesen und keineswegs so etwas wie ein automatisiertes Verfahren zur Ermittlung von Kausalbeziehungen.

Kelle (2003) kritisiert die Idealisierungen des Verfahrens anhand von Beispielen aus den Sozialwissenschaften, in denen die Vollständigkeits-

bedingung offensichtlich nicht erfüllt ist. Dort wird das Verfahren etwa im Vergleich verschiedener Länder eingesetzt, um damit herauszufinden, welche besonderen Bedingungen jeweils vorliegen müssen, damit in einem Land bestimmte Institutionen entstehen oder stabil bleiben. Diese Kritik wendet sich u.a. gegen Arbeiten von Charles Ragin (1987 und 2000), der sich im Rahmen der *qualitativen komparativen Analyse* (QCA) auf ein entsprechendes Verfahren stützte, um Kausalzusammenhänge zu ermitteln. Oft hilft hier schon unser weiteres Hintergrundwissen zu erkennen, dass die Vollständigkeitsbedingung nicht erfüllt ist. Wir können das Verfahren dann etwa als Heuristik nutzen, um herauszufinden, nach welchen Daten, wir gezielt Ausschau halten sollten. Neben den Übereinstimmungen sind das gerade die Fälle von Differenzen, in denen ein bestimmter Faktor als Unterschiedsmacher auftreten könnte. Können wir solche Daten nicht gewinnen oder müssen annehmen, dass unsere Daten unvollständig bleiben werden, so hilft uns die minimale Theorie natürlich nicht mehr weiter.

Kelle (2003, 242) meint, dass dann ein quantitativer statistischer Ansatz oft hilfreicher ist. Führt etwa A&F kausal zu W, aber die Hintergrundbedingung F ist uns nicht bekannt und sie ist in unserer Gesamtpopulation nur selten vertreten, so bliebe uns immer noch die Ungleichung P(W|A) > P(W) als Indikator für die Wirksamkeit von A für W. Doch auch das ist nicht so klar, denn erstens ist die Ungleichung vielleicht nur auf eine gemeinsame Ursache von A und W zurückzuführen und zweitens kann es hier ebenfalls andere Konstellationen geben, die die Ungleichung verhindern, obwohl A im Prinzip wirksam für W ist.

Kelle (2003, 243) diskutiert selbst ein mögliches Beispiel dafür. Es zeigte sich, dass der Faktor (A) *Erfolg in Ausbildung und Beruf* in vielen Fällen negativ korreliert ist mit dem Faktor (W) *Delinquenz*. Das heißt, Personen, die ihre Ausbildung nicht beenden und dann arbeitslos bleiben, zeigen eine größere Neigung zu kriminellen Verhaltensweisen als erfolgreichere Personen. So weit, so gut. Aber zugleich zeigte sich auch eine überraschende Neigung zu kriminellem Verhalten bei bestimmten Jugendlichen mit besonders erfolgreichen Berufsbiographien. Damit führt zum einen non-A zusammen mit bestimmten Hintergrundbedingungen zu Delinquenz, aber ebenso führt A mit anderen Hintergrundbedingungen zu Delinquenz. Nehmen wir an, der erste

Effekt wäre der stärkere, dann gilt: $P(W|A) < P(W)$, aber dennoch ist zugleich A positiv kausal relevant für W, was in unserer Ungleichung nicht aufscheint. Das heißt, der eine der beiden Kausalzusammenhänge würde durch unsere Ungleichung systematisch unterschlagen. Der statistische Ansatz kann also weiterhelfen, muss es aber keineswegs, solange uns die weiteren relevanten Hintergrundbedingungen verborgen bleiben. Häufiger sind dabei vermutlich die Probleme durch »common causes« oder *gemeinsame Ursachen* zu erwarten, auf die wir im nächsten Kapitel wieder zurückkommen werden.

Das kausale Schlussverfahren der minimalen Theorien lässt sich schließlich verallgemeinern auf den Fall von n Prüffaktoren und dann ausdehnen auf die Ermittlung ganzer Kausalketten (vgl. Baumgartner 2009a & 2009b und Baumgartner & Epple 2014). Für kausale Ketten werden mehrere minimale Theorien konjunktiv zu einer größeren Theorie zusammengesetzt. Jede minimale Theorie selbst betrifft dagegen nur den Zusammenhang zweier Ebenen eines kausalen Netzes. Wie in den anderen Ansätzen finden wir aber auch hier wieder Fälle von *beobachtungsäquivalenten Strukturen*, d.h. Koinzidenzstrukturen, die durch unterschiedliche kausale Strukturen hervorgebracht werden können. Einige kausale Strukturen sind grundsätzlich so beschaffen, dass sie generell zu denselben Koinzidenztabellen führen. Baumgartner & Graßhoff (2004) sprechen hier von *verschränkten Kausalfaktoren*. In diesen Fällen hat wieder unser Hintergrundwissen mit zu entscheiden, welches vermutlich die richtige Kausalstruktur hinter den Daten ist.

Für den Fall, dass wir schon die Wirkung W kennen und endliche viele potentielle direkte Ursachen (also ohne kausale Ketten) A, B, C, … vorliegen, die untereinander kausal unabhängig sind (wie die Wurzelfaktoren eines kausalen Netzes), wird gerne der Algorithmus von Quine-McCluskey eingesetzt, um die genaue Kausalstruktur zu ermitteln. Eine entsprechende Vorgehensweise wollen wir uns an einem einfachen Beispiel ansehen. Zunächst haben wir etwa eine Datentabelle wie die folgende, in der alle Kombinationen aller Wurzelfaktoren enthalten sind:

	A	B	C	W
1	1	1	1	1
2	1	1	0	1
3	1	0	1	1
4	0	1	1	1
5	1	0	0	0
6	0	1	0	0
7	0	0	1	1
8	0	0	0	0

Tabelle 7.5: Beispieldaten für 8 Situationen

Im ersten Schritt wählt man nun alle Zeilen mit W=1 und bestimmt so die hinreichenden Bedingungen für W, die wir hier gerade in den Zeilen 1–4 und 7 finden. (Dabei dürfen natürlich keine weiteren identischen Zeilen vorliegen, in denen nur W=0 vorkommt, aber alle anderen Faktoren ebenfalls vorliegen, denn dann wären die Bedingungen nicht wirklich hinreichend.) Wenn wir die Negationen der Faktoren mit kleinen Buchstaben bezeichnen, sind das also die folgenden Faktorbündel hinreichend für W: ABC, ABc, AbC, aBC und abC. Dann können wir durch Zusammenziehen bestimmter Bedingungen nun die *minimal* hinreichenden Bedingungen ermitteln. Der entscheidende Schritt ist, dass wir zwei Zeilen finden vom Typ:

$$\left.\begin{array}{l}\text{ABCDE}\\ \text{ABcDE}\end{array}\right\} \text{AB(c/C)DE}$$

Dadurch wird klar, dass der Faktor C bzw. c einen *redundanten* Teil der beiden Faktorbündel verkörpert und damit letztlich weggelassen werden kann. Er stellt in unseren Situationen jedenfalls keinen Unterschiedsmacher dar. Dieses Verfahren ist zu wiederholen, bis man die tatsächlich minimalen Faktoren gefunden hat. In unserem Beispiel ergibt sich aus den Zeilen 1 und 2: AB(c/C), aber keine weitere Reduktion dieses Terms. Aus aBC und abC erhalten wir a(b/B)C und zusammen mit AbC ergibt sich, dass auch der erste Faktor irrelevant ist: (a/A)(b/B)C. Also ergibt die Minimierung zwei *minimal hinreichende* Faktorbündel: AB und C für W.

Dann geht es noch darum, welche der Disjunktionen aus den minimal hinreichenden Faktorbündeln – z.B. X, Y und Z – nun zusammen für W *minimal notwendig* sind. Das heißt, es muss gelten: W→X∨Y∨Z (und wir können kein Disjunkt dabei weglassen). Liegt also eine Instanz von W

vor, so muss zumindest eines der minimal hinreichenden Faktorbündel realisiert sein, und für jedes Bündel muss es eine Situation geben, in der es alleine dafür sorgt, dass W auftritt. Das heißt, X, Y und Z sind alle möglichen Ursachenbündel von W und wir dürfen erst dann auf ¬W schließen, wenn sie *alle* nicht vorliegen. Um das zu ermitteln, können wir natürlich auch herausfinden, welche Kombination aus ¬X, ¬Y und ¬Z minimal hinreichend für ¬W ist.

Um in unserem Beispiel zu überprüfen, ob auch AB∨C minimal notwendig für W sind, müssen wir nur in der Tabelle 7.4 nach den entsprechenden Zeilen suchen, die anhand der Vollständigkeitsannahme zeigen, dass die beiden Faktoren tatsächlich notwendig sind. Wenn W vorliegt, so muss auch mindestens eines der beiden Bündel vorliegen. Zeile 2 belegt, dass AB für W tatsächlich erforderlich ist und die Zeilen 3, 4 und 7 belegen das für das Bündel C. (in der ersten Zeile liegen sowieso beide Bündel vor.) Dadurch zeigt sich also in unserem Beispiel, dass keine weiteren Minimierungen möglich sind und wir erhalten die *minimale Theorie*: AB∨C ⟺ W.

Wenn wir etwas komplexere Fälle zulassen, und wir es etwa mit der folgenden kausalen Struktur K1 zu tun haben: A⇛C⇚B⇛D, in der *zwei* Wirkungen C und D auftreten, müssen wir etwas anders schließen. Baumgartner (2009b) gibt ein komplexeres Verfahren an, dass für solche Fälle ebenfalls anwendbar ist. Das Problem ist hierbei, dass B immer gleich D zur Folge hat und daher nach dem oben geschilderten Minimierungsverfahren B und D gemeinsam minimal für C zu sein scheinen. Wenn wir also D nicht als mögliche Ursache von C von vornherein ausschließen können, stellt sich die Frage, ob D nicht ein Kofaktor von B bei der Erzeugung von C ist. Dann erhielten wir die alternative kausale Struktur K2 (A⇛C⇚BD), bei der A allein und BD als Kofaktoren Ursachen von C wären. In der Struktur K2 erhielten wir in der Kontingenztabelle dann aber die folgende Zeile: aBcd. Obwohl B vorliegt, wird C nicht realisiert, weil der Kofaktor D nicht realisiert wurde. Diese Zeile kann aber nicht in der Kontingenztabelle für K1 stehen, denn dort wird immer sogleich C auftreten, wenn B vorliegt. Weil also in der Kontingenztabelle von K1 die Zeile aBcd *fehlt*, können wir darauf schließen, dass B bereits allein für C hinreichend ist. Damit können wir das Faktorbündel BD weiter zu B minimieren.

Wir müssen nun also auch das *Fehlen bestimmter Tabellenzeilen* mit heranziehen und schauen, ob wir Fälle finden, in denen die Wirkung nicht auftritt, wenn wir bestimmte Faktoren weglassen, was für ihre Redundanz sprechen würde. Baumgartner (2008b) liefert uns dafür die folgende Art von *Minimierungsregel*:

Verstärkte Minimierungsregel
Sind ABD hinreichend für C, so ist A nur dann *erforderlich* in diesem Faktorbündel, wenn es eine Zeile aBDc in der Kontingenztabelle gibt, die aufzeigt, dass bei Fehlen von A auch C nicht mehr auftritt.

Fehlt dagegen die Zeile aBDc in unserer Kontingenztabelle, so ist also A *redundant* und wir können den Minimierungsschritt von ABD zu BD vollziehen.

Durch wiederholte Minimierung erhalten wir so schließlich wieder die minimal hinreichenden Faktoren in diesen komplexeren Fällen. Allerdings wird dabei deutlich, dass wir mit dieser stärkeren Regel besonders auf eine Vollständigkeit der Kontingenztabellen angewiesen sind, denn wir schließen aus dem *Nichtvorhandensein* bestimmter Zeilen darauf, dass ein Faktor redundant ist. Die obige Vorgehensweise schließt dagegen nur aus vorhandenen Zeilen auf die Redundanz bestimmter Faktoren. Sollte die Kontingenztabelle also an dieser Stelle einfach nur unvollständig sein, ergäbe sich daraus im zweiten Verfahren sofort ein Fehlschluss, der uns nicht so schnell auffallen dürfte. Wir hätten fälschlicherweise damit einen relevanten Kofaktor übersehen. Das dürfte in den meisten Fällen ein unangenehmerer Fehler sein, als fälschlicherweise A für einen relevanten Kofaktor zu halten. Das belegt die besonderen Gefahren der komplexeren Minimierungsregel. Außerdem verbleiben auch hier Mehrdeutigkeiten für noch komplexere Strukturen, was Baumgartner (2009b) an einem Beispiel aufdeckt.

Allerdings finden wir auch schon für sehr einfache, aber unterschiedliche Strukturen übereinstimmende Kontingenztabellen. So hat die kausale Struktur A⇒C⇐B⇒D dieselbe Kontingenztabelle wie A⇒C⇐D⇒B, bei der wir einfach B und D ausgetauscht haben, weil D zunächst nur von B verursacht wird und damit die D- und die B-Spalte übereinstimmen. Baumgartner und Graßhoff können allerdings argumentieren, dass sie

für jede Wirkung mindestens zwei Ursachen verlangen und dann die Unterbestimmtheit wieder aufgelöst würde, sobald noch eine weitere Ursache E für D in unserer Struktur hinzukäme.

Für diese komplexen Verfahren zur Aufdeckung von Kausalbeziehungen anhand von Kontingenztabellen innerhalb der minimalen Theorie gibt es inzwischen im Rahmen der freien Statistik-Programmiersprache R ein Anwendungspaket mit Namen »cna« (s. Ambuehl et al. 2015).

Das kausale Schließen im deterministischen Fall zeigt schon wesentliche Grundprobleme des kausalen Schließens auf. So ist manchmal nur eine bestimmte *Kombination* von Faktoren kausal aktiv bzw. wir müssen kausale *Interaktionen* zwischen unseren Faktoren berücksichtigen. Diese Struktur von Kofaktoren wird in manchen Ansätzen übersehen bzw. ausgeklammert oder zumindest nicht in ihrer vollen Bedeutung berücksichtigt. Der große Vorteil der minimalen Theorie besteht m.E. gerade darin, dass sie die Hilfsmittel bietet, die logische Struktur dieser komplexeren Kausalzusammenhänge übersichtlich aufzuschreiben, wodurch sie uns genau erklärt, welche Art von Regularität wir im Falle bestimmter Kausalzusammenhänge erwarten dürfen bzw. wann wir auf das Vorliegen von Kausalbeziehungen aus einer Regularität schließen dürfen. In anderen Kausaltheorien liegt das Augenmerk oft auf anderen Problemen wie dem, dass die Kausalbeziehung selbst nur probabilistisch ist, und z.B. die Kofaktorenbeziehung bleibt dann meist unbeachtet, weshalb die zugrundeliegenden Kausalstrukturen nicht immer zuverlässig aufgedeckt werden.

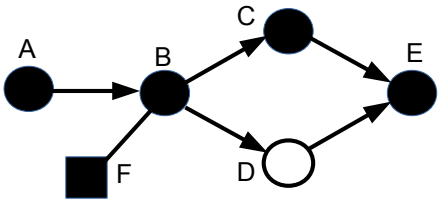
Um kausal schließen zu können, sind wir aber in jedem Fall auf vollständige und zuverlässige Daten bzw. hier Kontingenztabellen angewiesen. Doch selbst wenn die vorliegen, deuten sie nicht immer in eindeutiger Weise auf eine ganz bestimmte kausale Struktur. Man kann nicht oft genug betonen, an wie vielen Stellen wir daher auf kausales Vorwissen angewiesen sind, um weitere Kausalstrukturen aufzudecken. Wir müssen zunächst potentielle Ursachen ausmachen und eine Liste aller potentiellen Ursachen dafür aufstellen. Dann benötigen wir *vollständige* Kontingenztabellen und müssen bei dem Schließen selbst eventuell auf weiteres kausales Hintergrundwissen vertrauen, mit dem sich bestimmte mögliche kausale Hintergrundstrukturen ausschließen lassen. Diese Suche nach einer Kausalstruktur, die unsere Daten am

besten erklärt, zeigt wieder alle Merkmale eines Abduktionsschlusses und ist keinesfalls von deduktiver Natur, wie uns Baumgartner & Grasshoff (2004) optimistisch suggerieren möchten.

Eine weitere Komplikation kommt noch dadurch ins Spiel, dass einige Faktoren quantitativ sind. So kann ein Medikament in bestimmten Dosierungen hilfreich sein und zur Heilung beitragen, und in noch höheren Dosierungen bereits schädlich wirken. Zum Teil können wir das auf die bisherigen Erörterungen zurückführen, indem wir unterschiedliche Dosierungen als unterschiedliche Faktoren betrachten, aber es ergeben sich auch einige neue Phänomene, auf die wir weiter unten eingehen werden.

7.2.6 Komplexe Ursachenketten

Beim kausalen Schließen geht es aber nicht nur darum, *direkte* Ursachen zu ermitteln, sondern auch darum, in komplexeren Situationen mit längeren Kausalketten die *indirekten Ursachen* zu ermitteln. Das hat sich inzwischen als ein schwieriges Geschäft erwiesen, weil es sich nicht einfach auf die Kenntnis der direkten Ursachen zurückführen lässt und unsere Vorstellungen von Verursachung dabei auch nicht immer völlig eindeutig sind. Man kann wohl sagen, dass alle bekannten Ansätze damit gewisse Probleme haben. Das liegt vor allem daran, dass man nicht mehr voraussetzt, dass die Kausalbeziehung *transitiv* ist, wodurch sich die indirekte Kausalität auf die direkten Verursachungsbeziehungen reduzieren ließe. Einen wesentlichen Anteil an der Ablehnung der Transitivitätsannahme hatte dabei das folgende Beispiel, das wir als *Schalterbeispiel* bezeichnen können:



Graphik 7.1: Neuronendiagramm (Schalter) aus Baumgartner 2013

In solchen Neuronendiagrammen sollen die ausgefüllten Kreise aktivierte Neuronen darstellen und die leeren Kreise deaktivierte Neuronen. Pfeile stellen Kausalbeziehungen dar und geben die Aktivierung weiter, während Verbindungen mit kleinen kreisen an der Spitze deaktivierend wirken und sogar eine Aktivierung übertrumpfen. Mit Hilfe der Neuronendiagramme lassen sich nun kausale Beziehungen schematisch vereinfacht darstellen, weshalb sie inzwischen gerade auch in Kreisen der kontrafaktischen Ansätze gerne eingesetzt werden (s.z.B. Paul & Hall 2013, Baumgartner 2013, Hall 2007, Hitchcock 2009).

Man könnte sich das Beispiel etwa durch Stromkreise realisiert denken oder durch Züge, die am Punkte B von einer Weiche, die von F gestellt wird (dem Schalter) entweder auf dem Weg über C oder auf dem Weg über D nach E gesandt werden. Die Frage ist nun, ob die Schalterstellung (Weichenstellung) von F eine Ursache von E (dem Ankommen des Zuges in E) darstellt.

Die Leitidee soll auch hier sein, dass Ursachen Unterschiedsmacher zumindest in bestimmten Situationen sind. Zunächst bestimmt F, ob der Zug in C auftaucht oder nicht. F ist damit Unterschiedsmacher und erkennbare Ursache für das Ereignis C. C ist nun seinerseits Unterschiedsmacher für E, denn wenn zu dem entsprechenden Zeitpunkt etwa schon klar ist, dass kein Zug durch D kommt, ist das Fahren des Zuges durch C eine Ursache dafür, dass er dann bei E auftaucht. Würden wir die Kausalbeziehung als transitiv ansehen, sollte damit F auch eine Ursache von E sein. Doch das passt nicht zu unserer Konzeption, denn für das Auftreten von E macht es keinen Unterschied, welche Stellung unser Schalter F einnimmt, da der Zug ja auf beiden Wegen zu E gelangt. Intuitiv sollten wir daher sagen, dass F keine Ursache von E darstellt.

Damit stehen wir allerdings vor der neuen Aufgabe, die indirekten Ursachen eigens zu bestimmen. Baumgartner (2013) schlägt dazu vor, auch *indirekte minimale Theorien* zu ermitteln, die nun nach demselben Schema erstellt werden wie die direkten minimalen Theorien, sich aber auf weiter entfernte Faktoren in unserem Diagramm beziehen bzw. auf zeitlich früher instantiierte Faktoren. Zusätzlich zu den direkten minimalen Theorien müssen wir also die indirekten anführen, um auch die komplexeren kausalen Zusammenhänge in Fällen komplexerer Kausalbeziehungen zu ermitteln.

In unserem Beispiel ergibt sich damit die gesamte minimale Theorie als:

$$(A \Longleftrightarrow B) \,\&\, (BF \Longleftrightarrow C) \,\&\, (B\,(\text{non-}F) \Longleftrightarrow D) \,\&\, (C \vee D \Longleftrightarrow E) \,\&$$
$$(AF \Longleftrightarrow C)_i \,\&\, (A(\text{non-}F) \Longleftrightarrow D)_i \,\&\, (A \Longleftrightarrow E)_i \,\&\, (B \Longleftrightarrow E)_i$$

Dabei sind die *indirekten minimalen Theorien* durch den Index i gekennzeichnet. Hier wird schon deutlich, dass sich F nicht unter den (indirekten) Ursachen von E befindet – wie das eigentlich gewünscht erschien. Dazu schreiben wir $F \in MT(E)$, dafür, dass es eine (direkte oder indirekte) minimale Theorie für E gibt, für die F unter den Faktoren der linken Seite befindet. Dann gilt in unserem Fall: $F \notin MT(E)$.

Durch die fehlende Transitivität treten auch beim Übergang auf die *singuläre* Ebene neue Probleme für komplexere Kausalketten auf, die in Baumgartner (2013) wie folgt behandelt werden:

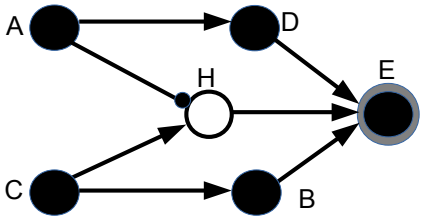
> **Singuläre Ursache:** Damit $a \in A$ eine Ursache von $e \in E$ ist, sollten zwei Bedingungen erfüllt sein: 1. $A \in MT(E)$ und 2. es gibt eine Kette (einen Pfad) von Faktoren $A=A_1, A_2, \ldots, A_n=E$, so dass gilt: $A_i \in MT_{direkt}(A_{i+1})$ und es sind de facto alle jeweils erforderlichen Kofaktoren der A_i instantiiert, so dass die Instanzen der A_{i+1} auch tatsächlich verursacht werden, und es somit einen *aktiven kausalen Pfad* von A nach E gibt.

Mit dieser Konzeption lassen sich viele Problemfälle der kontrafaktischen Ansätze wie das Preemptionproblem oder das der Überdetermination oder das eines »Kurzschlusses« auf recht einfache Weise behandeln (s. Baumgartner 2013 und vergleiche mit Hall 2007). Die kontrafaktischen Ansätze mit Strukturgleichungen (wie der von Hitchcock 2001) geraten dagegen schon bei unserem Schalterfall ins Straucheln (s. Hall 2007 und dazu wiederum Hitchcock 2009). Sie verlangen typischerweise, dass es einen kausalen Pfad von A nach E gibt und dass, wenn wir die Zustände der Knoten außerhalb des Pfades festhalten, eine kontrafaktische (de facto) Abhängigkeit zwischen A und E vorliegt.

Damit kommen diese Ansätze allerdings schon in unserem Schalterbeispiel zu einem anderen Ergebnis als dem gewünschten. Wenn wir

nämlich festhalten, dass D inaktiv ist, dann führt erst die gewünschte Schalterstellung von F dazu, dass der Weg über C aktiviert wird. Damit wäre auch F eine Ursache von E, obwohl nach dem intuitiven Differenzkriterium in der eigentlichen Situation die Schalterstellung von F keinen Einfluss auf die Aktivierung von E haben kann, und damit keine Ursache von E darstellen sollte. Die ausgefeilteste Variante der kontrafaktischen Ansätze, bei der man anhand eines kausalen Modells, das die allgemeinen kausalen Zusammenhänge zwischen bestimmten Faktoren mit Hilfe von Strukturgleichungen beschreibt, auf die Ursachenbeziehungen im konkreten Einzelfall schließt, findet sich aber wohl in dem Aufsatz von Halpern & Pearl (2005), in dem viele komplexe Beispiele daraufhin untersucht werden, welche singulären Kausalbeziehungen jeweils vorliegen. Man wird also weiter erforschen müssen, welcher Ansatz zum Schluss die Nase vorn hat.

Diese strikte Orientierung am Differenzkriterium scheint allerdings in anderen Fällen auch problematisch zu sein. Baumgartner diskutiert dazu die folgende etwas komplexere Situation:



Graphik 7.2: Der sture Knoten nach Baumgartner

Hierbei ist E ein sogenannter sturer oder widerspenstiger Knoten (»stubborn neuron«), der erst dann aktiviert wird, wenn er von zwei anderen Knoten einen positiven Impuls bekommt. Gemäß dem Diagramm sieht es also intuitiv betrachtet so aus, als ob A und C die Ursachen für die Aktivierung von E darstellen, weil die weitere Wirkung von C über H dort von der entsprechenden Deaktivierung von A her ausgeschaltet wird, was durch die kleine Kugel am Ende der Verbindung von A nach H repräsentiert wird. Aber was sagt uns das Differenzkriterium über den Einfluss von A? Wäre A inaktiv, so würde C noch zusätzlich H aktivieren

und damit würden wiederum zwei aktivierende Pfeile bei E eingehen und E somit aktivieren.

Gemäß dem reinen Differenzprinzip macht A also keinen Unterschied für das Auftreten von E und sollte daher nicht als Ursache angesehen werden. Das kommt auch in der minimalen Theorie zum Ausdruck, in der als einzige indirekte minimale Theorie C $\Longleftrightarrow$ E auftritt. Der minimale Theorie Ansatz erklärt also, dass A keine Ursache von E darstellt. Doch das passt nicht gut zu unserer intuitiven Beschreibung der Situation, die wir oben schon gegeben haben. Danach bewirken A und C zusammen, dass E aktiviert wird. Das scheint die tatsächlichen kausalen Verhältnisse zu beschreiben und so sieht es auch der Strukturgleichungenansatz. Wenn wir nämlich erst einmal zur Kenntnis nehmen, dass H inaktiv ist, dann hängt die Aktivierung von sowohl von A wie von C ab.

Diese Analyse wird natürlich auch gestützt durch Überlegungen, die an die Transitivität von Kausalität anknüpfen. Offensichtlich sind D und B Ursachen von E. D wird aber seinerseits unbestreitbar von A verursacht. Unsere Vorstellung, dass Kausalbeziehungen im Normalfall transitiv sein sollten, stützt daher wiederum die Analyse, dass A auch eine Ursache von E darstellt (vgl. dazu Paul & Hall 2013, 89 ff.)

An dieser Stelle ist unsere intuitive Kausalkonzeption anscheinend nicht so klar, dass wir den Fall einfach entscheiden könnten. Damit ist auch nicht klar, welche Konzeption in solchen Beispielen die Nase vorn hat. Wir müssen vermutlich die Gesamtleistung der verschiedenen Ansätze (auch anhand vieler klarer Beispiele) gegeneinander abwägen. Das sieht bisher ganz gut für die minimale Theorie aus.

Wir werden im nächsten Kapitel sehen, dass das sture Neuron zugleich ein Beispiel für einen Verstoß gegen die sogenannte Graphentreue abgibt, und es wird im Falle des probabilistischen Ansatzes A genauso wenig als Ursache anerkannt, weil dieser sich ebenfalls auf das Differenzprinzip stützt. Nach diesem Prinzip muss es zumindest eine Situation geben, in der eine Ursache auch einen Unterschiedsmacher für die Wirkung darstellt. A wäre demnach eher als *Epiphänomen* einzustufen, obwohl das Diagramm zunächst eine andere Sprache spricht. Dieser Fall bietet uns wohl einen Grenzfall für unsere Kausalvorstellungen und wir sollten ihn vermutlich nicht als Stolperstein für die Kausaltheorien betrachten, sondern ihn umgekehrt anhand unserer besten Kausaltheorien auflösen.

Inzwischen kursieren viele weitere Beispiele, in denen wir uns nicht so sicher sind, was unsere Kausalvorstellungen dazu sagen. Einige stützen sich etwa auf negative Faktoren. Die beurteilen wir in einige Fällen klar als Ursachen: »Weil Franz nicht gebremst hat, ist er seinem Vordermann aufgefahren.« Aber in anderen Beispielen sind wir uns nicht mehr so sicher, wie bei Angela Merkel, die meine Blumen nicht gegossen hat.

Es werden auch gerne etwas komplexere Beispiele konstruiert, die uns intuitive Probleme bereiten. Nehmen wir an, Ute wirft einen Stein auf eine Scheibe. Franz fängt ihn ab. Dann ist dieses Abfangen eine Ursache dafür, dass die Scheibe nicht zerspringt. Nehmen wir aber an, dass nach Franz noch John ins Spiel kommt, der als perfekter Fänger, den Stein in jedem Fall abgefangen hätte, hätte Franz nicht zugegriffen. Fängt Franz den Stein aber weiterhin ab, bleibt er wohl die Ursache für Unversehrtheit der Scheibe und wir können John eigentlich nicht dafür loben. Wird aber John nun durch eine stabile Mauer ersetzt, die ebenfalls vor der Scheibe steht und den Stein zuverlässig abgewehrt hätte, scheint sich der Fall wieder zu verändern. Der tatsächliche Abfänger Franz scheint nun keine Ursache mehr für die Unversehrtheit der Scheibe zu sein, denn Ute kann sowieso nicht durch die Mauer werfen. Hier scheint also die Mauer nun zum entscheidenden Faktor für die Unversehrtheit der Scheibe zu werden, obwohl sie doch kausal keine andere Funktion übernimmt als der perfekte Fänger John.

Man sieht hieran, dass unsere intuitive Kausalkonzeption nicht so weit ausgearbeitet ist, dass sie komplexere Situationen immer in konsistenter Weise eindeutig interpretieren kann. Es bleibt noch einiger Raum für entsprechende Debatten. Mir geht es aber vordringlich um das einfache kausale Schließen und nicht darum, welche Kausaltheorie unser Verständnis von Kausalität am besten wiedergibt. Für das kausale Schließen war vor allem die Homogenitätsforderung wesentlich. Wie kann man die sicherstellen?

7.2.7 Randomisierung als Allheilmittel?

Verhilft uns die Randomisierung immer zu homogenen Testsituationen? Die wird in der modernen Wissenschaft oft als Allheilmittel gepriesen, das all unsere Probleme sicher löst. Nach den klassischen Regeln

des Experimentierens versuchen wir zwei homogene Gruppen (Experimentalgruppe oder Versuchsgruppe E und Kontrollgruppe K) so herbeizuführen, dass alle relevanten Faktoren bzw. alle Störfaktoren (auch und vor allem die unbekannten) zumindest gleich auf die beiden Gruppen verteilt sind und daher ein Vergleich der beiden Gruppen kausal aufschlussreiche Ergebnisse liefert. Dann wird etwa der Faktor A nur in der Gruppe E realisiert und nicht in K. Damit hätten wir zumindest auf der Ebene ganzer Gruppen eine Form der Homogenisierung erzielt, da sie in Bezug auf die Häufigkeit der Faktoren in beiden Gruppen ähnlich sind.

Man beachte aber auch, dass damit die Forderung der Homogenisierung eine neue Gestalt erhalten hat. Wir nennen dann zwei Gruppen von Objekten oder Systemen homogen für W, wenn sich alle möglichen Störfaktoren für die Beziehung (A,W) in gleicher Anzahl auf die beiden Gruppen verteilen. Damit muss noch nicht einmal sichergestellt sein, dass es zwei Objekte in den beiden Gruppen gibt, die genau dieselben Störfaktoren aufweisen. Trotzdem erhoffen wir uns von einer derartigen Homogenisierung natürlich schon Hinweise auf die effektiven Unterschiedsmacher. Ist A kausal relevant für W nehmen wir an, dass sich in den meisten Fällen ein Unterschied in den relativen Häufigkeiten zeigen wird, mit denen W in den beiden Gruppen auftritt. Aber es könnten sich in E sogar bestimmte positive Wirkungen von A unter bestimmten Bedingungen mit negativen Wirkungen von A unter anderen Bedingungen gerade aufheben. Für den Normalfall nehmen wir aber an, dass das nicht passiert und sich daher die Wirksamkeit von A in entsprechenden Unterschieden der Gruppen manifestieren wird.

Doch selbst diese Form der Homogenisierung muss keinesfalls immer gelingen, denn die Zufallsauswahl zweier Stichproben ist nur ein statistisches Verfahren, das nur mit einer bestimmten Wahrscheinlichkeit die gewünschten Ergebnisse liefert. Als Beispiel möchten wir wieder einmal herausfinden, ob A eine Ursache von W ist. Nehmen wir dazu an, wir hätten eine große Gesamtpopulation G in einer deterministischen Welt, in der ein Faktor A tatsächlich eine Ursache von W darstellt. Allerdings ist dazu ein uns nicht bekannter Kofaktor N notwendig, der nur in 5% aller Elemente aus G vorliegt. Außerdem gebe es einen Störfaktor F, der ebenfalls und unabhängig von A zu W führt. F liege in 40% aller Fälle vor

und sei statistisch unabhängig von A und N. Wenn es keine anderen Ursachen von W gibt, dann finden wir folgende Situation in G: Für 43% der Elemente in G tritt W auf, wobei in 38% aller Fälle F allein die Wirkung W verursacht, in 2% der Fälle verursachen sowohl A wie auch F in einer Form der Überdetermination W und in 3% der Fälle wird W allein durch A (und seinen Kofaktor N) hervorgebracht.

Nun ziehen wir per Zufallsauswahl zwei Stichproben E und K zu je 100 Elementen aus G. In E sorgen wir dafür, dass A für alle Elemente vorliegt (etwa ein Medikament genommen wird) und in K dafür, dass non-A vorliegt. Nehmen wir außerdem an, dass die Stichprobenwahl für den Kofaktor N perfekt funktioniert hat und somit in genau 5% der Stichprobenelemente N vorliegt (also jeweils in 5 Fällen). Dann folgt, dass in 5% der Fälle von E A tatsächlich wirksam wird. Für E gehen wir also einfach von dem Idealfall aus, dass W in 43 Fällen auftritt. Sollte die Zufallsauswahl für K genauso exakt funktionieren, hätten wir dort nur 40-mal W zu erwarten und hätten immerhin noch einen Unterschied von 3 Fällen aufzuweisen, der einen Hinweis auf die Wirksamkeit von A geben würde.

Nun berücksichtigen wir aber noch, dass es sich nur um ein statistisches Verfahren handelt. Es ist nicht sichergestellt, dass F in genau 40 Fällen von K vorliegt. Was bedeutet das für unseren Unterschied? Dazu wollen wir die Frage beantworten: Wie oft müssen wir im Durchschnitt damit rechnen, dass auch in K 43-mal oder sogar häufiger W auftritt, wodurch die Wirksamkeit von A vollkommen unsichtbar würde? Der Störfaktor F ist in K binomialverteilt mit dem Parameter p=0,4. Davon dürfen wir ausgehen, da die Population G sehr groß sein sollte, ansonsten müssten wir mit der hypergeometrischen Verteilung arbeiten. Man beachte, dass unsere vorgestellte Welt hier im Wesentlichen deterministisch ist, bis auf die Zufallsauswahl der Elemente von E und K. Tatsächlich würden in ca. 24 von 100 Fällen mindestens 43 W-Elemente in K zu finden sein. Das *statistische Rauschen* kann also schnell bestimmte Unterschiede verwischen, zumal wenn diese klein sind. Dass sie klein sind, heißt aber nicht, dass A nur schwach wirksam wäre. In den Fällen, in denen der Kofaktor N vorliegt, wirkt A sogar einhundertprozentig.

Wir könnten noch zusätzlich berücksichtigen, dass auch für E eine Zufallsauswahl stattfindet, aber das würde uns z.T. begünstigen (der

Effekt von A schiene deutlicher) und z.T. gegen uns arbeiten (es würde zur Verwischung beitragen). Also würden wir auch dabei in ca. 24% der Fälle bei einer Randomisierung den Effekt von A nicht mehr erkennen können. Zusätzlich müssten wir eigentlich noch berücksichtigen, dass ein Unterschied schon eine bestimmte Größe annehmen muss, um überhaupt als signifikant gelten zu können, sonst gewinnt die Nullhypothese die Oberhand. Damit vergrößert sich der Bereich, in dem die Wirksamkeit von A nicht erkannt würde. Das mag als Beispiel ausreichen, um zu zeigen, dass die Randomisierung kein Allheilmittel für alle Fälle ist. Weitere Probleme werden später noch genannt. Trotzdem bleibt sie natürlich ein wichtiges Hilfsmittel dort, wo sie überhaupt einsetzbar ist.

Hilfreich sind für das Experimentieren auch die Kombinationen von aktiver Homogenisierung und Randomisierung. Kennen wir bereits bestimmte relevante Kausalfaktoren B für W, können wir Experiment- und Kontrollgruppe bewusst bzgl. dieser Faktoren homogenisieren und zusätzlich randomisieren, d.h., wir sorgen dafür, dass genauso viele Elemente mit B in E wie in K sind, in dem wir die B-Elemente heraussuchen und dann per Zufallsauswahl auf beide Gruppen verteilen.

Um die Anzahl der verzerrten Experiment- und Kontrollgruppen zu reduzieren, können wir in manchen Fällen die Stichprobengrößen erhöhen, aber es können immer noch weitere Störfaktoren auftreten, die in vielfältiger Weise interagieren. Also dürfen wir uns keineswegs blindlings auf die Randomisierung verlassen. Außerdem können wir sie in der Praxis oft nicht einsetzen, sondern sind schlicht auf die vorliegenden beobachtbaren Daten angewiesen, um auf Kausalität zu schließen. Das Beste, was wir dann haben, ist die Abduktion auf die Hypothesen, die unsere Daten am besten erklären können.

7.2.8 Ist minimales Schließen abduktiv?

Die Fragestellungen im Rahmen des Schließens anhand der minimalen Theorien entsprechen denen des Schlusses auf die beste Erklärung. Wir finden wieder die typischen Merkmale des abduktiven Schließens, nämlich eine Liste von konkurrierenden Hypothesen und eine möglichst weitgehende Elimination der Konkurrenzhypothesen, bis nur noch eine übrig bleibt. Die Liste der Hypothesen wird durch die ins Auge gefassten

Faktoren und die demnach möglichen minimalen Theorien vorgegeben. Dann suchen wir nach speziellen Daten, die einen Unterschied zwischen solchen minimalen Theorien ausmachen, um bestimmte Theorien eliminieren zu können. Dazu vergleichen wir homogenisierte Situationen miteinander, die sich nur im Prüffaktor A unterscheiden. Wenn W dann nur auftritt, wenn ebenfalls A vorlag und ohne A abwesend bleibt, dann können die minimalen Theorien, in denen A keine Ursache von W ist, das nicht mehr erklären und werden im Sinne einer eliminativen Abduktion ausgeschieden.

Diese speziellen abduktiven Schlüsse sind allerdings nur anhand spezieller kausaler Hintergrundannahmen möglich, das ist ganz analog zu den Schlüssen auf die beste Erklärung. Die Annahmen bestehen u.a. in der Vermutung, dass eine Homogenisierung vorliegt und dass empirische Vollständigkeit gegeben ist. Außerdem ist die kausale Theorie selbst als Vermutung zu nennen, nach der eine Ursache genau ein Faktor einer minimalen Theorie ist. Dafür gibt es zwar gute Gründe, aber letztlich handelt es sich auch hierbei um eine empirische Vermutung und wohl nicht nur um eine analytische Behauptung. Jedenfalls hilft sie uns, in bestimmten Situationen abduktiv zu erschließen, dass eine Ursache-Wirkungs-Beziehung vorliegt.

7.3 Probabilistische Kausalität und bayessche Netze

7.3.1 Probabilistische Ursachen und Reichenbachs Schlussregel

Einen anderen Zugang zur Kausalität, der zumindest nicht zwingend auf den Determinismus angewiesen ist, sondern mit der Vorstellung verträglich ist, dass Ursachen einen probabilistischen Charakter haben können, bieten die probabilistischen Kausalkonzeptionen. Sie gehen vor allem davon aus, dass das Auftreten einer Ursache A die Wahrscheinlichkeit dafür erhöht, dass auch die Wirkung W auftritt. Wir bleiben dafür zunächst bei dichotomen Zufallsvariablen bzw. qualitativen kausalen Faktoren A, B, … etc, die die Merkmale einer Situation beschreiben, die als Ursachen oder Wirkungen in Frage kommen. Wir bleiben außerdem auf der Ebene der generellen Kausalität:

A ist (prima facie) Ursache von W gdw.
(1) P(W|A) > P(W) oder äquivalent dazu
(2) P(W|A) > P(W|¬A)

Demnach muss eine Ursache vor allem die Wahrscheinlichkeit dafür erhöhen, dass die Wirkung auftritt, was wir hier durch die bedingten Wahrscheinlichkeiten ausdrücken. Man sagt auch, dass A und W nun *korreliert* sind. Das ist allerdings zunächst eine symmetrische Beziehung zwischen A und W, aus der sich nicht erschließen lässt, was Ursache und was die Wirkung ist.

Normalerweise werden wir dabei an objektive physikalische Wahrscheinlichkeiten denken und natürlich nicht an die subjektiven Wahrscheinlichkeiten des Bayesianers. Es geht uns darum, dass das Auftreten von A die objektive Wahrscheinlichkeit von W erhöht und nicht nur darum, dass A einen Grund dafür bietet, anzunehmen, dass nun auch W vorliegen wird. Insbesondere erscheinen dafür Propensitäten geeignet, die allerdings selbst schon recht nah an bestimmten Kausalkonzepten liegen, so dass jedenfalls die Gefahr besteht, mit (1) und (2) keine reduktive Erläuterung von Kausalität zu erhalten (vgl. Kap. 7.3.4). Wir sollten daher zumindest immer im Blick behalten, dass es eine enge Verbindung auch der Propensitätenkonzeption zu relativen Häufigkeiten gibt. Das beschreibt dann einen Weg, auf dem Kausalität mit beobachtbaren Größen zusammenhängt.

Eine interessante Variante zu (1) und (2) bietet Humphreys (1989) an, wonach wir die Erhöhung der Wahrscheinlichkeit gleich auf einen *neutralen Zustand* N beziehen sollten. Ob und welchen Beitrag ein Medikament tatsächlich zur Heilung liefert, wird dann besser dadurch beschrieben, dass wir fordern, das Medikament müsse besser abschneiden als ein Placebo, und die Gabe eines Placebos könnte z.B. als solch ein neutraler Zustand gelten.

(3) P(W|A) > P(W|N)

Es ist nicht immer klar, ob ein solcher Zustand existiert und ob er von (¬A) verschieden ist, aber in vielen Fällen existieren naheliegende neutrale Zustände und bieten eine Alternative zur klassischen Formulierung an. Doch ich werde dieser Komplikation hier nicht weiter nachgehen,

sondern mich einfach auf ¬A oder den allgemeinen Zustand ohne weitere Angaben über A beziehen, wie das in (2) und (1) oben der Fall ist.

Jedenfalls finden wir die Idee, dass wir von solchen Korrelationen aus auf Kausalzusammenhänge schließen dürfen, oft in der Praxis wieder und schließen häufig sogar vorschnell aus Korrelationen auf kausale Beziehungen. Vielleicht können wir z.B. beobachten, dass die Personen, die viel Zucker zu sich nehmen (Z) auch häufiger eine bestimmte Krebsart entwickeln (K) als andere. Es gilt also:

(4) P(K|Z) > P(K)

Aber deutet das bereits auf einen kausalen Zusammenhang hin, nach dem Zucker Krebs verursacht? Von einer derartigen Schlussfolgerung sind wir natürlich noch weit entfernt. Zunächst einmal ist eine Korrelation wie schon gesagt *symmetrisch*. Die Bedingung (4) ist daher äquivalent zu:

(5) P(Z|K) > P(Z)

Es könnte also auch umgekehrt sein, dass der beginnende Krebs ein Verlangen nach Zucker auslöst. Die Kausalbeziehung ist hingegen *asymmetrisch* und lässt sich schon deshalb aus probabilistischen Korrelationen nicht einfach herauslesen. Aber es gibt auch noch ganz andere Möglichkeiten. Es kann überdies zu Korrelationen kommen, ohne dass sie von einer Kausalbeziehung erzeugt werden, die die Faktoren miteinander verbindet. So war über viele Jahre der Brotpreis in England mit dem Wasserspiegel in Venedig korreliert (vgl. Sober 2001, Arntzenius 2010), einfach weil beide kontinuierlich ansteigen, aber wir vermuten dahinter natürlich keinen kausalen Zusammenhang zwischen diesen beiden Größen.

Einen weiteren einfachen Zusammenhang finden wir in *gemeinsamen Ursachen* (»*common causes*«). So treten etwa gelbe Finger und Lungenkrebs korreliert auf, ohne dass die gelben Finger krebsverursachend wären. Man nennt das manchmal etwas verwirrend *Scheinkorrelation*, doch die Korrelation ist echt, nur lässt sie sich nicht als Kausalbeziehung deuten. Wir sollten lieber von *Scheinkausalität* sprechen, denn die Kausalbeziehung liegt nur scheinbar vor. Keiner der beiden Faktoren ist

Ursache des anderen. Vielmehr steht im Hintergrund eine gemeinsame Ursache von beiden Phänomenen, die zu dieser Korrelation geführt hat, nämlich das vorhergehende Rauchverhalten der betreffenden Personen. Diese Einsicht hat zu Reichenbachs methodologischer Regel (seinem *Common-Cause-Prinzip*) geführt, die man ungefähr wie folgt formulieren kann:

(RR) Reichenbachs Regel: Wenn zwei Faktoren A und B miteinander korreliert sind, so liegt einer der drei folgenden Fälle vor:
(1) A verursacht B oder
(2) B verursacht A oder
(3) Es gibt eine gemeinsame Ursache (einen »Common Cause«) C für beide Faktoren.

Das ist sicher eine naheliegende anleitende Regel für die wissenschaftliche Forschung, die sich wieder auf eine einfache Regularität stützt. Im Normalfall werden wir bei auftretenden Korrelationen davon ausgehen, dass es dafür eine Ursache und damit auch eine Erklärung gibt. Das könnte zugleich als typische methodologische Regel abduktiven Schließens betrachtet werden: Sobald eine Korrelation auftritt, nehmen wir an, dass es dafür eine (kausale) Erklärung gibt. Dann beschreiben die drei Fälle unsere typischen Optionen. Wir könnten natürlich als weitere Hypothese noch eine Art von Nullhypothese hinzunehmen, die schlicht besagt, dass die Korrelation ein reines Zufallsprodukt ist.

Oft können wir anhand weiteren Hintergrundwissens schon entscheiden, welche der Fälle plausibel sind und welche nicht. Allerdings stellt die Regel selbst kein einfaches Schlussverfahren zur Verfügung, mit dem wir definitiv entscheiden könnten, welcher der drei Fälle nun vorliegt. Insbesondere wissen wir oft nicht, welche mögliche gemeinsame Ursache hier verantwortlich für die Korrelation sein soll. Die stellt eventuell eine unbekannte und unerwünschte Störvariable für unsere Untersuchung des Zusammenhangs von A und B dar.

Auf vergleichbare Weise schließen wir dann in konkreten *Einzelfällen*, in denen uns besondere Ähnlichkeiten auffallen. Erhalten wir von zwei Studenten a und b genau wortgleiche Klausuren, so werden wir etwa

annehmen, dass entweder a von b abgeschrieben hat oder umgekehrt oder dass es eine gemeinsame Quelle c für beide Texte gibt. Wir nehmen wiederum an, dass es eine Erklärung für diese Übereinstimmung gibt. Alles andere wie ein reiner Zufall würde uns (zumindest in diesem Fall) doch sehr überraschen.

Ein Problem der oben genannten Regel ist, dass sie zunächst die Möglichkeit vernachlässigt, dass die Gemeinsamkeiten bzw. die Korrelation durch *Zufall* entstanden sein kann. Selbst im Falle der Klausuren ist das nicht logisch ausgeschlossen, es erscheint uns hier nur sehr unwahrscheinlich. Wir müssen also die Zufallshypothese möglichst ausschließen, indem wir etwa zeigen, dass das Resultat sehr unwahrscheinlich wäre, wenn kein kausaler Zusammenhang vorgelegen hätte. Wir möchten die Zufallshypothese durch die entstehende Erklärungsanomalie ganz im Sinne der eliminativen Induktion bzw. des Schlusses auf die beste Erklärung eliminieren. Dabei ist aber wieder der Fehlschluss der probabilistischen Falsifikation zu vermeiden, den wir in Kapitel 6 kennengelernt haben.

Typischerweise wird jedenfalls die Zufallshypothese unsere Nullhypothese in einem Signifikanztest sein. Dann werden wir erst wieder der reichenbachschen Regel folgen, wenn wir diese Nullhypothese signifikant widerlegt haben, d.h., wenn sich die Korrelation schon als signifikant erwiesen hat. Die Nullhypothese würde also etwa besagen, dass in der Grundgesamtheit keine Korrelation zwischen A und B vorliegt, sondern nur durch Zufall in unserer Stichprobe eine Korrelation zu beobachten ist.

Ist die vorliegende Korrelation jedoch recht stark und damit signifikant, können wir diese »Erklärung« für unsere beobachtete Korrelation als nicht stichhaltig im Sinne eines probabilistischen Schlusses zurückweisen. Dabei sind allerdings wiederum die weiteren schon genannten Einschränkungen für derartige Hypothesentests zu berücksichtigen. Im Falle der beiden Schüler werden wir die Zufallshypothese umso eher zurückweisen, umso identischer und länger die beiden Texte sind. Bereits das Alltagswissen sagt uns, dass es sehr unwahrscheinlich ist, dass zwei Menschen einen längeren in großen Teilen wortgleichen Text (womöglich noch mit denselben Rechtschreibfehlern) unabhängig voneinander ver-

fassen, auch wenn das natürlich nicht definitiv ausgeschlossen werden kann.

Allerdings zeigt schon das obige Venedigbeispiel, dass dieses Verfahren nicht immer zu korrekten Schlüssen führen wird. Auch in der Quantenmechanik finden wir systematische Verstöße der Natur gegen die reichenbachsche Regel. In EPR-artigen Situationen (mit verschränkten Quantenobjekten) stoßen wir auf strikte Korrelationen, die aber keinen Common Cause aufweisen. Das sagt jedenfalls die Quantenmechanik, und das wird gestützt durch entsprechende Experimente, in denen die Bellsche Ungleichung getestet wurde, die dafür sprechen, dass es keine verborgenen Parameter gibt, die für die entsprechenden quantenmechanischen Vorgänge die Rolle einer gemeinsamen Ursache übernehmen könnten. Das gilt jedenfalls unter der Annahme, dass keine echten Fernwirkungen existieren, sondern Ursachen ihre Wirkungen nur mit endlichen Geschwindigkeiten ganz im Sinne der Relativitätstheorie verbreiten.

Man kann so letztlich zeigen, dass für genuin indeterministische Systeme die reichenbachsche Regel nicht mehr gilt. Es gibt noch weitere (z.T. eher technische) Probleme mit der reichenbachschen Regel, die u.a. Hitchcock (2010, Kap. 2.3) auflistet. Trotzdem scheint sie der beste Weg zur Ermittlung und Auszeichnung von Kausalbeziehungen für viele makroskopische Bereiche der Wissenschaften zu sein. Sie dient uns daher normalerweise als gute Richtschnur für das kausale Schließen in empirischen Untersuchungen und wird uns im Rahmen der Bayesschen Netze in Form der kausalen Markovannahme wieder begegnen.

Man könnte anhand der reichenbachschen Regel schließen, A sei Ursache von B, wenn eine Korrelation zwischen A und B vorliegt, für die es keinen Common Cause gibt und wir anhand unseres Hintergrundwissens ausschließen können, dass B A verursacht. Dazu müssen wir noch genauer bestimmen, was wir unter einem *Common Cause* C verstehen wollen bzw. wie wir feststellen können, dass etwas einen Common Cause darstellt:

(CC) Common Cause:
C ist ein *Common Cause* für zwei Faktoren A und B gdw:
(1) P(B|A) > P(B) und
(2) P(B|A&C) = P(B|C) und
(3) P(B|A&¬C) = P(B|¬C)

Das heißt, wenn wir schon wissen, dass der Common Cause C vorliegt, gibt uns A *keine zusätzliche Information* mehr über das Auftreten von B (und umgekehrt). (Man sagt auch, C schirmt A und B voneinander ab.) Die Bedingung (3) wird manchmal weggelassen, aber erst sie besagt, dass die Zufallsvariablen A und B unabhängig sind für alle Werte der Zufallsvariablen C.

Sonst könnten wir zu seltsamen Schlussfolgerungen gelangen, wonach eine gewöhnliche Ursache in den Verdacht gerät, ein Common Cause zu sein. Nehmen wir z.B. an, A und C verursachen jeder für sich B. Klassisches Beispiel sind die zwei Todesschützen, die beide erfolgreich und damit tödlich auf ihr Opfer schießen (Schüsse: A und C), dass dabei zu Tode kommt (B). Nehmen wir dabei an, wir hätten einen klassischen Fall von echter *Überdetermination* vor uns, d.h., beide Ursachen reichen für sich genommen aus, um B herbeizuführen.

Dann scheint es zunächst klar zu sein, dass wir intuitiv sowohl A wie auch C als Ursachen von B betrachten müssen; obwohl selbst das bisweilen angezweifelt wird. So behaupten Vertreter des kontrafaktischen Ansatzes manchmal, die Fälle von Überdetermination wären unklar, weil ihr Ansatz damit Probleme hat. Genau genommen müssen wir zunächst auf die Ebene der konkreten Instanzen unserer Faktoren gehen (also zur *singulären Kausalität*), denn nur dort finden wir die Überdetermination. Ein Argument für ein Vorliegen einer echten kausalen Überdetermination sieht dann ungefähr so aus: Wir können wohl zunächst sagen, dass unser B-Ereignis nicht unverursacht stattfand. Unser B-Ereignis ist nicht eines ohne kausale Vorgeschichte (wie das etwa für sogenannte Vakuumfluktuationen der Fall sein könnte), sondern wird ganz offensichtlich durch die vorausgehenden Ereignisse beeinflusst. Weiterhin stehen das entsprechende A-Ereignis und das entsprechende C-Ereignis in einer symmetrischen Situation zueinander, weshalb wir entweder

beide oder keines als Ursache betrachten sollten. Kommen wir nun nicht auf die Idee, seltsame disjunktive Ereignisse wie ein A∨C-Ereignis einzuführen, bleibt uns eigentlich keine andere Wahl, als beide als gewöhnliche Ursachen von B zu betrachten, sonst bliebe B schließlich doch unverursacht.

In so einem Fall können die Anforderungen (1) und (2) von (CC) erfüllt sein. Das Auftreten von A-Ereignissen wird im Normalfall das Auftreten von B-Ereignissen wahrscheinlicher machen. Aber, wenn schon C-Ereignisse vorliegen, trägt A nichts Neues mehr zu unserem Wissen über das dann sichere Auftreten von B-Ereignissen bei. Also würde C fälschlicherweise als ein Common Cause im Sinne der obigen Definition (CC) eingestuft, wenn wir nicht noch Bedingung (3) hätten (vgl. auch Baumgartner & Graßhoff 2004, Kap. 4). Tatsächlich trägt unsere Information, dass ein A-Ereignis stattfand, dazu bei, dass wir nun auf ein B-Ereignis schließen dürfen, wenn wir schon wissen, dass kein C-Ereignis stattfand. Es ist also gerade Bedingung (3), die dafür sorgt, dass C hier nicht fälschlicherweise für einen Common-Cause gehalten wird, was dazu führen würde, dass dann die Kausalbeziehung von A nach B nicht mehr erkannt würde.

Leider ist die Regel (CC) nicht immer so erfolgreich. Typischerweise werden durch sie z.B. *Zwischenursachen* fälschlicherweise als gemeinsame Ursachen eingestuft. Nehmen wir an, eine Meldung über die angebliche Zahlungsunfähigkeit des Bankhauses Superschlau (A) führe zu einer Angst der Anleger, dass diese Bank in Insolvenz gehen könnte (C) und das verursache wiederum, dass die Kunden massiv Gelder bei Superschlau abheben (B). Dann ist A eine Ursache von C und C eine von B, aber C erfüllt offensichtlich die Gleichungen (2) und (3). Das Wissen um die negative Meldung über das Bankhaus Superschlau hilft uns nicht, A zu prognostizieren, wenn wir C schon kennen oder wissen, dass C nicht vorliegt, also etwa schon wissen, dass die Meldung keine negativen Auswirkungen auf die Gemüter der Kunden gehabt hat. (Auch hier schirmt C A und B voneinander ab.)

Das Entsprechende gilt auch in der anderen Richtung. Trotzdem ist C keine gemeinsame Ursache von A und B, sondern nur eine Zwischenursache zwischen A und B. Hier stoßen wir auf das Problem, dass es zwei unterschiedliche Kausalstrukturen gibt: (a) A⇒C⇒B und

(b) A⇐C⇒B, die probabilistisch gesehen korrelationsäquivalent sind, d.h., die zumindest dasselbe Muster von Korrelationen bzw. statistischen Unabhängigkeiten zwischen den Zufallsvariablen A, B und C erzeugen. Allerdings kann es u.U. gelingen, die Äquivalenz wieder aufzuheben, indem man weitere Zufallsvariablen ins Spiel bringt. Gelingt das nicht, müssen wir weiteres Hintergrundwissen etwa über das zeitliche Auftreten der Instanzen unserer Faktoren ins Spiel bringen, um zu klären, ob Fall (a) oder (b) vorliegt.

Man könnte die Definition der gemeinsamen Ursache also noch dadurch ergänzen, dass C *keine Zwischenursache* zwischen A und B sein darf. Das können wir etwa dadurch erreichen, dass die Instanzen von C jeweils *zeitlich vor* denen von A und B liegen müssen. Viele Vertreter von Kausaltheorien zieren sich allerdings, zeitliche Beziehungen in die Definition der Kausalbeziehung aufzunehmen, weil sie die Kausalbeziehung für grundlegender als die zeitlichen Beziehungen halten und eine entsprechende Explikationsreihenfolge einhalten möchten.

Für das kausale Schließen sollten wir uns aber keinesfalls davon abhalten lassen, solche zeitlichen Beziehungen zu nutzen, wenn sie denn für uns erkennbar sind. Das muss nicht immer leicht sein. So gilt zwar die Bewegungsarmut oft als Ursache für unser Übergewicht, aber es ist nicht so klar, was zuerst da war. Vielleicht führt auch unser Übergewicht dazu, dass wir uns weniger bewegen, weil es natürlich die Bewegung erheblich erschwert. Oder es findet eine gegenseitige Beeinflussung im Sinne eines Teufelskreises statt. In der Praxis des kausalen Schließens helfen also auch zeitliche Beziehungen nicht immer weiter.

7.3.2 Simpsons Paradox und probabilistische Verursachung

Ein weiteres Problem für das probabilistische kausale Schließen mit praktischen Auswirkungen finden wir im sogenannten *Simpson Paradox*. Hierbei überlagern sich zwei (oder mehr) Ursachen so, dass sie zu überraschenden Ergebnissen führen. Nehmen wir dazu ein einfaches fiktives Beispiel: Bei der Beobachtung von 100 regelmäßigen Alkoholtrinkern (A) (etwa als Zufallsstichprobe unter allen Alkoholtrinkern auszuwählen) stellen wir fest, dass nur 10 von ihnen im Laufe von drei Jahren einen Krebs (K stehe für einen bestimmten Krebs) entwickeln, während bei der

Beobachtung von 100 Abstinenzlern 20 in demselben Zeitraum einen Krebs entwickeln.

Also scheint das Alkoholtrinken überraschenderweise eine starke Schutzfunktion gegen Krebs aufzuweisen, halbiert es doch die Krebsrate. Jedenfalls gilt: $P(K|A) < P(K|\neg A)$. Aber so schön ist das Leben meistens nicht. Dahinter verbergen sich vielleicht ganz andere Zusammenhänge. Nehmen wir z.B. an, unter den 100 Alkoholtrinkern wären 80 Jugendliche, weil das Alkoholtrinken bei denen gerade besonders in Mode ist, während nur noch 20 ältere Menschen dabei sind, weil bei denen das Alkoholtrinken viel seltener vorkommt. Bei den Abstinenzlern sei das Verhältnis dann genau anders herum (80 Ältere und 20 Jüngere). Die Gelegenheitstrinker lassen wir beiseite. Dann könnte folgendes passiert sein: Die hohen Krebsraten bei den Abstinenzlern könnten eine Folge davon sein, dass unter ihnen viel mehr ältere Menschen zu finden sind und diese eben deutlich häufiger Krebs bekommen als jüngere Menschen. Die Zahlenverhältnisse könnten so wie in unserer Tabelle aussehen:

	Alkoholtrinker		Abstinenzler	
	jung	alt	jung	alt
	80	20	20	80
Krebsfälle	5	5	1	19
	(6,25%)	(25%)	(5%)	(24%)
zusammen	10		20	

Tabelle 7.6: Ein Beispiel für das Simpson-Paradox

Schauen wir uns nun noch einmal diese Tabelle genauer an. Tatsächlich ist in jeder der beiden Gruppen für sich genommen (sowohl bei den Jungen und wie bei den Alten) die Krebsrate der Alkoholtrinker leicht erhöht gegenüber der von den Abstinenzlern. Die Tabelle gibt also bei genauerer Betrachtung keinen Anhaltspunkt mehr dafür, dass Alkohol gegen Krebs (K) helfen könnte, denn die Wahrscheinlichkeiten (bzw. relativen Häufigkeiten) für Krebs sind in den beiden Untergruppen jeweils für die Alkoholtrinker ganz leicht erhöht. Im Gegenteil scheint er also eher (ein wenig) schädlich zu sein.

Das deutet wieder darauf hin, dass wir unsere einfache Gleichung für Ursachenverhältnisse $P(K|A) > P(K|\neg A)$ in unserem Fall ergänzen müssen. Wir müssen andere für die Krebsrate relevante Faktoren berücksichtigen,

vor allem, wenn sie eine Korrelation mit unserem Prüffaktor A (dem Alkoholtrinken/Nicht-trinken) aufweisen, was in unserem Fall vorliegt: Hier stellt das Alter (B) einen solchen Faktor dar, weil in unserem Beispiel de facto mehr Junge als Alte Alkohol trinken und sie mit ihren niedrigen Krebsraten die Krebsrate der Trinker deutlich verbessern. Die älteren Menschen trinken zwar weniger, aber entwickeln natürlich viel häufiger einen Krebs. Hier sind zwei kausale Faktoren miteinander korreliert und sorgen so für eine mögliche Konfusion. Berücksichtigen wir den Faktor Alter (B) erhalten wir dagegen in jeder der beiden neuen Untergruppen wieder eine positive Krebsrelevanz für den Alkohol. Als neue Anforderung für die Wirksamkeit von non-A für eine hohe Krebsrate K sollte daher stehen:

(6) $P(K|A\&B) > P(K|\neg A\&B)$ und $P(K|A\&\neg B) > P(K|\neg A\&\neg B)$,

d.h. bei Vorliegen und bei Abwesenheit des zweiten Faktors B sollte jeweils A seine Wirksamkeit in einer erhöhten Wahrscheinlichkeit für K zeigen. Erst dann dürfen wir ihn als Ursache von K einstufen. Faktor B könnte hierbei sogar ein Common Cause von A und K sein, was wiederum deutlich macht, dass es sich um einen *potentiellen Störfaktor* handelt.

Bedingung (6) stellt somit wieder eine Form der *Homogenisierung bzgl. K* dar (d.h., wir *kontrollieren* den Faktor B), die wir benötigen, um die Wirkung von A relativ zu non-A jeweils in einer ansonsten bzgl. K homogenen Umgebung zu ermitteln. Statistiker sprechen hier davon, dass A nicht unbedingt in der Gesamtpopulation eine niedrige Krebsrate nach sich ziehen muss, aber doch in den relevanten Subpopulationen, die hier durch B und non-B gegeben sind. Nur dann könnten wir A als eine Ursache von K betrachten. Da das in unserem Beispiel nicht gegeben ist, spricht das dagegen, dass A hier gegen K hilft.

Außerdem wäre es bereits ein Hinweis auf eine kausale Wirkung von A auf K, wenn nur eine der beiden Ungleichungen Bestand hätte, denn dann hätte A zumindest noch eine wahrscheinlichkeitserhöhende Wirkung in einer bestimmten Umgebung, wobei auch wieder zwei homogene Situationen miteinander verglichen würden. Dann wäre A zumindest wohl unter diesen Bedingungen eine Ursache von K. Wenn

etwa der Alkohol nur bei den Jüngeren zu Krebs führt (das Alter wäre dann also ein Kofaktor für den Alkohol in Bezug auf Krebs), dann würden wir natürlich trotzdem behaupten, dass der Alkohol kausal relevant für Krebs wäre.

Alkohol könnte nun bei den älteren Menschen vor Krebs schützen und dann könnte insgesamt sogar gelten: $P(K|A) = P(K|\neg A)$. Trotzdem wäre dann der Alkohol kausal relevant für Krebs, aber eben in unterschiedlicher Weise in verschiedenen Subpopulationen. Solche seltsamen Konstellationen hatten wir im Rahmen der minimalen Theorie schon beim Vierertest diskutiert und zugelassen. Sie dürften aber eher die Ausnahme sein.

Mit unserer Forderung (6) sind wir aber noch nicht am Ende der Fahnenstange, denn es könnte noch weitere Störfaktoren $F_1,\ldots,F_n$ geben, die unser Ergebnis wiederum irreführend gestalten. Die müssten wir dann ebenfalls berücksichtigen. Wir suchen schließlich nach der *größten* gerade noch *homogenen Subpopulation*, um die Wirkung von A auf K zu ermitteln. Erst wenn A seine Wirkung auf K jeweils in all diesen Subpopulationen (oder zumindest in einigen davon) entfaltet, dürfen wir A als Ursache von K annehmen. Hierbei gehen wir wiederum (wie im Falle der Propensitätendebatte) davon aus, dass sich die Menge aller für K kausal relevanten Faktoren F_i als endliche Menge erweist, die zumindest im Prinzip in unserer Ungleichung angebbar sind.

Wählen wir noch ein weiteres konkretes Beispiel, in dem wir schon wissen, dass tatsächlich eine Kausalbeziehung vorliegt. Rauchen (R) verursacht Lungenkrebs (L). Damit das nachgewiesen werden kann und wiederum nicht nur eine Simpson-Paradox-Situation vorliegt, müssen wir im Prinzip für alle Faktoren F_i und ihre Kombinationen (hier gegeben etwa durch die Vollkonjunktionen der Aussagen F_i), die für das Auftreten von L möglicherweise relevant sind, verlangen, dass eine entsprechende Ungleichung vorliegt. Das heißt, wir müssen wie im deterministischen Fall eine *weitergehende Homogenisierung* der Bedingung vornehmen, für die dann zwei Situationen miteinander verglichen werden, nämlich einmal mit R und einmal ohne R:

(7) **R ist Ursache von L gdw.**
$\forall F_i$ in allen möglichen $\pm$-Konstellationen gilt:
$P(L|R\&\pm F_1\&\ldots\&\pm F_n) > P(L|\neg R\&\pm F_1\&\ldots\&\pm F_n)$

In dieser Richtung geht auch Humphreys (1989). Damit würden wir das Auftreten von gemeinsamen Ursachen bereits mitberücksichtigen, denn die wären schließlich ebenfalls bestimmte Faktoren F_i für die (7) nicht mehr gelten würde. Wir würden somit Fehlschlüsse aufgrund von Scheinkausalität vermeiden.

Leider müssen wir (7) sogleich wieder einschränken. *Zwischenfaktoren*, die in der Kausalkette zwischen R und L liegen, dürfen natürlich nicht homogenisiert werden, weil sonst die Wirkung von R an dieser Stelle wieder gestoppt würde. Geht es z.B. darum, ob Aspirin eine schmerzstillende Wirkung besitzt, dürfen wir nicht über den Faktor der ASS-Konzentration im Blut homogenisieren. Halten wir die nämlich konstant, wird sich kein Unterschied zeigen zwischen der Situation, in der wir Aspirin nehmen, und der, in der wir das nicht tun. Die Bedingung (7) verlangt aber, dass wir tatsächlich *alle* anderen für L kausal relevanten Bedingungen F_i im Blick haben, weshalb diese Vorgehensweisen oft explizit auf eine vorgegebene Menge Φ von Faktoren relativiert werden.

Wir können dann wieder allgemeiner formulieren: Welche Bedingungen müssen erfüllt sein, damit wir sagen können, dass A eine (probabilistische) Ursache von W ist? Nehmen wir zunächst an, wir hätten schon eine Menge Φ von Faktoren, die vermutlich alle für L kausal relevanten Faktoren zumindest mit enthält (außer den Zwischenfaktoren zwischen A und W), und wir möchten nun nur noch untersuchen, welche davon tatsächlich wirksam sind. Von dieser Grundannahme geht praktisch jeder Ansatz zum kausalen Schließen mehr oder weniger aus, denn wenn gänzlich unbekannte Faktoren im Spiel sind, wird es sehr schwierig, korrekte kausale Schlüsse zu ziehen. Haben wir nicht einmal eine Vorstellung von dieser Menge Φ der in Frage kommenden Faktoren, bleibt als unser einziger Trick dann nur noch die Randomisierung, die aber leider nicht zuverlässig das leistet, was wir wünschen (s. Kap. 7.3.13). Es sei nun also $\Phi = \{A,F_1,\ldots,F_n\}$, dann können wir etwa festlegen:

(KR) Kausale Relevanz: Ein Faktor A ist *kausal relevant* für einen anderen Faktor W, wenn Folgendes gilt:

(i) $\Phi = \{A,F_1,\ldots,F_n\}$ ist eine Menge von Faktoren, die (unter anderem) alle für W direkt kausal relevanten Faktoren enthält (außer den Zwischenfaktoren zwischen A und W).

(ii) Es gibt zumindest *eine* Konstellation $S \equiv \pm F_1 \& \ldots \& \pm F_n$ von Faktoren, für die gilt: $P(W|A\&S) > P(W|S)$.

Man könnte sagen, dass diese Definition von kausaler Relevanz diese auf die Menge Φ *relativiert*, aber wir möchten uns schließlich wieder möglichst von solchen Relativierungen befreien. Deshalb habe ich gleich eine recht große Menge Φ gewählt, für die nur noch verlangt wird, dass alle potentiell relevanten Faktoren tatsächlich dort enthalten sind. In vielen Fällen verfügen wir zumindest über eine begründete Vermutung, welche Faktoren überhaupt relevant für eine Wirkung W sein *könnten* und können die dann einbringen. Dadurch, dass wir Φ sehr groß wählen, dürfte die Relativierung auf Φ nicht mehr ganz so gravierend ausfallen, da wir nicht genau wissen müssen, welche Faktoren tatsächlich kausal wirksam sind (das wollen wir schließlich erst herausfinden), sondern nur noch, welche überhaupt als mögliche Ursachen in Frage kommen, und dabei können wir großzügig vorgehen.

Ein weiterer wichtiger Schritt ist hier die *Abschwächung* der früheren Anforderungen an Kausalbeziehungen in (7). Wir verlangen in dieser Definition nicht mehr, dass A unter allen Umständen, bzw. in allen Subpopulationen kausal relevant ist, sondern nur noch, *dass es zumindest eine solche Subpopulation gibt, in der A seine Wirksamkeit zeigt*. Damit können wir zum einen der Tatsache Rechnung tragen, dass einige der Faktoren Kofaktoren oder Zwischenursachen sein könnten. Zum anderen könnte es der Fall sein, dass ein Faktor zwar unter bestimmten Umständen zur Wirkung kommt, aber nicht unter allen.

Unter einigen Umständen könnte er sogar kontraproduktiv sein, wie wir das schon im Alkohol-Krebs-Beispiel diskutiert hatten. Dann würden wir ihn trotzdem weiterhin als *kausal relevant* ansehen. Die Forderung von Humphreys (1989) nach einer Wirksamkeit unter allen Bedingungen stellt unnötig starke und unrealistische Anforderung an eine kausale

Relevanzbeziehung. Die schwächere Forderung wird manchmal so beschrieben, dass kausale Relevanz *kontextuell* sein kann, d.h., je nach spezieller Situation zum Tragen kommt oder in anderen Situationen eben nicht zum Tragen kommt. Wir dürfen jedenfalls nicht bereits in einer Analyse von Kausalität voraussetzen, dass kausale Faktoren in jedem Kontext ihre Wirksamkeit entfalten, wenn sie überhaupt jemals wirksam sind.

Haben wir z.B. ein Medikament M, das für bestimmte Personengruppen P Grippe heilen kann, aber für andere nicht oder die Krankheit im Extremfall für andere Personengruppen P* sogar verschlimmert, so werden wir trotzdem sagen, dass M kausal relevant für die Grippe ist. Auch an einem solchen Medikament wären wir natürlich sehr interessiert. Allerdings kann es leicht zu epistemischen Problemen kommen, denn wenn wir normalerweise in unseren Populationen immer einen Mix von Personen aus den Gruppen P und P* vorfinden, könnte die positive Wirkung von M dadurch völlig überdeckt werden. Die mit M behandelten Personen genesen dann vielleicht im Durchschnitt genauso langsam oder sogar noch langsamer als die Personen, die nur einen Placebo erhalten haben. Sollten wir also die unterscheidenden Merkmale für Personen vom Typ P und solche vom P* nicht kennen und daher keine entsprechenden Experimente nur für diese Untergruppen durchführen können, droht unser Medikament schnell wieder verworfen zu werden (es droht ein Simpson-Paradox). Hier hilft natürlich auch keine Randomisierung.

Andererseits verhindert das aber nicht, dass wir M als wirksames Medikament betrachten würden, wären uns die tatsächlichen Fakten bekannt. Mehr verlangen wir von kausal relevanten Faktoren nicht. Sie müssen zumindest in bestimmten Situationen eine bestimmte Wirkung haben. Also sind auch *kontextuell kausal relevante* Faktoren kausal relevante Faktoren.

Die Menge Φ sollte so gewählt sein, dass zunächst zumindest alle kausal relevanten Faktoren für die Wirkung W mit enthalten sind. Als Approximation würde es aber schon genügen, wenn die gewichtigeren Faktoren enthalten sind. Typischerweise nehmen wir in der Wissenschaft erfolgreich an, dass für einen bestimmten Effekt nur wenige Faktoren einen gewichtigen Einfluss haben. Das ist praktisch eine transzendentale

Annahme für unsere Forschungstätigkeit (s. Kap. 5.4.7). Doch das Problem der approximativ vollständig gewählten Menge Φ möchte ich hier nicht weiter verfolgen. Denken wir uns Φ lieber gleich als besonders groß gewählt und damit vollständig. Wir können die Menge dann gegebenenfalls wieder verkleinern, wenn sich bestimmte Faktoren als irrelevant für unsere Wirkung gegeben andere Faktoren erwiesen haben.

Nehmen wir etwa an, die gesamte kausale Struktur sei durch eine Reihe von Kausalketten organisiert, die sich zwar kreuzen dürfen, aber keine Zirkel enthalten. Dann benötigen wir in Φ genau genommen nur die direkten Ursachen von W (seine kausalen Eltern) und können die indirekten (die kausalen Vorfahren der Eltern von W) außen vor lassen, denn wir denken uns Kausalität im Regelfall so, dass vorhergehende Ursachen nur über die Zwischenursachen auf spätere Ereignisse wirken, so dass nicht noch zusätzlich die vorhergehenden Ursachen das Ergebnis beeinflussen. Das ist die Idee der *kausalen Markov-Bedingung*, die wir später noch kennenlernen werden. Nur die direkt wirkenden Faktoren sind zu berücksichtigen. Das werden wir noch weiterverfolgen, um wieder etwas mehr Übersicht in unsere Überlegungen zu bringen. Wir müssen dann in (KR) nur für die direkten kausalen Ursachen von W die entsprechenden Homogenisierungen betrachten.

Es läuft also darauf hinaus, dass wir in (KR) möglichst große, aber bzgl. W noch homogene Subpopulationen vom Typ $S_j \equiv \pm F_1 \& \ldots \& \pm F_n$ betrachten, wobei wir nur die kausalen Eltern von W berücksichtigen müssen, denn eine weitere Unterteilung dieser Populationen würde keine Veränderung der Wahrscheinlichkeit von W bewirken. So würden wir als Elementarzellen unserer Zerlegung Φ die größten relativ zu W objektiv homogenen Zellen aus der Gesamtpopulation auswählen. Das setzt natürlich ein weiteres Hintergrundwissen voraus, über das wir manchmal nicht verfügen und deshalb würde es schon genügen, wenn Φ zumindest alle für W relevanten Faktoren enthält und eventuell noch einige mehr, die genau genommen überflüssig sind.

7.3.3 Die Stärke der probabilistischen Verursachung

Für solche probabilistischen Konzeptionen wird verständlicherweise häufig nach einem probabilistischen Maß für die Stärke der Ursachen

in Bezug auf die Wirkung gefragt. Leider ist die Antwort nicht ganz so einfach, und es wurden etliche unterschiedliche Maße vorgeschlagen, von denen einige etwa bei Fitelson (2011) aufgeführt und verglichen werden. In unserem Fall, in dem A kausal relevant für W ist und in dem es eine Situation $S \equiv \pm F_1 \& \ldots \& \pm F_n$ geprägt durch die Faktoren F_i gibt, für die wir dann $P(W|A\&S) > P(W|S)$ finden, liegt es nahe, zunächst die spezielle *Wahrscheinlichkeitserhöhung* für W durch A vorzuschlagen:

Stärke der Verursachung: $\text{Stärke}(A,W;S) = P(W|A\&S) - P(W|S)$

Dieser Wert ist natürlich relativiert auf die spezielle Situation S. In S hat A eine bestimmte wahrscheinlichkeitserhöhende Wirkung für das Auftreten von W. Man kann das Ganze natürlich ebenso für eine Population Q angeben, in der unterschiedliche Situationen S_j mit unterschiedlichen relativen Häufigkeiten auftreten und erhalten dann den entsprechenden gewichteten Durchschnittswert der unterschiedlichen Wirkungen von A:

$$\text{Stärke}(A,W;Q) = \sum_j (P(W|A\&S) - P(W|S))\ P(S_j)$$

Leider haben die so gewonnenen Maße einen kleinen Schönheitsfehler. Sie geben nicht unbedingt die Stärke an, mit der A W hervorruft, denn es hängt davon ab, in welchem Ausmaß W bereits durch S hervorgerufen wird. Je eher S dazu neigt, W hervorzurufen, umso kleiner wird die Stärke(A,W;S) ausfallen. Das könnte man einen *Basiseffekt* nennen, der verhindert, dass wir den unverfälschten Einfluss von A auf W bestimmen. Patricia Cheng (1997) hat das Maß etwa noch durch den Wert $P(\neg W|\neg A)$ geteilt, wodurch der Basiseffekt umgangen werden könnte. Allerdings hat das ursprüngliche Maß schöne Eigenschaften, die Chengs Maß nicht mehr aufweist, wie z.B. dass es sich für kausal unabhängige Ursachen additiv verhält (vgl. Fitelson 2011).

Wir bleiben daher hier bei dem Maß: Stärke(A,W;S) und geben einfach die konkrete Relativierung S mit an, da es zumindest für eine konkrete Situation S den zusätzlichen Beitrag von A für die Erzeugung von W beschreibt. Es weist auch einen gewissen Zusammenhang zu unserer Konzeption abduktiven Schließens auf, bei der uns die Likelihoods P(W|H) für eine Hypothese H einen Aspekt der Erklärungsstärke liefern

sollen. Damit hängen nun die kausale Stärke von A und die Erklärungskraft einer Erklärung mit Hilfe von A in enger Weise zusammen, was wiederum für dieses spezielle Maß spricht.

7.3.4 Relative Häufigkeiten in einer Population

Wichtig ist sicher noch die Frage, von welcher Art von Wahrscheinlichkeit wir hier jeweils mit »P« sprechen. Idealerweise sollte es sich um eine *objektive Wahrscheinlichkeit* im Sinne von *Propensitäten* handeln. Doch die sind uns nur sehr indirekt zugänglich. Zunächst einmal haben wir es bei unseren Daten mit realen relativen Häufigkeiten in einer realen Grundgesamtheit G zu tun, die allerdings selbst wieder die kausalen Verhältnisse verzerrt wiedergeben kann. So könnten zufälligerweise in G bestimmte *Kofaktoren* fast völlig fehlen oder es könnten Korrelationen vorliegen, die entweder bloß zufällig sind oder auf irgendwelche Faktoren zurückzuführen sind, die außerhalb unseres Blickfeldes liegen. Dann haben wir kaum eine Chance, die entsprechenden Kausalverhältnisse in G aufzudecken.

Oder Krankheit K mag in Wahrheit etwa mit einer Tendenz von 80% tödlich sein, aber in der Grundgesamtheit G gibt es vielleicht nicht viele Fälle von K und die sind zufälligerweise de facto nur in 20% der Fälle tödlich. Für genuin probabilistische Kausalzusammenhänge kann so etwas schließlich passieren. Doch uns geht es nicht um die 20% in G, sondern um die 80% Tendenz, die eigentlich für K besteht. Sie wäre auch der richtige Wert für Vorhersagen und nicht die de facto vorfindbare 20%-Quote von Todesfällen bei K. Sie bietet auch den richtigen Wert für die Stärke der Kausalbeziehung, die uns interessiert.

Um uns hier (ähnlich wie im deterministischen Fall) schlicht auf eine Grundgesamtheit G beziehen zu können, könnten wir wieder eine Art von *empirischer Vollständigkeitsannahme* ins Spiel bringen, wonach G die tatsächlichen kausalen Tendenzen korrekt widerspiegelt. Alle kausal möglichen Konstellationen müssten verwirklicht sein, und sie müssen zusätzlich mit den richtigen relativen Häufigkeiten verwirklicht sein. Wir könnten das als die *kausale Repräsentativitätsannahme* für G bezeichnen. Unsere Grundgesamtheit muss *kausal repräsentativ* sein, dann können wir mit »P« hier die relativen Häufigkeiten in der Grundgesamtheit G

bezeichnen, und nur dann können wir auch erwarten, allein anhand der relativen Häufigkeiten in G anhand von (KR) auf die kausalen Beziehungen der Faktoren schließen zu können.

Viele Grundgesamtheiten sind recht groß, weshalb wir dann vielleicht erwarten dürfen, dass G approximativ kausal repräsentativ ist. In jedem Fall kann uns kein Verfahren mehr als das garantieren. Sollte die Natur uns »systematisch hinters Licht führen«, indem eigentlich zu 80% tödliche Krankheiten tatsächlich nur in 20% der Fälle tödlich enden, so haben wir kaum Chancen, das mit irgendeinem kausalen Schlussverfahren zu korrigieren.

Wir können uns das so vorstellen, dass unsere Grundgesamtheit G eine Art von großer »Stichprobe« aus einer imaginären und unendlich Hypergesamtheit H darstellt, in der alle kausalen Beziehungen mit den richtigen Zahlenverhältnissen repräsentiert sind. Bei großen Anzahlen sollte unsere »Stichprobe« G einigermaßen repräsentativ für H sein.

Jedenfalls beobachten wir nur die relativen Häufigkeiten in der realen Grundgesamtheit G und müssen die zugrundeliegenden Wahrscheinlichkeiten P in Form der zugrundeliegenden Propensitäten daraus erst noch erschließen. Außerdem müssen wir uns normalerweise selbst für G wieder auf eine deutlich kleinere *Zufallsstichprobe* S stützen, und haben von S auf die Verhältnisse in G (bzw. auf die basalen Propensitäten) zu schließen. Das zeigt noch einmal die unterschiedlichen Fehlerquellen auf, die beim Schließen aus den Daten (also aus relativen Häufigkeiten) auf die sie erzeugenden Kausalzusammenhänge zu berücksichtigen sind.

Dazu sind jeweils Mittel der schließenden Statistik vonnöten. Um eine Ungleichung wie P(W|F*&S) > P(W|S) zu überprüfen, werden wir etwa die Nullhypothese P(W|F*&S) = P(W|S) mit unseren Daten vergleichen, und nur wenn wir sie statistisch falsifizieren können, die Ungleichung als gegeben akzeptieren. Oder wir arbeiten mit entsprechenden Konfidenzintervallen.

Mit (KR) ist also eine erste Definition für die kausale Relevanz eines einzelnen Prüffaktors gelungen. Aber verfügen wir damit bereits über ein Schlussverfahren? In Analogie zum Differenztest des deterministischen Falls können wir auch hier schließen. Wenn es uns gelingt, zwei bzgl. der Ursachen von W homogene Situationen S1 und S2 herzustellen, die sich nur in A unterscheiden, können wir wie im Falle des Differenztests vor-

gehen. Allerdings müssen wir hier Differenzen von Wahrscheinlichkeiten feststellen und das gelingt letztlich nur durch zahlreiche Wiederholungen. Während wir im deterministischen Fall die Wiederholungen nur als Hinweise benötigten, dass die Homogenitätsbedingung erfüllt ist, benötigen wir die Wiederholungen beider Situationen im probabilistischen Fall, um auf die zugrundeliegenden Wahrscheinlichkeiten zurück zu schließen, die zu den relativen Häufigkeiten in unseren Stichprobenergebnissen geführt haben.

Dazu können wir etwa auf die schon erwähnten Signifikanztests zurückgreifen. Oder wir arbeiten mit Konfidenzintervallen zum Niveau 1–α und gehen dann davon aus, dass (KR) erfüllt ist, wenn das Konfidenzintervall zur wiederholten Situation S1 (mit A) oberhalb des Konfidenzintervalls von S2 liegt und sich die Intervalle somit nicht überschneiden. Dadurch kommen also neue Komplikationen gegenüber dem deterministischen Fall ins Spiel, die vor allem in der schließenden Statistik behandelt werden, aber die Grundidee bleibt hier dieselbe: *Wir vergleichen zwei bzgl. W ansonsten homogenisierte Situationen, die sich nur in A unterscheiden, daraufhin, ob sie sich auch in W (bzw. der Wahrscheinlichkeit für W) unterscheiden.* Ist das in den (signifikant) meisten Instanzen der beiden Situationen der Fall, so halten wir die kausale Relevanz von A für W für gegeben.

7.3.5 Probabilistische singuläre Verursachung

Das Schließen auf *singuläre Kausalbeziehungen* wird allerdings noch einmal deutlich erschwert durch die Aufgabe des Determinismus. Im Falle des Determinismus konnten wir schließen: Wenn A kausal relevant für W ist und eine Instanz a von A vorliegt, für die auch alle Kofaktoren von A instantiiert sind, dann verursacht diese Instanz von A definitiv das Auftreten einer Instanz w von W. Gibt es nur ein solches W-Ereignis w in dem entsprechenden Raum-Zeit-Gebiet, so ist a eine Ursache von w. Es könnte sich zusätzlich noch um einen Fall von Überdetermination handeln, wonach zusätzlich eine andere Ursache b von e vorliegt, aber wir wissen trotzdem definitiv, dass a eine Ursache von e ist.

Das ist im probabilistischen Fall viel schwieriger. Gibt es mehrere potentielle probabilistische Ursachen, werden wir oft nicht ermitteln

können, welche davon tatsächlich wirksam war. So könnte es passieren, dass A zwar die (singuläre) Wahrscheinlichkeit für W erhöht hat und W auch eingetreten ist, aber dabei durch eine alternative Ursache C verursacht wurde. Die Wahrscheinlichkeitserhöhung ist dann keineswegs direkt mit einer Ursächlichkeit zu identifizieren.

Dass das durchaus praktische Bedeutung haben kann, offenbart das folgende Beispiel: Im Jahre 1999 wurde das Schmerzmedikament Vioxx in den Markt eingeführt, generierte schon bald Milliardenumsätze und wurde 2004 wieder vom Markt genommen, weil sich gezeigt hatte, dass es u.a. die Wahrscheinlichkeit von Herzinfarkten erhöht. Die Angehörigen derjenigen Patienten, die nach längerer Einnahme von Vioxx am Herzinfarkt gestorben waren, klagten gegen den Hersteller Merck auf Schadenersatz. In den USA bekamen sie Millionensummen zugesprochen, nicht aber in Deutschland.

Das Problem war jedes Mal der Nachweis im Einzelfall, dass die Einnahme von Vioxx die tatsächliche Ursache des jeweiligen Herzinfarktes war. Es gibt neben Vioxx viele andere Risikofaktoren für Herzinfarkt, und wir wissen, dass selbst junge Menschen manchmal überraschend daran versterben. Bei älteren Personen sind mehrere Risikofaktoren meist schon deutlicher erkennbar – wie etwa erhöhte Cholesterinwerte – so dass es im Einzelfall kaum noch zwingend nachweisbar ist, dass der eine oder der andere Faktor einen konkreten Herzinfarkt ausgelöst hat, denn beide kennen wir nur als probabilistische Faktoren, die einen Herzinfarkt begünstigen, aber nicht unbedingt jedes Mal herbeiführen. Ob sie also tatsächlich für das Auftreten des Herzinfarktes wirksam waren, wissen wir daher normalerweise nicht. Diese immer verbleibende Unsicherheit ist den Klägern in Deutschland regelmäßig zum Verhängnis geworden.

Am besten kann in solchen Fällen vermutlich der Prozessansatz der Kausalität die singulären Zusammenhänge beschreiben. Wir müssten im Prinzip aufzeigen, welcher genaue Prozess stattgefunden hat, der von a nach w führte. Haben wir darüber aber keine speziellen Informationen, und es gibt mehrere mögliche probabilistische Ursachen, können wir scheinbar nicht mehr definitiv entscheiden, welche davon tatsächlich wirksam wurde und welche nur potentiell blieb.

Machen wir uns das noch einmal an einem einfachen Beispiel klar. Nehmen wir an, wir hätten einen *gezinkten Würfel* mit einem kleinen

eingebauten Bleiplättchen auf Seite der Eins, so dass die Wahrscheinlichkeit eine Sechs zu würfeln 1/5 wäre. Nun würfeln wir tatsächlich eine Sechs. Ist in diesem konkreten Fall das eingebaute Bleiplättchen eine (Mit-) Ursache für diese Sechs? Auch ohne das Plättchen würden wir hin und wieder eine Sechs werfen, woher wissen wir also, dass diese Eigenschaft des Würfels bzw. des Wurfes für unser Ergebnis kausal wirksam wurde? Ohne exaktere Informationen über den Vorgang dieses speziellen Wurfes lässt sich das nicht entscheiden. Haben wir die, könnte es deshalb entscheidbar sein, weil wir das Beispiel eigentlich als einen deterministischen Vorgang deuten.

Stellen wir ihn uns aber *genuin indeterministisch* vor, scheint die Frage nicht mehr sinnvoll beantwortbar zu sein. Wir könnten eigentlich nur sagen, dass wir dann die Verfälschung des Würfels in jedem Fall als eine Teilursache betrachten sollten, wenn keine speziellen Informationen dagegen sprechen. Nur in der komfortablen Situation, dass wir alle anderen potentiellen Ursachen ausschließen können, können wir ein Ereignis bzw. einen Faktor dann definitiv als Ursache bezeichnen, obwohl er nur probabilistisch wirkt. Es bleibt für diese Entscheidung schließlich wieder nur eine Abwägung im Sinne eines Schlusses auf die beste Erklärung übrig, wonach unser bester Tipp für eine Verursachung das Ereignis primär nennt, dass die Wirkung am besten erklärt, das also kausal deutlich am stärksten dazu beiträgt.

Aber selbst in diesem Fall wird man die anderen potentiellen Ursachen vermutlich als *beitragende Ursachen* für den Einzelfall ansehen müssen, wenn es nicht spezielle Hinweise darauf gibt, dass sie dort nicht zum Tragen kamen. Es scheint in solchen Fällen also noch nicht einmal immer Tatsachen zu geben, die die Frage entscheiden würden, welche der genuin probabilistischen Ursachen zum Zuge kam. Eine Unterscheidung zwischen nur potentiellen probabilistischen und den tatsächlich wirksamen probabilistischen Ursachen ist dann vermutlich nicht mehr sinnvoll.

Wir müssen hier schließlich festlegen, was wir eigentlich unter *singulärer probabilistischer Kausalität* verstehen wollen. Denken wir uns diese probabilistischen Ursachen zumindest als irgendwie beitragend, wäre es jedenfalls nur konsequent und passend zum deterministischen Fall sie schlicht als Ursachen einzustufen, wie wir das auch im de-

terministischen Fall für die Kofaktoren taten. Die Kläger im Vioxx-Fall sind dann womöglich u.a. an einem ungeeigneten Verständnis von singulärer probabilistischer Verursachung gescheitert. Es wurde von ihnen etwas Unmögliches verlangt. Man kann in diesen Fällen höchstens noch fragen, in welchem Maße die Teilursachen jeweils zum Herzinfarkt beigetragen haben und sich dafür auf die Stärke der verschiedenen Kausalfaktoren beziehen. Man kann jedoch nicht den Nachweis verlangen, dass bestimmte potentielle probabilistische Ursachen auch aktuale Ursachen waren, denn das ist in solchen Fällen unmöglich zu beweisen und beruht wohl eher auf einem falschen Verständnis von singulärer probabilistischer Verursachung.

Gerade für den Bereich der singulären Kausalität scheinen klassische *kontrafaktischen Ansätze* zur Kausalität (vgl. Lewis 1973, Kvart 2001 oder Paul & Hall 2013) besser geeignet zu sein, zu explizieren, was wir mit der direkten Verursachung eines Ereignisses meinen. Schwieriger ist dann allerdings die Frage zu beantworten, wie wir diese konkreten kontrafaktischen Zusammenhänge ermitteln sollen. Daher stehen hier die kontrafaktischen Ansätze auch nicht im Vordergrund. Eine intuitiv hilfreiche Idee ist aber jedenfalls, dass e_1 Ursache von e_2 ist, wenn gilt, dass e_2 nicht so stattgefunden hätte, wenn e_1 nicht der Fall gewesen wäre. Allerdings bleiben die sogenannten Preemption-Probleme, die Überdetermination und andere Fragen weiter zu diskutieren, wozu ich an anderer Stelle schon Einiges gesagt habe (vgl. Bartelborth 2007). Außerdem handelt es sich wie gesagt zu einem Teil schon um ein Problem der Definition der Kausalbeziehung in solchen Fällen und nicht nur um eines des kausalen Schließens (vgl. auch Paul & Hall 2013).

Schon im Falle der *Überdetermination* zeigen sich die Grenzen des kontrafaktischen Ansatzes. Verursachen etwa zwei weggeworfene Zigaretten (z_1 und z_2) einen Waldbrand (w), wobei jede der Zigaretten für sich bereits den Waldbrand ausgelöst hätte (vgl. Woodward 2003), dann wird die kontrafaktische Abhängigkeit des Ereignisses w von z_1 und z_2 nicht sichtbar, weil, auch wenn eines der beiden Ereignisse nicht stattgefunden hätte, trotzdem der Waldbrand ausgebrochen wäre, da schließlich die andere Zigarette dafür genügte. Um die hier versteckte kontrafaktische Abhängigkeit doch wieder sichtbar werden zu lassen, müssen wir schon einige Verrenkungen anstellen (vgl. Collins 2006).

Woodward (2003) diskutiert einen Vorschlag, wie man sie dadurch sichtbar werden lässt, dass man kontrafaktische Situationen betrachtet, die gegenüber der aktuellen Situation nur Veränderungen in ganz bestimmten Variablen und in bestimmten Bereichen erlauben. In unserem Beispiel wird die kontrafaktische Abhängigkeit natürlich dann erkennbar, wenn wir uns fragen, welche Auswirkungen z_1 versus non-z_1 hätte, sobald wir vergleichbare Situationen nur eben ohne z_2 betrachten. Aber das ist ein komplizierter Ansatz zur Analyse von Kausalität, den ich hier nicht weiterverfolgen möchte.

Wir könnten schließlich für das probabilistische kausale Schließen auch noch das Analogon zum Vierertest des deterministischen Schließens entwickeln und so versuchen, etwas komplexere Kausalstrukturen aufzudecken. Doch zu diesem Zweck werden wir später noch einen anderen Ansatz kennenlernen, der dazu dienen wird, kausale Strukturen anhand bayesscher Netze zu entdecken.

Einen direkteren singulären Zugang zur probabilistischen Kausalität finden wir etwa in Kvart (2001), der insbesondere noch weitere Komplikationen durch Zwischenfaktoren erörtert. Allerdings eignet er sich noch weniger als die hier diskutierten Ansätze für das kausale Schließen, da er auf singulären Wahrscheinlichkeiten beruht, die epistemisch nicht besonders gut zugänglich sind.

Tatsächlich genügt im genuin probabilistischen Fall (also in einer indeterministischen Welt) von singulärer Kausalität unsere bisherige Ungleichung $P(W|A\&S) > P(W|S)$ für die kausale Relevanz nicht mehr, um Kausalbeziehungen eindeutig zu kennzeichnen. Im konkreten Fall (mit objektiven Einzelfallwahrscheinlichkeiten) denken wir uns $P(W|A\&S)$ meist als die singuläre Wahrscheinlichkeit zum Zeitpunkt, an dem gerade A eintritt. Doch Kvart (1997) hat uns gezeigt, dass das im Fall einer indeterministischen Welt nicht ausreicht.

Schauen wir uns dazu das folgende Beispiel an: Nehmen wir an, ich investiere zu t_1 viel Geld in die Aktien einer Pharmafirma (A). Da die keine guten Mitarbeiter hat, ist deren Erfolgswahrscheinlichkeit klein und damit die Wahrscheinlichkeit, dass ich zu t_2 reich sein werde (R). Es könnte etwa gelten: $0{,}1 = P(R|A) < P(R|\neg A) = 0{,}3$. Aber der unwahrscheinliche Fall kann eben doch eintreten, dass die Firma zwischen t_1 und t_2 durch Zufall eine grandiose Entdeckung (E) macht und hohe Profite

einfährt. Dann hat sich mein Aktienkauf rentiert und war tatsächlich die Ursache dafür, dass ich nun (zu t_2) reich bin.

Kvart nennt dann E einen »Increaser« (Erhöher). Der kann im Zeitraum zwischen dem Auftreten von A und dem von R eintreten und entscheidet dann erst darüber, ob sich A als Ursache von R erweist. Das können jedenfalls *stabile* Increaser, die nicht durch andere Ereignisse (sogenannte »Decreaser«) wieder zunichte gemacht werden. Zunächst wächst durch den Increaser die Wahrscheinlichkeit an: (*) P(R|A&E) > P(R|E), aber wenn dann nach dem Auftreten von E die Zulassungsbehörde die Zulassung verweigert (V), gilt wiederum: P(R|A&E&V) < P(R|E&V), wodurch A als Ursache meines Reichtums wieder ausfällt. Es muss sich bei E also um einen stabilen Increaser handeln, der stabil die Erhöhung der Wahrscheinlichkeit liefert, weil wir zu allen potentiellen Decreasern immer wieder geeignete Increaser finden, die sie ausgleichen, so dass die Ungleichung (*) letztlich Bestand hat. Unsere probabilistische Definition von Ursachen, müsste dann noch diese Zwischenereignisse berücksichtigen.

Ganz ähnlich ist das beim Kauf eines Lotterieloses. In den meisten Fällen ist es eine Geldverschwendung. Sollte aber durch Zufall gerade doch mein Los gezogen werden, erweist sich das als Increaser für die Wahrscheinlichkeit meines Reichtums. Ob dann also das Los eine Ursache meines späteren Reichtums war, lässt sich nicht schon zum Kaufzeitpunkt des Loses feststellen, sondern erst später bei der Ziehung der Lose.

So lassen sich auch andere bekannte, vermeintliche Gegenbeispiele gegen die probabilistische Konzeption aufklären. Wenn der Golfer den Ball schräg trifft, wird das im Normalfall seine Chancen auf ein Ass (»hole in one«) verringern. Wenn der Ball aber im konkreten Einzelfall einen Baum trifft und daran so abprallt, dass er direkt ins Loch geht, so war er dann doch die Ursache für das Ass. Das kann man entweder so beschreiben, dass die aktuale *Einzelfallwahrscheinlichkeit* für ein Ass sogleich mit dem schrägen Schlag angestiegen ist, oder so, dass der Ball erst unterwegs etwa durch den Wind so getrieben wurde, dass das Ass zustande kam. Dann war der treibende Wind ein Increaser im Sinne von Kvart. So wird erklärlich, wieso das Golfbeispiel auf der generischen Ebene als Problemfall erscheint, aber auf der singulären Ebene doch

korrekt beschrieben werden kann. Das zeigt zugleich, wie komplex die kausalen Verhältnisse (in einer indeterministischen Welt) werden können und wie schwierig es dann sein kann, die Kausalverhältnisse zu ermitteln.

7.3.6 Quantitative Zufallsvariablen

Wir können schließlich noch quantitative Zufallsvariablen für bestimmte Faktoren einführen, um der Tatsache Rechnung zu tragen, dass unsere Faktoren oft viele unterschiedliche Ausprägungen mit unterschiedlichen Wirkungen annehmen können. Diese Zufallsvariablen sind Funktionen $X{:}\Omega \rightarrow \mathbb{R}$ eines Grundraumes von Ereignissen Ω in einen Zahlenraum $\mathbb{R}$ etwa die Menge der reellen Zahlen. Sie dienen u.a. dazu, bestimmte Ergebnisse von Zufallsprozessen zu beschreiben, woher der Name rührt. So kann einem Wurf mit drei Würfeln z.B. die *Summe der Augen* der Würfel zugeordnet werden. Dann erhalten wir eine Wahrscheinlichkeitsverteilung für X als Verteilung auf der Menge der Ergebnisse von X. Dafür schreiben wir etwa (X=3) für das Urbild der 3 unter X, also für die Menge aller Würfe mit insgesamt drei Augen: $\{\omega \in \Omega;\ X(\omega)=3\}$. Da es dafür nur ein Elementarereignis gibt, gilt $P(X=3) = 1/216$, wenn es sich um faire Würfel handelt.

Oft wird der Grundraum Ω überhaupt nicht mehr angegeben, sondern wir wenden uns direkt den Wahrscheinlichkeiten für die Zufallsvariable X zu. Die Variable X gibt einfach ein bestimmtes Merkmal unserer Situation wieder, indem sie eine bestimmte Zahl annimmt. Für die verschiedenen Ausprägungen dieses Merkmals müssen wir dann nur noch wissen, wie groß die Wahrscheinlichkeiten dafür sind, dass X diese Werte annimmt, um mit diesen Zufallsvariablen rechnen zu können. So ersparen wir uns gerne die Konstruktion eines ursprünglichen Grundraums Ω, der oft nicht leicht zu konstruieren ist. Wir haben vielleicht andere Gründe für die Annahme, dass die Werte z.B. normalverteilt sind und können die speziellen Parameter (μ und σ) der Verteilung anhand einer Stichprobe schätzen oder kennen sie bereits aus anderen Untersuchungen. Die Zufallsvariablen stehen im Folgenden meist für kausale Faktoren, die bestimmte Ausprägungen annehmen können. So könnte X für die Stärke oder Häufigkeit der Nebenwirkungen stehen oder für die Anzahl der eingenommenen Tabletten etc.

Unsere bisherigen dichotomen Faktoren F sind eine Art Sonderfall davon und lassen sich durch eine Zufallsvariable mit den Werten 0 und 1 charakterisieren. Das Vorliegen von F wird dann etwa durch X = 1 und die Abwesenheit von F durch X = 0 wiedergegeben. Außerdem werden bestimmte Konstellationen von Zufallsvariablen $X_1,\ldots,X_n$ auf einfache Weise dargestellt, so schreiben wir vereinfachend:

$$P(x_1,\ldots,x_n) = P(X_1=x_1 \;\&\ldots\&\; X_n=x_n)$$

für die gemeinsame Verteilung der Zufallsvariablen $X_1,\ldots,X_n$. Insbesondere müssen wir nun nicht mehr nur einfache Korrelationen von Zufallsvariablen berücksichtigen, sondern haben auch eine komplexere Definition der *statistischen Unabhängigkeit* anzuwenden:

> Danach sind zwei Zufallsvariablen X und Y *statistisch unabhängig*, wenn für *alle Werte x und y* aus ihren jeweiligen Wertebereichen gilt: $P(x|y) = P(x)$ bzw. äquivalent dazu: $P(X=x \;\&\; Y=y) = P(X=x)\cdot P(Y=y)$
> Das schreiben wir auch manchmal einfacher: $P(X,Y) = P(X)\cdot P(Y)$

Entsprechend sagen wir von n Zufallsvariablen $X_1,\ldots,X_n$, dass sie statistisch unabhängig sind, wenn gilt: $P(x_1,x_2\ldots,x_n) = P(x_1)\cdot\ldots\cdot P(x_n)$ für alle Werte $x_1,\ldots,x_n$, die unsere Zufallsvariablen $X_1,\ldots,X_n$ annehmen können, bzw wenn gilt: $P(X_1,X_2\ldots,X_n) = P(X_1)\cdot\ldots\cdot P(X_n)$.

Wenn wir nun über eine Liste von Zufallsvariablen $X_1,\ldots,X_n$ verfügen, die alle Faktoren darstellen, die potentiell auf W wirken, dann können wir unsere Definition für kausalen Einfluss im Prinzip übertragen auf den Fall quantitativer Zufallsvariablen:

> **(KRQ) X_1 ist kausal relevant für W gdw.**: Es gibt reelle Zahlen
> $w, x_1, x_1^*, x_2,\ldots,x_n$ mit: $P(w|x_1,x_2,\ldots,x_n) > P(w|x_1^*,x_2,\ldots,x_n)$

Das soll bedeuten, dass es zwei Werte x_1 und x_1^* für den Faktor X_1 gibt, die zu unterschiedlichen Wahrscheinlichkeiten für das Auftreten des Wertes w für den Faktor W führen, wenn alle anderen für W kausal relevanten Faktoren konstant bleiben. Normalerweise würden wir das bei kontinuierlichen Größen X_i auch nicht nur für isolierte Werte verlangen,

sondern eher schon davon ausgehen, dass die Wirkungen zumindest in kleinen Intervallen von entsprechenden Werten auftreten, doch wir ersparen uns hier die daraus resultierenden technischen Probleme. Die Idee dürfte genügend klar sein.

Das Ermitteln solcher Beziehungen ist natürlich wieder mit den oben genannten Schwierigkeiten behaftet. Erstens meinen wir mit »P« genau genommen die tatsächlichen objektiven Wahrscheinlichkeiten, haben stattdessen aber nur bestimmte relative Häufigkeiten zur Verfügung, und zweitens ist es sehr schwer, die geforderte Homogenisierung zu finden oder herbeizuführen. Trotzdem bietet (KRQ) zumindest einen prinzipiellen Weg zum Schließen auf kausale Beziehungen.

Das möchte ich noch aus einer etwas anderen Perspektive beleuchten, nämlich mit Hilfe des sogenannten *interventionalistischen Ansatzes*, der auf den ersten Blick einen anderen Zugang zur Kausalität bietet, nämlich über die Veränderung bestimmter Variablen, aber schließlich doch einen ähnlichen Grundgedanken verfolgt. Er beschreibt auf etwas andere Weise, wie wir die Änderung an der Variable X_1 in der Gleichung (KRQ) so herbeiführen können, dass alle anderen Faktoren konstant gehalten werden, um damit die isolierte Wirkung von X_1 auf W bestimmen zu können.

7.3.7 Der interventionalistische Ansatz zur Kausalität

Der interventionalistische Ansatz der Kausalität stellt in den Vordergrund, dass ich mit einem *Eingreifen* in die Welt (einer Intervention), wenn dieses lokal erfolgt und nur ganz bestimmte Faktoren (nämlich die in Frage stehenden potentiellen Ursachen) ändert, andere Faktoren (die Wirkungen) damit beeinflussen und somit ebenfalls *ändern* kann. Das ist die wichtigste Eigenschaft der Kausalbeziehung, die sie für uns so bedeutsam macht und sie insbesondere von bloßen Korrelationen unterscheidet. Gelbe Finger mögen mit dem Auftreten von Lungenkrebs positiv korreliert sein, aber das gibt uns keinen Hinweis darauf, was wir gegen Lungenkrebs unternehmen können. Das Verändern der Gelbheit der Finger ändert nicht die Wahrscheinlichkeit für einen Lungenkrebs. Das Rauchen ist hingegen Ursache des Lungenkrebses und diese Information sagt zugleich, was wir gegen diese Krankheit tun können, nämlich

mit dem Rauchen aufzuhören. Diese Änderung würde auch meine Wahrscheinlichkeit für einen Lungenkrebs verändern. Diese Übertragung von (kontrafaktischen) Änderungen von der Wirkung auf die Ursache ist nun das charakteristische Merkmal kausaler Beziehungen gemäß dem interventionalistischen Ansatz.

> **Die Idee des interventionalistischen Ansatzes**: X ist *Ursache* von Y, wenn ich durch eine geeignete Änderung von X (eine Intervention an X) eine Änderung von Y herbeiführen kann.

Diese geeigneten Änderungen sollen Interventionen sein, die nur den Wert von X verändern, ohne dabei die Werte der anderen direkten Ursachen von Y zu verändern. Dazu gibt Woodward uns zunächst die Idee:

> (TC) X is a **total cause** of Y if and only if there is a possible intervention on X that will change Y or the probability distribution of Y. (Woodward 2003, 51)

Wesentlich ist es nun zunächst, den Begriff der Intervention genauer zu fassen. Woodward (2003: Kap. III) führt zu diesem Zweck sogenannte *Interventionsvariablen* I ein (bei Pearl (2000) finden wir stattdessen den do-Operator, der schlicht und einfach per do(X=x) eine Variable X auf einen Wert x setzt). Die Interventionsvariablen I müssen bestimmte Eigenschaften aufweisen. Eine Intervention ist einem idealen Experiment nachgebildet. So wie in einem idealen Experiment der genaue Einfluss der Variable X auf Y bestimmt werden soll, wobei eine Vermengung mit anderen Einflüssen verhindert wird, indem die anderen Einflussfaktoren konstant gehalten werden, so ist das auch hier sicherzustellen. Dazu definiert Woodward, wann eine Größe I eine Interventionsvariable für X bezüglich Y ist. Hierzu gehört zunächst, dass I wie ein Schalter X auf einen bestimmten Wert bringen kann und X dabei von anderen Einflüssen abschneidet. Außerdem darf I die Größe Y nur auf dem Weg über X beeinflussen und nicht direkt oder auf anderen Wegen

(Nebengleisen) ohne X. Außerdem sollte I nicht schon mit anderen Größen Z korreliert sein, die ihrerseits Y beeinflussen. Mit Hilfe von I können wir dann unsere gewünschten Interventionen an X im Hinblick auf Y vornehmen. Woodward (2000, 98) beschreibt das so:

> (**IV**)
> I1. I causes X.
>
> I2. I acts as a switch for all the other variables that cause X. That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I.
>
> I3. Any directed path from I to Y goes through X. That is, I does not directly cause Y and is not a cause of any causes of Y that are distinct from X except, of course, for those causes of Y, if any, that are built into the I-X-Y connection itself; that is, except for (a) any causes of Y that are effects of X (i.e., variables that are causally between X and Y) and (b) any causes of Y that are between 7 and X and have no effect on Y independently of X.
>
> I4. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X.
> ('Cause' in this characterization always means 'contributing cause' rather than 'total cause.')

Mit Hilfe dieser Definition von Interventionsvariablen lässt sich dann auch festlegen, was eine Intervention ist:

> (**IN**) I's assuming some value $I = z_i$, is an *intervention* on X with respect to Y if and only if I is an intervention variable for X with respect to Y and $I = z_i$ is an actual cause of the value taken by X.

Woodward unterscheidet hier noch zwischen dem »total cause« und dem »contributing cause«. Falls X auf mehreren Wegen Y beeinflusst, soll der »total cause« die Gesamtbeeinflussung beschreiben, während die »beitragenden Ursachen« die Einflüsse der einzelnen Wege beschreiben. Hier werden die Größen auf den anderen Wegen von X nach Y auch noch konstant gehalten. Das ist besonders für den recht speziellen Fall gedacht,

dass sie sich gegenseitig aufheben. In derartigen Fällen, Woodward (2003: 49 u. 64) spricht von einem »failure of faithfulness« (das Konzept erläutere ich später in Kap. 7.3.9), kann sich die Gesamtwirkung der zwei Wege aufheben. X wirkt direkt positiv auf Y, aber auch positiv auf Z, was seinerseits einen negativen Einfluss auf Y hat. Um die direkte Wirkung von X auf Y sichtbar zu machen, müssen wir in einem Experiment zusätzlich Z kontrollieren (festhalten auf einem bestimmten Wert), während wir X verändern.

Ein einfaches Beispiel von Woodward kann die allgemeine Idee verdeutlichen. Wir möchten wissen, ob ein Medikament M eine Krankheit K heilt. X besagt, ob jemand M bekommt (1) oder nicht (0), und Y beschreibt, ob der Betreffende geheilt wird (1) oder nicht (0). (Unsere Variablen nehmen also nur zwei konkrete Werte an.) Wir experimentieren nun im Prinzip mit einer Versuchsperson (bzw. zwei Personen oder sogar zwei Gruppen von Versuchspersonen, der Behandlungs- und der Kontrollgruppe, weil das Experiment praktisch nur so durchführbar ist), indem wir ihr einmal das Medikament bei Krankheit verabreichen (I) und schauen, was passiert, und einmal in *derselben Situation* das Medikament nicht verabreichen.

So möchten wir den Einfluss des Medikaments (also den der Variable X) kennen lernen und zum Beispiel von störenden Faktoren unterscheiden, wie dass das Verabreichen selbst (Placeboeffekt) ohne den Weg über X bereits die Krankheit heilt. Wir möchten auch nicht, dass das Medikament speziell nur den widerstandskräftigeren Patienten verabreicht wird (I soll also unabhängig von anderen Faktoren Z, die Y beeinflussen, eingreifen). Alles das könnte unser Ergebnis verfälschen. I soll eine Intervention darstellen, die nur den isolierten Einfluss von X auf Y aufzeigen soll (vgl. auch Bartelborth 2007). Dazu dient die Intervention, die nur genau eine Änderung an X vornimmt, während alle anderen Einflussgrößen, die noch auf Y wirken, konstant gehalten werden, was wir in der Regel durch einen speziellen Versuchsaufbau mit Randomisierung zu erreichen hoffen.

Bisher hatten wir nur Woodwards Definition von »total cause«, allgemeiner gilt für *beitragende Ursachen*:

> (M) A necessary and sufficient condition for X to be a (type-level) *direct cause* of Y with respect to a variable set ***V*** is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in ***V***. A necessary and sufficient condition for X to be a (type-level) contributing cause of Y with respect to variable set ***V*** is that (i) there be a directed path from X to Y such that each link in this path is a direct causal relationship; that is, a set of variables $Z_1 \ldots Z_n$ such that X is a direct cause of Z_1, which is in turn a direct cause of Z_2, which is a direct cause of $\ldots Z_n$, which is a direct cause of Y, and that (ii) *there be some intervention on X that will change Y when all other variables in* ***V*** *that are not on this path are fixed at some value*. If there is only one path P from X to Y or if the only alternative path from X to Y besides P contains no intermediate variables (i.e., is direct), then X is a contributing cause of Y as long as there is some intervention on X that will change the value of Y, for some values of the other variables in V. (Woodward 2000, 59; kursiv von mir)

Zunächst wird hier eine Relativierung auf eine bestimmte Variablenmenge ***V*** vorgenommen, womit wir explizit die Menge anführen, von der wir annehmen, dass sie zumindest alle gemeinsamen Ursachen von Faktoren innerhalb von ***V*** enthält. Dann wird definiert, was eine direkte Ursache ist und mit Hilfe dieser Bestimmung werden auch Ursachenketten und damit *indirekte Verursachung* definiert. Die erkennen wir daran, dass wir alle Variablen, die nicht auf dem aktiven Pfad von X zu Y liegen, festhalten und dabei eine Veränderung von X zu einer von Y führt.

Das aktive Element der *Intervention* ist zwar neu, aber es dient letztlich vor allem dem Zweck, wieder zwei Situationen herbeizuführen (einmal mit und einmal ohne Intervention), die homogen sind im Hinblick auf Y mit der Ausnahme von X und so den Effekt von X auf Y zu untersuchen gestatten. Interventionen an X relativ zu Y werden gerade so beschrieben, dass sie alle anderen für Y kausal relevanten Faktoren unverändert lassen, die nicht auf einem Pfad von X zu Y liegen. Die Grundidee bleibt daher dieselbe wie in unserer obigen Definition (KRQ) aus dem letzten Abschnitt. Es wird nur etwas anschaulicher dargestellt, wie wir derartige kausale Vergleichssituationen herbeiführen können. Wenn

wir eine Situation s haben, so können wir sie durch eine Intervention I in eine geeignete Vergleichssituation s* überführen. Das kann man sich zunächst sogar als zeitliche Abfolge vorstellen. Hat jemand über einen gewissen Zeitraum eine Krankheit, die nicht heilt, und wird dann gesund, nachdem er eine bestimmte Medizin erhalten hat, ohne dass sich ansonsten etwas an seinen Lebensumständen ändert (das ist die Intervention I), dann dürfen wir schließen, dass die Medizin die Heilung bewirkt hat.

Leider ist dieser Schluss nicht ganz so überzeugend, wie er auf den ersten Blick zu sein scheint. Wir erwarten z.B. nicht, dass die neue Situation in allen Aspekten der alten Situation vor der Einnahme gleicht. Vielleicht hat sich inzwischen das Immunsystem auf die Krankheit eingestellt und hätte auch ohne das Medikament die Krankheit nun besiegt. Vielleicht war es auch nur der Glaube daran, nun eine wirkungsvolle Medizin zu erhalten, der zu der Heilung geführt hat (im Sinne eines Placebo-Effekts). Vielleicht war es auch nur die Zuwendung, die mit der Gabe der Medizin verbunden war. Vielleicht haben sich auch noch andere Umstände geändert, die uns nur nicht bekannt sind.

Um diesen Dingen Rechnung zu tragen, wird die Intervention oft nur als *kontrafaktische Möglichkeit* eingeführt. Hätte sie stattgefunden, während alle anderen Faktoren gleich geblieben wären, dann hätte sie eine bestimmte Änderung von Y oder der Wahrscheinlichkeiten für Y zur Folge gehabt. Damit werden zumindest bestimmte zeitliche Entwicklungen (wie hier die zu erwartende Veränderung des Immunsystems) ausgeschlossen. Allerdings ist der Preis zu bezahlen, dass man diese kontrafaktische Situation nicht tatsächlich beobachten kann. Also müssen wir doch wieder auf einen Vergleich realer Situationen zurückgreifen. Dafür denkt auch Woodward (2003) an einen Vergleich von Experimental- und Kontrollgruppe, wie er oben schon beschrieben wurde.

Wir haben allerdings noch ein Problem außen vor gelassen, das noch wesentlich ist. Im Unterschied zum deterministischen Fall müssen wir auch noch berücksichtigen, dass sich vielleicht nur die Wahrscheinlichkeiten verschieben, ohne dass es im Einzelfall gleich sichtbare Effekte gibt. Im Falle unserer Heilung kann es schon immer eine gewisse Wahrscheinlichkeit für eine Selbstheilung gegeben haben, die

nur zufällig nach der Gabe des Medikaments zum Tragen kam. Oder es hat die Medizin zwar die Heilungschancen für den Patienten deutlich erhöht, aber der ist trotzdem nicht genesen. Wir müssen also noch berücksichtigen, dass es sich um einen probabilistischen Effekt handelt, der nur im Durchschnitt einer größeren Gruppe auch beobachtbar ist. Pearl (2000) beschreibt das so, dass nur die jeweiligen (bedingten) Erwartungswerte sich unterscheiden. Wenn man X zunächst auf x und dann auf x* setzt, dürfen wir das Ausmaß des Effekts also so beschreiben:

Maß für den Effekt: $E(Y|do(X=x^*)) - E(Y|do(X=x))$

Das entspricht unserem obigen Maß für die Stärke der Kausalbeziehung für eine bestimmte Population im Durchschnitt, wobei hier der do-Operator sicherstellen soll, dass jeweils nur homogene Situationen miteinander verglichen werden. Um solche probabilistischen Effekte jedoch feststellen zu können, sind wir natürlich wieder auf größere Stichproben angewiesen, wobei die Elemente der beiden Gruppen sich untereinander ähnlich sein sollten.

Schauen wir noch einmal auf ein Beispiel: Wir geben den Kranken Franz und Karla beiden eine für bestimmte Personen (bestimmter Kofaktor) wirksame Medizin. Dann kann es aber sein, dass sie bei Franz positiv wirkt, aber speziell bei Karla sogar schädlich ist, so dass nur Franz gesund wird. Eine solche 50%-tige Heilung sei aber auch ohne Medizin normal. Daher werden wir dann fälschlicherweise schließen, dass die Medizin unwirksam ist. Um solchen Fehlschlüssen zu entgehen, müssen die Personen in unseren Gruppen gewisse Ähnlichkeiten untereinander aufweisen. Auch das versucht man mit einer Randomisierung zu erzielen, doch das Beispiel weist schon den Weg wie das schiefgehen kann.

Ist in unserer Grundgesamtheit G der Kofaktor K in 50% der Fälle gegeben und non-K also in den anderen Fällen, aber nur zusammen mit K ist unsere Medizin wirksam und sonst nur schädlich, so dass die natürliche Genesung, die im Durchschnitt in 50% aller Fälle auftritt, verhindert wird, so liegt K auch bei einer idealen Randomisierung in unseren Gruppen jeweils in 50% der Fälle vor. So tritt eine Genesung dann in der Experimentalgruppe und der Kontrollgruppe in 50% aller Fälle auf und das Medikament wird als völlig unwirksam eingestuft,

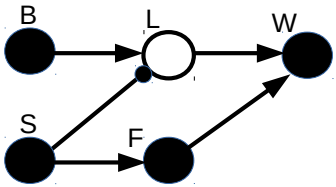
obwohl es bei 50% der Bevölkerung eine 100%-tige Wirkung aufweist und bei den anderen 50% zu 100% schädlich ist! Das belegt wieder, dass eine bewusste Homogenisierung der beiden Gruppen nicht immer durch eine Randomisierung zu ersetzen ist. Hier wird auch der Nachteil sichtbar, den wir in Kauf nehmen müssen, wenn wir die Stärke eines Effekts nur als Durchschnittswert (bzw. Erwartungswert) festlegen.

Gegen den interventionalistischen Ansatz ist zudem der Einwand erhoben worden (vgl. Baumgartner 2009, 2009a), dass sie als Definition und damit zugleich als Schlussverfahren ungeeignet sei, weil sie nicht wirklich *reduktiv* ist. Die Entscheidung, ob X Ursache von Y ist, setzt bereits Wissen darüber voraus, ob I eine Interventionsvariable für X relativ zu Y ist, ob also I Ursache von X ist und I keine Ursache von anderen Ursachen von Y ist. Der Einwand scheint mir zwar korrekt zu sein, aber die anderen Ansätze genau genommen ebenso zu treffen. Wir kehren damit wieder zurück zu der Einsicht von Nancy Cartwright: »No causes in, no causes out«.

Der interventionalistische Ansatz nutzt gerne Gleichungssysteme zwischen bestimmten Zufallsvariablen, um damit die direkten kausalen Beziehungen zu beschreiben. Wir müssen demnach ein *strukturelles kausales Modell* unserer Situation aufstellen, indem wir für die direkten Kausalbeziehungen (bzw. direkten funktionalen Abhängigkeiten) Gleichungen formulieren, und können dann daraus ablesen, welche indirekten Kausalbeziehungen vorliegen. So erhofft man sich, auch die Preemption-Probleme lösen zu können.

Dazu kurz das Standard-Beispiel: Suzy wirft einen Stein in Richtung einer Flasche (S), die dann zerspringt (W). Billy steht bereit, ebenfalls zu werfen (B), lässt aber davon ab (L), als er sieht, wie Suzy den Stein wirft. Nun gilt aber: Hätte Suzy den Stein nicht geworfen, hätte Billy den Stein geworfen und damit wäre die Flasche ebenfalls zerstört worden. (Wir gehen hier vereinfachend davon aus, dass beide perfekte Werfer sind und wir nicht noch einen probabilistischen Fehlerterm benötigen). Das lässt sich mit folgendem Neuronendiagramm beschreiben, wobei S eine hemmende Wirkung auf L ausübt und wir eine Zwischenvariable F einfügen, die Suzys fliegenden Stein beschreibt. Das ist ein typisches Verfahren, um die Zusammenhänge besser beschreiben zu können. Wir

müssen nur zeigen, wie sich unsere Kausalzusammenhänge angeben lassen, wenn wir geeignete Zwischenvariablen finden.



Graphik 7.3: Einfaches Preemption-Beispiel

Wir finden dazu das folgende Gleichungssystem bzw. strukturelle Modell (s. Hall 2007), wenn wir von zweiwertigen Variablen (0 und 1) ausgehen, wobei 1 den Aktivierungszustand beschreibt:

$W = L + F - LF$
$F = S$
$L = B(S-1)$

Die Gleichungen beschreiben nun die kontrafaktischen Zusammenhänge in unserem System, wobei die Gleichungen nur von rechts nach links gelesen werden: Wäre S=1, so wäre auch F=1 gewesen. Oder: Wären L=1 und F=1, so wäre auch W=1 gewesen usf. Wie lassen sich dann größere Kausalzusammenhänge finden?

Eine wichtige Idee ist nun, dass die kontrafaktischen Abhängigkeiten auf einem bestimmten Pfad dann wieder sichtbar werden, wenn wir die Variablen außerhalb dieses Pfades auf ihren aktuellen Werten festhalten (vgl Hitchcock 2001, Hall 2007 und Halpern und Pearl 2005, Part I). In unserem Fall geht der Pfad von S über F nach W und wir dürfen L=0 festhalten. Die anderen Variablen erhalten alle den Wert 1. Setzen wir aber nun L=0 voraus, so gilt: Wäre S=0 gewesen, so wäre auch W=0 gewesen. Damit ist die Abhängigkeit von W von S wieder sichtbar geworden und damit S wieder als Ursache von W erkennbar. Diese Idee ist zunächst einmal recht intuitiv und scheint den richtigen Weg zu weisen.

Allerdings funktioniert sie bereits in Fällen einfacher Überdeterminiation nicht mehr, bei dem zwei Variablen A und B gleichzeitig aktiv

sind und gleichzeitig W aktivieren. W ist nicht kontrafaktisch abhängig von A, wenn wir den Wert von B festhalten. Nun könnte man noch denken, dass die Überdetermination eben ganz besondere Probleme aufwirft, das Verfahren aber womöglich in anderen Fällen zuverlässig die Ursachen bestimmt. Leider hat Ned Hall (2007) dazu auch andere komplexe Gegenbeispiele konstruieren können, die den ganzen Ansatz weiter unterminieren. Er selbst entwickelte zwar einen Verbesserungsvorschlag, um diese Gegenbeispiele wieder ausschalten zu können, aber Hitchcock (2009) konnte nachweisen, dass es dagegen wiederum entsprechend trickreiche Gegenbeispiele gibt, so dass auch der neue Ansatz nicht wirklich zufriedenstellen kann. Hier ist also bisher für komplexe indirekte Kausalbeziehungen kein perfektes Beschreibungs- und Entdeckungsverfahren zu finden, das wir nun einfach übernehmen könnten.

Die Beurteilung von Kausalbeziehungen ist letztlich ein holistisches Verfahren, mit dem wir komplettere Kausaltheorien für einen bestimmten Bereich miteinander vergleichen müssen. Ganz ohne kausale Grundannahmen wird uns dabei kein Verfahren zu weitergehenden Kausalhypothesen leiten. Kausalität ist vermutlich einer unserer Grundbegriffe oder zumindest ein sehr basaler Begriff, der nicht strikt auf basalere Begriffe reduziert werden kann. Wir können nur die Verbindungen zu anderen Begriffen explizieren, die aber z.T. selbst schon in der Nähe unserer Kausalbegriffe liegen, wie der der Einzelfallwahrscheinlichkeit oder Propensität. Trotzdem sind gerade die strukturellen Modelle (ergänzt um einen probabilistischen Fehlerterm) für die Praxis der Wissenschaften von besonderer Bedeutung. Sie ergeben sich etwa anhand von Regressionsverfahren oder aus anderen Betrachtungen von Korrelationen, und sie liefern uns zugleich aufschlussreiche Antworten darauf, was passieren würde, wenn wir auf bestimmte Weise in das System eingreifen würden (vgl. dazu Kap. 7.4).

7.3.8 Kausale Strukturen und gerichtete Graphen

Um komplexere kausale Strukturen darzustellen, bietet sich das Hilfsmittel der Kausalgraphen an (vgl. Kap. 5.3.8). Meist beschränkt man sich auf *gerichtete azyklische Graphen* G (sogenannte DAGs für »directed

acyclic graph«). Man nimmt also bereits an, dass die zu ermittelnde Kausalstruktur keine Rückkoppelungen kennt, sondern einfacher Art ist. Der Graph G = (V,E) besteht aus den Knoten (vertices) V = $\{X_1,\ldots,X_n\}$ und den gerichteten Kanten (edges) oder Pfeilen E ⊆ V×V, die jeweils durch ein Paar von Knoten gegeben sind. »X wirkt auf Y« wird im Graphen durch einen Pfeil X⇛Y dargestellt. Lässt man alle Ausrichtungen weg und arbeitet mit ungerichteten Kanten, erhält man das *Skelett* von G. Weiterhin gibt es viele anschauliche Konzepte auf einem solchen Graphen, die einem helfen. Ein *Pfad* von X nach Y ist ein Weg von X nach Y über möglicherweise weitere Knoten, der entlang der Kanten (mit oder gegen deren Richtung) läuft. Ein *gerichteter Pfad* läuft nur entlang der gerichteten Kanten in deren Richtung. In DAGs sind Zyklen, also gerichtete Pfade, die zum selben Knoten zurückführen verboten. Sind zwei Knoten durch einen Pfad verbunden, heißen sie *verbunden*.

Außerdem verwendet man in anschaulicher Weise die Begriffe für Verwandtschaftsbeziehungen: Gilt X⇛Y, so sagen wir, dass X Elter von Y ist bzw. Y Kind von X ist. Außerdem sprechen wir von Nachfahren und Vorfahren etc. Ein spezieller Graph ist der *vollständige Graph*, in dem zwischen je zwei Knoten eine ungerichtete Kante existiert.

Zusätzlich zur kausalen Struktur nehmen wir auch noch eine Wahrscheinlichkeitsverteilung $P(X_1,\ldots,X_n)$ auf der Menge V an. Das ist eine komplexe Funktion, die allen Kombinationen der Werte für die Variablen $X_1,\ldots,X_n$ eine Wahrscheinlichkeit zuweist. Hier droht wieder eine zahlenmäßige Explosion, aber für die beschriebenen DAGs lässt sich die Funktion deutlich vereinfachen. Wir nehmen dort an, dass die Wahrscheinlichkeitsverteilung für ein X_i jeweils nur von den Wahrscheinlichkeitsverteilungen der Eltern (PA für »parents«) von X_i abhängt. Man nennt solche Wahrscheinlichkeitsverteilungen markovsch relativ zu G. Das heißt, dass X_i statistisch unabhängig zu seinen Vorfahren (und anderen Nicht-Nachfahren) ist gegeben seine Eltern. Das ist gerade für kausale Beziehungen eine recht plausible Annahme.

Man kann das auch so beschreiben, dass wir durch die Kenntnis der Vorfahren keine neuen Informationen über X_i erhalten, wenn wir die Werte der Eltern bereits kennen. Dann vereinfacht sich die gemeinsame Verteilung der X_i erheblich. Zunächst gilt nach der Kettenregel für Wahrscheinlichkeiten für alle Werte der Variablen $X_1,\ldots,X_n$:

(KR) $P(x_1,\ldots,x_n) = \prod_j P(x_j|x_1,\ldots,x_{j-1})$

Dann können wir das weiter vereinfachen zu

Markov-Bedingung 1: $P(x_1,\ldots,x_n) = \prod_j P(x_j|pa_j)$,

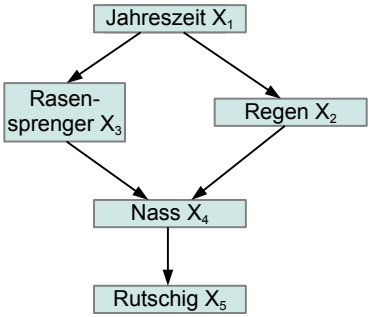
weil wir für die Wahrscheinlichkeiten von X_j nur die der Eltern heranziehen müssen (vgl. Kap. 5.3.8). So erhalten wir ein Bayessches Netz N = (G,P) (vgl. Neapolitan 2004, Williamson 2005). Das ist ein DAG, für das die Markov-Bedingung gilt.

Bayessches Netz N = (G,P)
N ist ein Bayessches Netz gdw. G ein gerichteter azyklischer Graph ist und P eine Wahrscheinlichkeitsverteilung auf seinen Knoten, die die Markov-Bedingung bezüglich G erfüllt.

Man kann die Markov-Bedingung für P auch äquivalent so schreiben, dass gilt:

Markov-Bedingung 2: Für alle X in V gilt: X ist statistisch unabhängig zu ND_X gegeben PA_X, was auch oft so geschrieben wird:
$(X \perp ND_X \mid PA_X)_P$

Das soll bedeuten, dass X statistisch unabhängig ist von allen Nicht-Nachfahren ND_X von X gegeben die Eltern PA_X von X (relativ zur Wahrscheinlichkeitsverteilung P). Das können wir uns nun an einem einfachen Beispiel von Pearl (2000, 15) verdeutlichen: Darin geht es darum, dass je nach Saison (Sommer, Herbst, Winter Frühling) sich andere Wahrscheinlichkeiten für Regen oder das Angehen des Rasensprengers einstellen und das jeweils dazu führt, dass ein Weg nass wird und deshalb rutschig wird oder auch nicht.



Graphik 7.4: Beispiel für ein kleines Bayessches Netz

Die Eltern X_2 und X_3 von X_4 schirmen hier X_4 von X_1 ab. Wenn wir also schon wissen, ob der Rasensprenger an ist und ob es regnet, dann wissen wir schon alles, was es gemäß diesem Netz darüber zu wissen gibt, ob der Weg nass wird. Jedenfalls gibt uns die zusätzliche Information, welche Jahreszeit X_1 wir haben, dann keine zusätzlichen Hinweise mehr darauf, ob der Weg nass wird, wenn unsere Wahrscheinlichkeitsverteilung für dieses kleine Netz markovsch ist, also zu der Kausalstruktur G passt.

Die Markov-Bedingung lässt sich für beliebige DAGs G und Wahrscheinlichkeitsverteilungen P formulieren und scheint für reale Kausalstrukturen sehr plausibel zu sein. Stellt G den Graphen zu einer realen Kausalstruktur dar und P die tatsächlichen Wahrscheinlichkeiten für das Auftreten von Werten der beteiligten Größen, die sich aus den kausalen Zusammenhängen ergeben, dann sollte P markovsch bzgl. G sein. Jedenfalls ist das der Fall, wenn die Menge V aller Variablen im Netz kausal vollständig ist, d.h., wenn es keine gemeinsamen Ursachen von Faktoren in V gibt, die nicht in V enthalten sind (sogenannte verborgene Common Causes), und auch keine vor uns verborgenen kausalen Pfade. Wir erwarten dann, dass N = (G,P) die kausale Markov-Bedingung erfüllt. Das ist der Grund, warum Bayessche Netze als geeignete formale Repräsentationen kausaler Zusammenhänge betrachtet werden.

Bayessche Netze sind vor allem durch ihre Unabhängigkeitsbeziehungen gekennzeichnet. Die gemeinsame Wahrscheinlichkeitsverteilung P der Faktoren aus V gibt uns hier an, welche Koinzidenzen wir zu erwarten haben und welche nicht auftreten werden. Sie ist der zumindest im Prinzip der Beobachtung zugängliche Teil des bayesschen Netzes N. Daher ist die Aufgabenstellung des kausalen Schließens vor allem, aus

den »Daten« P wieder auf die zugrundeliegende Kausalstruktur G zurück zu schließen. Insbesondere wird es darum gehen, aus den zu beobachtenden statistischen Unabhängigkeiten auf die zugrundeliegenden kausalen Unabhängigkeiten zu schließen und daraus den Graphen G zu rekonstruieren.

Anders als im Falle der deterministischen Kausalität versucht man hier also nicht direkt aus dem *gemeinsamen* Auftreten von Faktoren auf kausale Zusammenhänge zu schließen, sondern vielmehr aus dem *nicht gemeinsamen* Auftreten darauf zu schließen, dass auch keine Kausalbeziehungen vorliegen. Im Prinzip geht man zunächst davon aus, dass zwischen allen Variablen Kausalbeziehungen vorliegen könnten, und eliminiert dann nach und nach solche potentiellen Kausalbeziehungen, für die sich keine Koinzidenzen der beteiligten Faktoren finden lassen. Dazu werden aber nicht sogleich die umfangreichen kompletten Graphen betrachtet, sondern vielmehr Teilstrukturen von jeweils drei Faktoren.

Um das Verfahren umsetzen zu können, müssen wir zunächst drei *kausale Grundstrukturen* unterscheiden, die uns schließlich auch beim (probabilistischen) kausalen Schließen immer wieder als Teilstrukturen begegnen werden. Da haben wir zunächst die *kausale Kette*: X⇛Y⇛Z. In unserem Beispiel könnte das etwa sein: Regen⇛Nass⇛Rutschig. Für solche Verhältnisse gilt normalerweise: $(X \perp Z \mid Y)$. Wissen wir also schon, dass der Weg nass oder trocken ist, erfahren wir nichts Neues über den Rutschigkeitszustand des Weges, wenn wir etwas darüber erfahren, ob es geregnet hat. Das ist jedenfalls dann der Fall, wenn es hier nicht einen geheimnisvollen Einfluss von Regen auf die Rutschigkeit gibt, der nicht über den Nässezustand des Wegs vermittelt wird.

Ein Problem für das kausale Schließen ist nun, dass auch die *Common-Cause-Struktur* oder *(divergierende) Gabel* X⇚Y⇛Z dieselbe Unabhängigkeitsstruktur $(X \perp Z \mid Y)$ aufweist. Das hatten wir bereits in anderen Beispielen gesehen, findet sich aber natürlich auch in unserem Nässebeispiel wieder. Die Jahreszeit war dort eine solche gemeinsame Ursache von Regen und Rasensprenger, wobei sie hier eine negative Korrelation stiftet. Weiß ich schon, dass Herbst ist, hilft mir die Information, dass der Rasensprenger läuft, nicht weiter für die Frage, ob es regnet. Die Wahrscheinlichkeit für Regen bestimmt sich hier ganz aus

der Jahreszeit und ist davon unabhängig, ob der Rasensprenger an ist. Das gilt in unserem Beispiel auch umgekehrt, obwohl es nicht sehr plausibel zu sein scheint, dass jemand den Sprenger einschaltet, wenn es regnet. Das würde aber eine neue Kausalverbindung etablieren, die in unserer Geschichte nicht vorkommt. Der Sprenger schaltet sich in unserem Beispiel einfach nur nach Jahreszeit ein und nicht anhand der Entscheidung einer Person, die erst einmal nachschaut, ob es regnet. Das wäre eine andere Kausalstruktur, für die $(X \perp Z \mid Y)$ nicht mehr gelten würde.

Als dritte Struktur müssen wir uns noch der *konvergierenden Gabel* (dem »*collider*«) bzw. der *v-Struktur* $X \Rightarrow Y \Leftarrow Z$ zuwenden. Sie hat nun allerdings eine andere Abhängigkeitsstruktur als ihre Vorgänger, was sie so bedeutsam für das kausale Schließen macht. Unser Collider in unserem Beispiel ist: Sprenger$\Rightarrow$Nass$\Leftarrow$Regen. Wenn wir nur diese Struktur betrachten, so sind hier die äußeren Variablen zunächst als statistisch unabhängig zu betrachten: $(X \perp Z)$, ohne, dass wir dafür eine Bedingung benötigen. Im Gegenteil findet hier erst ein Informationsfluss von X zu Y statt, wenn Y bekannt ist. Wissen wir nämlich schon, dass es nass ist, dann können wir daraus, dass es nicht regnet, in unserem Netz schließen, dass der Sprenger an sein muss. Das heißt, die Kenntnis von Y schaltet hier den Informationsfluss von X nach Z frei, während diese Kenntnis in den beiden vorhergehenden Fällen den Informationsfluss blockiert hat. Man nennt in einem gerichteten Graphen die Knoten *d-separiert* (»d« steht für »directed«), zwischen denen kein kausaler Informationsfluss stattfindet. In unserer dritten Grundstruktur sind X und Z ohne weiteres d-separiert, während in den ersten beiden Strukturen der Weg von X nach Z durch Y blockiert werden muss. Erst wenn Y dort bekannt ist, liefert X keine weiteren Informationen mehr über Z und umgekehrt. Hier gelten X und Z dann durch Y als d-separiert.

Anhand dieser Regeln können wir aus einem Kausalgraphen die statistischen Unabhängigkeiten der gemeinsamen Verteilung P schnell herauslesen. Für alle Variablen, die in dem Kausalgraphen *d-separiert* durch bestimmte andere Variablen sind, gilt, dass sie dann auch statistisch unabhängig sind (gegeben diese anderen Variablen). Das ist nur eine andere Darstellung der Markov-Bedingung. Dazu können wir

zunächst etwas allgemeiner definieren, was es heißt, dass zwei Variablen durch andere Variablen d-separiert sind.

d-Separation: Zwei Variablen X und Y sind dann *d-separiert* durch eine Menge von Variablen ***Z***, wenn jeder Pfad p von X nach Y blockiert wird durch ***Z***, d.h., wenn gilt:

1) Für zumindest einen Collider X⇛M⇚Y auf dem Pfad p gilt, dass M$\not\in$***Z*** ist und auch kein Nachkomme von M in Z liegt. (So ein M bzw. ein Nachkomme von M würde den Collider freischalten, wie unser Beispiel zeigt.) oder

2) Für zumindest eine der anderen Strukturen auf p vom Typ X⇛M⇛Y oder vom Typ X⇚M⇛Y gilt: M$\in$***Z***.

Das heißt, dass auf jedem Pfad zwischen X und Y eine Blockade stattfindet, entweder weil zumindest ein Collider nicht durch Z freigeschaltet ist oder weil eine der anderen Verbindungen durch Z blockiert wurde. Hier finden wir das Theorem, dass wenn X und Y durch Z d-separiert sind, dann in einem bayesschen Netz gilt, dass X und Y statistisch unabhängig sind gegeben Z. Diese Definitionen und Zusammenhänge lassen sich auf Mengen X und Y von Variablen in entsprechender Weise ausdehnen.

7.3.9 Graphentreue Bayessche Netze

Man könnte also sagen, wenn es einen nicht blockierten Pfad zwischen X und Y gibt, dann findet ein kausaler Informationsfluss zwischen X und Y statt, während durch Z d-separierte X und Y keinen kausalen Informationsfluss mehr aufweisen, wenn Z gegeben ist. In einem Bayesschen Netz mit Wahrscheinlichkeitsverteilung P findet sich diese Struktur in den statistischen Unabhängigkeitsbeziehungen wieder (das besagt letztlich die Markov Bedingung), indem hier gilt: $(X \perp Y \mid Z)_P$. Das lässt sich auch kurz für alle Mengen X, Y und Z von Zufallsvariablen aus V so schreiben:

Strukturähnlichkeit 1: (Markov-Bedingung)
$(X \perp Y \mid Z)_G$ daraus folgt: $(X \perp Y \mid Z)_P$,

wobei $(X \perp Y \mid Z)_G$ bedeuten soll, dass Z gerade X von Y d-separiert in G. Man könnte das so ausdrücken: *Den kausalen Strukturunabhängigkeiten entsprechen bestimmte strukturelle statistische Unabhängigkeiten.*

Allerdings könnte es noch weitere nicht strukturelle Unabhängigkeitsbeziehungen in P geben, also probabilistische Unabhängigkeiten, die nicht durch die Struktur von G bedingt sind (nicht notwendig sind). Da wir letztlich anhand der statistischen Unabhängigkeiten, die wir aus den beobachtbaren Daten erschließen können, zurück auf die zugrundeliegende kausale Struktur schließen möchten, könnten uns diese nicht-strukturellen Unabhängigkeiten in die Irre führen. Deshalb verlangt man meist zusätzlich eine zweite Strukturähnlichkeit zwischen G und P:

Strukturähnlichkeit 2: (Graphentreue)
$(X \perp Y \mid Z)_P$ daraus folgt: $(X \perp Y \mid Z)_G$

Das heißt, es gibt nur noch die statistischen relativen Unabhängigkeiten, die tatsächlich durch die kausale Struktur repräsentiert in G vorgegeben sind, die man als strukturelle statistische Unabhängigkeiten bezeichnen kann. Diese Annahme der *Graphentreue* (oder bei Pearl »Stabilität«) der Verteilung P gegenüber dem Graphen G ist offensichtlich hilfreich und wesentlich für das Schließen auf den zugrundeliegenden Graphen G, denn schließlich sollen unsere kausalen Schlüsse genau in dieser Richtung laufen.

Die Idee für unser Schließen ist damit die folgende: Wir nehmen an, dass die wahre kausale Struktur in einem bestimmten Anwendungsbereich durch ein DAG G = (V,E) ausgedrückt werden kann. Davon kennen wir im Wesentlichen schon die (möglichen) Knoten V, wissen aber zunächst noch nicht, wie die Struktur E auf V beschaffen ist. Die versuchen wir zu ermitteln, indem wir die zwei obigen Annahmen der Strukturähnlichkeit voraussetzen und mit deren Hilfe aus den vorliegenden Daten (hier müssen wir annehmen, dass wir alle relativen statistischen Unabhängigkeiten ermitteln können, also alle relevanten möglichen Situationen auch tatsächlich aufgetreten sind), dann die Struktur E (also die Pfeile im Graphen) erschließen können. Man sieht hier wiederum, wie schwierig das Schließen auf kausale Zusammenhänge

ist und wie starke Annahmen wir bereits voraussetzen müssen, um überhaupt zum Schluss zu kommen. Trotzdem sind wir hier noch immer nicht am Ziel.

Doch bevor wir uns überlegen, was wir mit dem Verfahren im Idealfall erreichen können, sollen die Annahmen noch einmal kurz erörtert werden. Zunächst die *Markovbedingung*. Sie impliziert u.a., dass die Wurzelknoten untereinander statistisch unabhängig sind. Ein Wurzelknoten X soll schließlich von seinen Nicht-Nachkommen (also auch einem anderen Wurzelknoten Y) statistisch unabhängig sein, gegeben seine Eltern. Da er keine Eltern hat, müssen schlicht X und Y statistisch unabhängig voneinander sein. Das heißt, genau genommen, dass für die Annahme der Menge V keine *gemeinsamen Ursachen von Wurzelknoten* und damit auch von anderen Knoten im Netz übersehen wurden. Das setzt bereits ein gewisses kausales Hintergrundwissen voraus. Wir nehmen damit an, dass wir bereits über eine vollständige Liste aller potentiell wirksamen Faktoren in einem bestimmten Bereich verfügen.

Das entspricht in etwa den Forderungen nach einer vollständigen Liste von Hypothesen für die früheren induktiven Schlussformen, denn im Prinzip liefert uns die Menge aller Kombinationen von Faktoren aus V eine solche Liste von Hypothesen. Die zwei Hypothesen zu zwei Faktoren V und W sind eben gerade, dass (1) V kausal relevant für W ist und dass (2) W kausal relevant für W ist. Wir haben schon früher darauf hingewiesen, dass diese Annahme recht anspruchsvoll ist, hatten aber auch gesehen, dass sie für viele induktive Schlussverfahren unverzichtbar ist. Im konkreten Anwendungsfall ist sie auch häufiger plausibel einzulösen. Das belegen einige der früheren Beispiele.

Zugleich wird deutlich, dass hier bestimmte Unterminierer für das kausale Schließen anhand bayesscher Netze zu finden sind. Außerdem besagt die Markov-Bedingung noch, dass kausale Wirkungen von X auf Y, die nur über bestimmte Zwischenfaktoren Z vermittelt werden, dazu führen, dass z.B. die Großeltern im Netz von den Kindern probabilistisch unabhängig sind gegeben die Eltern, und das wird durch unsere Vorstellungen von Kausalität nahegelegt. Die Darstellung durch einen gerichteten azyklischen Graphen schließt aber noch nicht aus, dass die Großeltern daneben auch direkt auf die Kinder einwirken und damit

zugleich weitere Eltern der Kinder sind. Das muss dann nur durch einen weiteren Pfad gekennzeichnet werden.

Die zweite Strukturähnlichkeitsbedingung der *Graphentreue* wird manchmal für harmlos gehalten. Doch für deterministische Systeme ist sie nicht so ohne. Was muss für (probabilistische) Systeme passieren, damit sie verletzt wird? X könnte auf zwei Wegen über M und N zu Y führen, wobei die Wirkungen sich gerade aufheben. Ein Beispiel, das sich in ähnlicher Weise bei Nancy Cartwright (1989) findet, ist das folgende (vgl. a. Baumgartner & Grasshoff 2004). Nehmen wir an, X bezeichne die Einnahme der Antibabypille und M die Erhöhung des Östrogenspiegels. Das wiederum erhöhe das Thromboserisiko (Y). Andererseits reduziert die Einnahme der Pille die Anzahl der Schwangerschaften (N) und diese würden ihrerseits das Thromboserisiko erhöhen. Also könnte es der Fall sein, dass die Wirkungen der Einnahme der Pille für das Thromboserisiko für Frauen sich auf beiden Wegen gerade aufheben. Die Einnahme der Pille würde also insgesamt das Thromboserisiko nicht verändern.

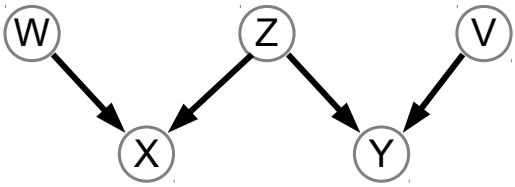
Damit wären X und Y *statistisch unabhängig*, obwohl X auf zwei Wegen *kausal relevant* für Y ist, was sich natürlich anhand geeigneter Eingriffe (etwa einer Verhütung ohne Pille) auch aufzeigen lässt. Allerdings verlangt diese Konstellation nach ganz besonderen Bedingungen für die zwei kausalen Wege. Die Risikoerhöhung und die Risikoverminderung auf den beiden Wegen müssen genau gleich groß sein. Daher argumentieren die Befürworter einer Annahme der Graphentreue, dass bei geeigneter Parametrisierung der speziellen funktionalen Zusammenhänge sich derartige Konstellationen nur für eine Nullmenge der Bedingungen finden (Spirtes & Glymour & Scheines, 41: Theorem 3.2) und wir sie daher als Ausnahmefälle beiseitelassen dürfen.

Das sieht allerdings anders aus für *deterministische Zusammenhänge* (Spirtes & Glymour & Scheines 2000, 53 f.). Betrachten wir den Graphen, in dem gilt: $A \Rrightarrow B \Rrightarrow C$ und zusätzlich direkt $A \Rrightarrow C$. Dann gibt es gemäß der d-Separation keine Unabhängigkeitsbeziehungen darin. Nehmen wir aber an, dass B den Wert von C determiniert (wir haben es schließlich mit deterministischer Kausalität zu tun), so gilt: $(A \perp C|B)_P$. Also gibt es statistische Unabhängigkeitsbeziehungen, die aber keine kausalen Unabhängigkeitsbeziehungen darstellen, denn wir haben schließlich angenommen, dass A auch direkt kausal relevant für C ist. Unsere

Kenntnis von B schneidet also den *kausalen* Einfluss von A auf C keineswegs ab. Die Graphentreue ist damit verletzt. Oder es determiniere A direkt C, dann gilt: $(B \perp C|A)_P$, obwohl auch B weiterhin kausal relevant für C ist.

In deterministischen Strukturen treten also sehr schnell nicht-graphentreue Unabhängigkeitsbeziehungen auf, während das für rein probabilistische Beziehungen zumindest als wenig wahrscheinlich gelten kann. In dem Fall könnten wir sagen, die Natur führt uns in die Irre, indem sie die unterschiedlichen Effekte der Pille auf das Thromboserisiko so gestaltet hat, dass keine Korrelation zwischen beidem auftritt, obwohl kausale Beziehungen vorliegen. Für das kausale Schließen (wie auch für andere induktive Schlüsse) sind wir aber auf ein gewisses Entgegenkommen der Natur angewiesen. Verbirgt sie ihre kausale Struktur zu geschickt, werden wir sie nicht korrekt ermitteln. Das ist das Induktionsrisiko, das wir beim induktiven Schließen nie ganz ausschalten können, und es liefert uns wiederum typische Unterminierer für die entsprechenden Kausalschlüsse. Noch problematischer sind diese Fälle natürlich für entsprechende Definitionsversuche der Kausalbeziehung, die sich dadurch als falsch erweisen.

Nehmen wir die Graphentreue jedoch einmal als gegeben an, so erhalten wir in folgendem Beispiel nur die der Graphenstruktur entsprechenden Unabhängigkeitsbeziehungen (vgl. Spirtes & Glymour & Scheines 2000, 43 f.).



Graphik 7.5: ein einfaches Beispielnetz

Für den Graphen (CCN) $W \Rightarrow X \Leftarrow Z \Rightarrow Y \Leftarrow V$ ergeben sich damit die Unabhängigkeitsbeziehungen:

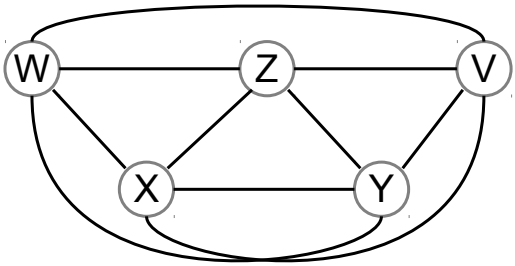
(1) $(W \perp \{Z,Y,V\})$
(2) $(X \perp \{Y,V\}|\{W,Z\})$

(3) (Z⊥{W,V})
(4) (Y⊥{W,X}|{V,Z})
(5) (V⊥{W,X,Z})

Dabei sind nun allerdings einige Unabhängigkeitsbeziehungen doppelt aufgeführt, da sie symmetrisch sind in den ersten beiden Argumenten.

7.3.10 Erschließen des Kausalgraphen

Im Prinzip können wir nun in der anderen Richtung fragen, welche kausale Struktur diese Unabhängigkeitsbeziehungen am besten erklären kann. Das wäre dann unser bester Tipp für unsere kausale Theorie dieses Gebiets. Unter Annahme der beiden Strukturzusammenhänge haben sich bestimmte Algorithmen für diesen Rückschluss etabliert, von denen ich kurz den SGS-Algorithmus (Spirtes & Glymour & Scheines 2000, 82) sowie vor allem den *IC-Algorithmus* (für »inductive causation«) von Pearl (2000, 50 ff.) betrachten möchte. Beide Verfahren gehen von einem *vollständigen ungerichteten Graphen* aus, bei dem jeder Knoten $V = \{X_1,\ldots,X_n\}$ mit jedem anderen verbunden ist und nun anhand von Unabhängigkeitsbeziehungen bestimmte Verbindungen gestrichen werden und zuletzt möglichst viele Verbindungen ausgerichtet werden.



Graphik 7.6: ein vollständiger Graph

Dieser vollständige und ungerichtete Graph entspricht hier einer vollständigen Hypothesenliste mit den Hypothesen H_{ij}:$X_i \Rightarrow X_j$ für alle ungleichen i und j, wobei der Pfeil für eine direkte Verursachungsbeziehung steht. Im Verfahren versuchen wir dann möglichst viele der Hypothesen zu eliminieren anhand der beobachteten statistischen Unabhängigkeitsbeziehungen, bis nur noch eine kompatible Hypothesenmenge übrigbleibt, die

uns dann nach Möglichkeit einen definitiven Kausalgraphen beschreibt. Allerdings werden wir hier wieder auf das Problem stoßen, dass nicht unbedingt alle Konkurrenzhypothesen bis auf eine eliminiert werden können und wir so mit einer gewissen Unterbestimmtheit zurückbleiben. Wir starten mit dem vollständigen ungerichteten Graphen H und fangen zunächst an, Kanten zu streichen und versuchen danach, von den verbleibenden Kanten so viele wie möglich zu orientieren:

SGS-Algorithmus

1. Für jedes Paar A, B von Variablen, für die es eine Menge von Variablen $S_{AB} \subseteq V\backslash\{A,B\}$ gibt, für die $(A \perp B|S_{AB})_P$ gilt, entferne man die Kante zwischen A und B. (Nur direkte Kausalbeziehungen werden im Graphen durch Pfeile repräsentiert.)

2. Für jedes Tripel A—C—B, bei dem A und B nicht verbunden sind, orientiere man A⇒C⇐B gdw. es *keine* Menge $S_{AB} \subseteq V\backslash\{A,B\}$ gibt, für die $(A \perp B|S_{AB})_P$ gilt. (Hier werden die v-Strukturen unseres gesuchten Graphen identifiziert.)

3. Wenn gilt A⇒C—B, aber A und B sind nicht verbunden, und es gibt keine Pfeilspitze, die auf B zeigt, so orientiere C⇒B. (Hier wird orientiert unter der Vermeidung weiterer v-Strukturen.)
Wenn es einen gerichteten Pfad von A nach B gibt und zusätzlich A und B direkt verbunden sind, so orientiere: A⇒B. (Hier wird orientiert unter der Vermeidung von Zyklen.)

Ganz ähnlich funktioniert der IC-Algorithmus. Er besteht aus den folgenden Schritten, wobei wir wie gesagt mit einem vollständigen ungerichteten Graphen starten und uns im ersten Schritt einer vereinfachten Beschreibung im Sinne des SGS-Algorithmus von Spirtes und Glymour und Scheines (2000, 82) bedienen:

IC-Algorithmus

1. Für jedes Paar A, B von Variablen, für die es eine Menge von Variablen $S_{AB} \subseteq V\backslash\{A,B\}$ gibt, für die $(A \perp B|S_{AB})_P$ gilt, entferne man die Kante zwischen A und B.

2. Suche alle Tripel A, B und C, bei denen A und B nur über C miteinander verbunden sind: A—C—B. Falls dann C nicht in der

Menge S_{AB} [mit $(A \perp B|S_{AB})_P$] enthalten ist, so orientiere das Tripel als v-Struktur: $A \Rightarrow C \Leftarrow B$.

3. In dem nun erhaltenen partiell gerichteten Graphen sollen so viele Kanten wie möglich ausgerichtet werden unter Einhaltung der zwei folgenden Bedingungen: (i) Produziere keinen neuen v-Strukturen und (ii) produziere keine gerichteten Zyklen.

Diese Regeln sind jetzt eigentlich relativ leicht verständlich. Die erste Regel nutzt die angenommene Graphentreue aus: Wenn A und B statistisch unabhängig sind gegeben S_{AB}, dann gibt es auch keine direkte kausale Verbindung zwischen A und B mehr, sondern nur noch die indirekten kausalen Verbindungen über die Variablen in S_{AB}. A ist eventuell Ursache von B über Zwischenfaktoren in S_{AB} oder es gibt gemeinsame Ursachen von A und B in S_{AB}. In jedem Fall liegt jedenfalls keine direkte Kausalbeziehung zwischen A und B mehr vor. Die entsprechende Kante kann also eliminiert werden, was einer Eliminierung der entsprechende Hypothese entspricht.

Im zweiten Schritt suchen wir nach verbundenen Tripeln A—C—B, für die B aber weder Zwischenursache noch gemeinsame Ursache von A und B ist. Dann bleibt nur noch die v-Struktur übrig. Damit nutzen wir die besondere Stellung der v-Struktur unter den drei Grundstrukturen aus. Sie wird durch eine spezielle probabilistische Unabhängigkeitsstruktur gegenüber den anderen beiden kausalen Grundstrukturen hervorgehoben.

Dann haben wir aber bereits alle v-Strukturen erwischt und müssen im dritten Schritt auf andere Strukturen hinaus, die jedoch weiter auf einen DAG hinführen sollen. Wir wissen nun schon, dass die weiteren Strukturen keine v-Strukturen sind (denn die hätten sich schon in Schritt 2 gezeigt), und wir nutzen unsere Grundannahme aus, dass wir es nur mit Fällen gerichteter Kausalität zu tun haben (und keine Zyklen auftreten), um die verbliebenen Kanten auszurichten. Das Ausrichten von Kanten entspricht wiederum der Elimination von einer der jeweils zwei verbliebenen Kausalhypothesen zwischen zwei Knoten.

Um das Verfahren im dritten Schritt noch zu algorithmisieren, gibt es weitere Regeln (s. etwa Pearl 2000, 51), die das leisten sollen. Sie sollen den dritten Schritt wie folgt systematisieren:

Regel 1: Ist A⇛B—C und A ist nicht direkt mit C verbunden (A und C sind nicht benachbart), so orientiere: A⇛B⇛C, da sonst eine v-Struktur auftreten würde.

Regel 2: Falls A⇛B⇛C und A—C gilt, so orientiere: A⇛C, da sonst ein Zyklus auftreten würde.

Regel 3: Falls man zwei Ketten findet: A—C⇛B und A—D⇛B, aber C und D sind nicht verbunden, so orientiere die Kante A—B zu A⇛B, da jede mögliche andere Ausrichtung sonst entweder einen Zyklus oder eine v-Struktur erzeugen würde.

Regel 4: Falls man zwei Ketten findet: A—C⇛D und C⇛D⇛B, aber C und B sind nicht verbunden, so orientiere die Kante A—B zu A⇛B, da jede mögliche andere Ausrichtung sonst entweder einen Zyklus oder eine v-Struktur erzeugen würde.

Die letzten beiden Fälle sind etwas komplexer, aber man kann sich anhand eines kleinen Beispielgraphen schnell davon überzeugen, dass es keine anderen Möglichkeiten gibt, als die gerade genannten.

Das Verfahren können wir nun auf unser kleines Beispielnetz CCN anwenden. Wir starten also mit dem vollständigen ungerichteten Graphen und fangen dann zunächst an zu streichen.

Gemäß den Unabhängigkeitsbeziehungen in (1) fallen alle direkten Verbindungen von nach Z, Y und V weg, so dass Z nur noch (ungerichtet) mit X verbunden ist. Analoges übernimmt (5) für V. (2) kappt die direkte Verbindung zwischen X und Y, da diese durch den Common Cause Z voneinander abgeschirmt werden. Weitere Verbindungen können nicht gestrichen werden. Nun geht es um die Ausrichtung der verbliebenen Kanten, die schon die Grundstruktur unseres angestrebten Netzes zeigen. Hier können wir zwei v-Strukturen erkennen, da die Variablen X und Y nicht unter den »Blockierern« in (1)–(5) zu finden sind. Also entsteht wieder unser ursprüngliches Netz. Das klappt allerdings nur so gut, weil wir gleich zwei v-Strukturen in unserem Netz vorfinden.

Im Allgemeinen können wir nicht mit eindeutigen Ergebnissen rechnen. Das besagt zumindest ein bekanntes Theorem dieses Ansatzes, das wir etwa bei Pearl (2000, 19: Theorem 1.2.8 *Observational Equivalence*) finden, wonach zwei Graphen G_1 und G_2 dann dieselben Unabhängigkeitsstrukturen generieren, wenn sie dasselbe Skelett aufweisen (gleich

viele Knoten und dieselben Nachbarschaftskanten) und über dieselben v-Strukturen verfügen. Sobald also die ungerichtete Grundstruktur gleich ist und dieselben v-Strukturen vorliegen, erzeugen zwei Graphen dieselben probabilistischen Unabhängigkeitsmuster und sind mit diesem Verfahren nicht mehr zu unterscheiden.

Damit lassen sich leicht etliche beobachtungsäquivalente Graphen konstruieren. Einfache Beispiele hatten wir schon kennengelernt: A⇛B⇛C, A⇚B⇛C und A⇚B⇚C sind demnach beobachtungsäquivalent, aber auch z.B. die umfassenderen Graphen: D⇛A⇛B⇛C⇛E und D⇚A⇚B⇛C⇛E und viele ähnliche Strukturen, die keine unterschiedlichen v-Strukturen (in unserem Fall z.B. überhaupt keine v-Strukturen) enthalten. Viele formale Resultate über solche markoväquivalenten Graphen finden sich in Mayo-Wilson (2014). Welche Bedueutng die für die praktische Anwendbarkeit dieser Verfahren haben, ist aber nicht so leicht zu klären.

Es ist sicher zu stark, in diesen Fällen gleich von *Beobachtungsäquivalenz* zu sprechen, denn wir können durch Beobachtung auch zu anderen Informationen etwa über die zeitliche Abfolge der Instanzen unserer Variablen gelangen, die selbst wieder zu bestimmten Eliminationen von Kausalvermutungen führen können. Auch sagt uns unser Hintergrundwissen, dass etwa eine Variable wie das Geschlecht nicht verursacht wird durch die Variable Rauchverhalten. Wir können unser Geschlecht nicht dadurch verändern, dass wir mehr oder weniger rauchen. Das heißt aber, dass auch bei *Markov-Äquivalenz* (wie die Beziehung manchmal genannt wird) durchaus noch empirisch zugängliche Informationen darüber vorliegen können, welche der möglichen Kausalgraphen, die alle unsere Unabhängigkeitsbeziehungen hervorgebracht haben können, unser bester Tipp ist.

Die Mehrdeutigkeit oder *Unterbestimmtheit der Theorie* (bzw. unseres vermuteten Kausalgraphen G) wird allerdings noch größer, wenn wir die Voraussetzungen abschwächen und z.B. noch verborgene Common Causes zulassen, die sich dann nicht in unserer Liste V von Faktoren befinden. Dazu gibt es wiederum bestimmte Verfahren wie den IC*-Algorithmus, die ich aber hier nicht verfolgen werde. Ein Weg, um die Unterbestimmtheit zu reduzieren, sind bestimmte Minimalitätsprinzipien, wonach wir die einfacheren latenten Strukturen unter den passenden

auswählen. Dieser Ansatz zeigt wiederum klare Zusammenhänge zur Abduktion, da die einfacheren Theorien ceteris paribus die besseren Erklärungen liefern sollen.

Ein Modell, das wir uns aber meist gerade nicht wünschen, ist etwa immer anwendbar, nämlich das, bei dem eine latente Variable Z (also eine Variable, deren Werte gerade nicht beobachtbar sind und die uns bisher nicht bekannt ist) alle anderen Variablen determiniert. Das droht etwa bei der *Gotteserklärung*, bei der Gott als Instanz ins Spiel gebracht wird, die alle Ereignisse in der Welt festlegt und damit *erklärt*(?). Das können wir mit dem genannten Verfahren nicht ausschließen, entspricht aber nicht unserer bisherigen Vorstellung von unserer Welt, in der lokale Ereignisse x an einer bestimmten Stelle andere lokale Ereignisse y in der Umgebung von x hervorrufen und ganz bestimmte Eigenschaften Ursachen ganz bestimmter anderer Eigenschaften sind, aber nicht gleich aller anderen Eigenschaften. Gewisse Vorstellungen von Lokalität gehen im Rahmen der »Gotteserklärung« also verloren. Die gestalten aber normalerweise Kausalbeziehungen erst so spannend für uns, weil wir nur dadurch wissen, wie wir lokal eingreifen können. Wir werden so in vielen Fällen über Forderungen nach möglichst großer Erklärungsstärke noch eine weitere Auswahl unter den markov-äquivalenten Strukturen treffen können. Der IC-Algorithmus nutzt demnach keineswegs bereits alle zugänglichen Informationen aus.

Sind die so beschriebenen Verfahren abduktiver Natur? Sie zeigen zumindest die typischen Merkmale abduktiven Schließens. Ausgangspunkt sind bestimmte Ereignisse E = (W=w), für die wir nach einer Erklärung suchen. Die unterschiedlichen Faktoren $X_1,\ldots,X_n$ und ihre Kombinationen bieten die unterschiedlichen Erklärungshypothesen als mögliche Ursachen des Ereignistyps E. Die Koinzidenztabellen bzw. statistischen Abhängigkeiten sind unsere weiteren Daten, die nun mithelfen sollen, bestimmte Hypothesen zu *eliminieren*. Jeder Schritt von IC kann als eine solche Elimination betrachtet werden. Die auftretenden Koinzidenzen stellen Erklärungsanomalien für bestimmte Hypothesen dar und werden deshalb ausgeschlossen.

Wenn wir dabei sorgfältig vorgehen und so Gründe dafür haben, dass wir keine möglichen gemeinsamen Ursachen übersehen haben, so haben wir auch prima facie Gründe dafür, anzunehmen, dass es sich

um Wissen handelt. Gibt es dann tatsächlich keine Unterminierer – also etwa vergessene Common Causes –, so sollte es sich tatsächlich um Wissen handeln.

7.3.11 Kausale Vorhersagen

Mit Hilfe eines kausalen Modells lässt sich anschließend bestimmen, welche Wirkung von bestimmten Variablen auf andere ausgeht und etwa vorhersagen, was bei bestimmten Eingriffen in ein System passieren würde. Das lässt sich aber nicht einfach mit den im Prinzip beobachtbaren bedingten Durchschnittswahrscheinlichkeiten identifizieren.

Nehmen wir an, wir möchten ermitteln, welchen Einfluss fettes Essen in einem konkreten Fall auf die Wahrscheinlichkeit für einen Herzinfarkt hat. Möchten wir dazu etwa herausfinden, wie hoch die Wahrscheinlichkeit für einen Sportler wie Franz (S) sein würde, wenn er ab heute immer fett essen (X) würde, einen Herzinfarkt Y zu bekommen, so dürfen nicht einfach rechnen: P(X|do(Y=fett)) = P(X|(Y=fett)). (Der do-Operator beschreibt dabei ein Intervention, die die Variable auf einen bestimmten Wert setzt, vgl. Kap. 7.3.7.) Wir können ihm nicht einfach den Wert zuschreiben, den die durchschnittlichen Fettesser haben, denn bei denen sind normalerweise weitere relevante Variablen im Spiel, die sich von denen unseres Sportlers unterscheiden, wie z.B. die Bewegungsarmut und andere herzschädigende Lebensumstände. Ist also der durchschnittliche Fettesser viel träger als Franz, so dürfte die zweite Wahrscheinlichkeit deutlich größer sein, als die für Franz, einen Herzinfarkt zu erleiden, wenn er (oder ein anderer Sportler wie er) nun auf fette Kost umsteigt. Die Auswirkungen einer Intervention lassen sich nicht ganz so einfach abschätzen. Wir müssen uns genauer überlegen, welche anderen Faktoren festgehalten werden bei der Intervention und welche sich ändern dürfen oder sogar müssen (weil sie etwa von Y beeinflusst werden).

Pearl u.a. haben Regeln dafür aufgestellt, wie sie sich die neuen Wahrscheinlichkeitsverteilungen nach Interventionen aus »beobachtbaren« Wahrscheinlichkeiten ermitteln lassen. Dazu müssen wir die grundlegende Kausalstruktur kennen. In unserem Beispiel haben wir eine Dreiecksstruktur bestehend aus einem Common Cause S von X und

Y und zusätzlich einem Zusammenhang X⇛Y. Auf die Berechnungsweisen kann ich hier nicht umfassend eingehen, aber einige Beispiele für derartige Kausalschlüsse mögen genügen:

(1) Gilt einfach nur X⇛Y, so ist P(x|do y) = P(x|y).
(2) Ist aber X⇚Z⇛Y gegeben, dann gilt: P(y|do x) = P(y), da X keine direkten Auswirkungen auf Y hat.

Ist aber 3. wie in unserem Beispiel sowohl X⇚S⇛Y, als auch X⇛Y gegeben, so gilt:

(DD) $P(y|do\ x) = \sum_s P(y|x,s)\ P(s)$,

d.h., die Wahrscheinlichkeitsverteilung für Y nach der Intervention ergibt sich in Abhängigkeit von der weiteren Variablen S, für weitere Lebensumstände wie Bewegungsarmut etc. Wenn wir also jemanden auf fettes Essen setzen würden (per Intervention), so ergäbe sich die Wahrscheinlichkeit für einen Herzinfarkt als Durchschnittswert für alle davon Betroffenen in den unterschiedlichen Lebensumständen s. Möchten wir die Wahrscheinlichkeit speziell von Franz möglichst genau ermitteln, sollten wir natürlich nicht mit dieser Durchschnittsverteilung arbeiten, sondern nur den für Franz spezifischen Wert s einsetzen: P(Franz erleidet Herzinfarkt|fettes Essen) = P(Y=1|X=fettes Essen & S=sportlich). Möchten wir dagegen den durchschnittlichen Schaden durch fettes Essen ermitteln, interessiert uns gerade die Wahrscheinlichkeitsverteilung in der Gesamtpopulation und wir könnten mit den entsprechenden Erwartungswerten arbeiten (s.o.):

E(Y|do(X=fettes Essen)) − E(Y|do(X=mageres Essen)).

Leider sind Interventionen nicht immer so einfach durchführbar. Meist wird das in Experimenten durch eine geeignete Randomisierung versucht. Wir hatten aber schon gesehen, dass sie nicht immer funktioniert, weil es sich nur um einen Zufallsprozess handelt. Es gibt aber noch andere Probleme.

Ein Beispiel von Howson & Urbach (1993) beschreibt den folgenden Versuch. Nehmen wir an, wir möchten ermitteln, ob Tabakrauchen zu

Lungenkrebs führt, und verteilen unsere Probanden zufällig auf die Experimentalgruppe, die nun stark rauchen muss, und die Kontrollgruppe, die nun nikotinabstinent bleiben muss. Stellen wir uns vor, es ergeben sich signifikante Unterschiede zwischen den Gruppen. Trotzdem es muss nicht der Tabak sein, sondern es könnten auch bestimmte Bestandteile im Papier der Zigaretten sein, die den Krebs auslösen. Das kann man auch durch eine Randomisierung nicht ausschließen. Wir müssten zunächst überhaupt auf die Idee kommen, dass es um das Papier gehen könnte. Außerdem könnten wir einen solchen Versuch schon aus moralischen Gründen wohl nie so durchführen, weil man die Probanden nicht zum Rauchen bewegen dürfte.

Ein anderes Problem könnte sich, wie oben schon erwähnt, aus der Verteilung von Kofaktoren ergeben. Nehmen wir an, es bekämen 10% der Probanden in unserer Kontrollgruppe Lungenkrebs, ohne geraucht zu haben. Tabak sei aber ein dominanter Faktor zusammen mit bestimmten genetischen Kofaktoren für den Lungenkrebs. In der Experimentalgruppe sei der Kofaktor so verteilt, dass genau 10% der Versuchspersonen ihn aufweisen und dann mit Sicherheit einen Krebs entwickeln und für die anderen 90% sei der Kofaktor nicht vorhanden, und für die entfalte der Tabak sogar eine wirksame Schutzfunktion vor Krebs. Dann finden wir in beiden Gruppen dieselbe Quote von 10% von an Lungenkrebs Erkrankten vor. Damit bleiben uns die genannten kausalen Zusammenhänge trotz Randomisierung verborgen. Das ist wiederum ein Fall von mangelnder Graphentreue der Verteilung. Wir benötigen hier tiefergehende Hypothesen über die Kofaktoren, die uns andere Experimente vorschlagen – etwa mit Subpopulationen mit denselben Kofaktoren. Doch dazu müssen wir erst einmal Hypothesen darüber entwickeln, welche Kofaktoren in unserem Fall am Werk sein könnten.

Dominik Janzing (2003, 56 ff.) spricht in einem Beispiel auf eine weitere praktische Schwierigkeit solcher Experimente an, nämlich die »Compliance«-Probleme, die auftreten können. Nehmen wir dazu einen Doppelblindversuch mit einer Medizin versus Placebo (Einnahmevariable X) der durch eine Interventionsvariable Z (Aufforderung zur Einnahme) verursacht wird. Die Heilung sei wieder durch Y dargestellt. Es gibt aber eine unbekannte Störvariable U, die hier etwa Aspekte der Persönlichkeitsstruktur betrifft, etwa ob jemand ein *Starrkopf* ist oder

nicht. Von den Starrköpfen, die sich nichts sagen lassen wollen, nehmen nun etliche ihre Medizin nicht ein. Dann funktioniert die Intervention nicht wirklich sauber, da auch U auf X einwirkt und nicht zuverlässig abgeschaltet wurde. Viele Starrköpfe könnten verhindern, dass eine an sich wirksame Medizin als solche in Erscheinung tritt, aber es könnte auch noch schlimmer kommen.

Sollte U zusätzlich noch Auswirkungen auf die Heilung haben, wäre es eine gemeinsame Ursache von X und Y und unsere Randomisierung Z hätte wiederum nicht funktioniert. Die Randomisierung wäre keine saubere Intervention mehr und das Experiment könnte den durchschnittlichen Effekt der Medizin nicht sauber bestimmen. Hier sind unterschiedlichste Konstellationen denkbar. Wenn es viele Starrköpfe gäbe und die zudem noch besonders kränklich wären, könnte das Experiment auch für eine an sich wirksame Medizin ungünstig ausgehen.

Fisher (1951) setzt in seinen Regeln für das Experimentieren ganz auf die Randomisierung und sie ist sicher ein sehr wichtiges Instrument für das Experimentieren, allerdings ist es nicht ganz so einfach und erfolgreich, wie es manchmal unkritisch angenommen wird. Ganz deutlich wird schon bei Fisher, dass er es vor allem als entscheidendes Instrument betrachtet, um damit bestimmte kausale Hypothesen eliminieren zu können. Im Idealfall entsprechen unsere Versuchsgruppe und die Kontrollgruppe durch eine Randomisierung einer idealen Intervention und unterscheiden sich in ihrer Zusammensetzung nur durch den Prüffaktor X. Sollte sich dann im Hinblick auf Y ein Unterschied zwischen den Gruppen ergeben, bleibt uns nur eine Erklärung für diesen Unterschied übrig, nämlich dass der auf den Unterschied im Hinblick auf X zurückgeht. Wir hätten so den idealen Fall eines Schlusses auf die beste Erklärung vor uns, in dem es uns gelungen ist, alle anderen Erklärungen bis auf die eine zu eliminieren. Die Randomisierung leistet im idealen Fall genau das.

Die Hypothese, ein unbekannter Faktor U hätte diesen Unterschied herbeigeführt, wird eliminiert durch den Hinweis auf die Randomisierung, wonach U in E und K gleich häufig vorkommt und daher keinen Unterschied bewirken kann. Allerdings verbleiben in dieser Elimination alle Unsicherheiten des ganzen Verfahrens, das eben nur ein

Zufallsverfahren ist. Einige Probleme haben wir oben bereits geschildert. Weitere beschreibt etwa Freedman (1999).

7.3.12 Ein Beispiel aus der Praxis

Dass diese Verfahren zur Ursachensuche in der Praxis nicht ganz so einfach funktionieren, zeigt das folgende Beispiel. Spirtes & Glymour & Scheines (2000, 239 ff.) bieten eine umfassende Rekonstruktion eines Teils der heftig geführten Debatte um die Frage, ob Rauchen Lungenkrebs verursacht, deren wichtigste Werke auch heute noch zitiert werden. Die Bewertung dieser Debatte durch die drei Autoren zeigt m.E. zugleich, dass man die Verfahren des kausalen Schließens nicht als alleinige Kriterien heranziehen sollte. Ausgangspunkt waren die Artikel von Doll und Hill von 1950 und 1952, wonach es eine starke Korrelation zwischen Zigarettenrauchen und Lungenkrebs gab. Das konnte in anderen Studien bestätigt werden und wurde als starkes Indiz dafür betrachtet, dass (H) Rauchen Lungenkrebs verursacht.

Vor allem Sir Ronald Fisher (1959) betätigte sich als Kritiker dieser Hypothese und hielt die sogenannte *konstitutionelle Hypothese* (K) entgegen, wonach ein Common Cause für Beides verantwortlich wäre, also vermutlich eine entsprechende genetische Disposition zum einen Ursache für Nikotinsucht und zum anderen für Lungenkrebs wäre. Außerdem wurde kritisch angemerkt, dass die Studien retrospektiv waren. In den 50er, 60er und 70er Jahren gab es zahlreiche Artikel, die die einfache kausale Hypothese H vertraten und die Vertreter der konstitutionellen Hypothese K waren eine Minderheit, die aber nach Spirtes et al. eigentlich gute Argumente hatten, denn die Daten der Vertreter von H konnten nach den Regeln kausalen Schließens H nicht wirklich etablieren. Das mag zwar stimmen, aber die Frage bleibt, ob sie nicht trotzdem schon damals die besseren Argumente auf ihrer Seite hatten; auch ohne ein entsprechendes Verfahren wie die genannten Algorithmen. Dazu möchte ich die wesentlichsten Argumente kurz aufnehmen und nicht so sehr auf die u.a. sehr heftig geführte Debatte mit ihren persönlichen Angriffen eingehen.

Zu den Kritikern der kausalen Hypothese H gehörte u.a. Burch (1983), der die Unzulänglichkeiten der Daten für den Schluss auf einen kausalen

Zusammenhang betont, während unter den Vertretern von H noch Cornfield et al. (1959) zu nennen sind, die einen Überblick über die positiven Indizien geben. Fisher (1959) verwies darauf, dass Korrelationen Kausalbeziehungen nicht festlegen würden, und konnte seine Common-Cause-Spekulation K noch durch einige Belege untermauern. So scheinen sich eineiige Zwillinge in ihrem Rauchverhalten weniger zu unterscheiden als zweieiige. Das schien dafür zu sprechen, dass das Rauchverhalten auch durch die Gene beeinflusst wird. Außerdem gab es Hinweise darauf, dass einige Krebsarten ebenfalls genetische Ursachen haben. Cornfield et al. (1959) stellten dagegen noch einmal die Belege für die kausale These zusammen:

(1) Starke Korrelationen zwischen Rauchen und Lungenkrebs auch in den Subpopulationen. Raucher schienen 26-mal so häufig Lungenkrebs zu bekommen wie Nichtraucher.
(2) Die Lungenkrebsraten stiegen genau parallel (allerdings mit einer ca. 20-jährigen Verzögerung) zu dem Anstieg der Raucher an.
(3) Die Lungenkrebsraten waren in städtischen Populationen höher als in den ländlichen mit weniger Rauchern.
(4) Männer waren stärker vom Lungenkrebs betroffen als Frauen und rauchten auch stärker.
(5) Die Unterschiede zwischen den Betroffenen in anderen Lebensumständen schienen zu gering, um die große Kluft zwischen Rauchern und Nicht-Rauchern zu erklären.
(6) Die Lungenhärchen von Kühen, Ratten und Hasen zeigten nach dem Rauchen Beeinträchtigungen ihrer Funktion.
(7) Zigaretten-Teer konnte die Zellen von Hunden verändern.
(8) Rauchen führte bei Mäusen zu Zellveränderungen, aber nicht zu Lungenkrebs.
(9) Einige Inhaltsstoffe des Rauches waren als karzinogen bekannt.

Die Autoren waren der Meinung, dass man die Hypothese K zwar logisch gesehen nicht zwingend widerlegt hatte, aber diese eben doch immer schwieriger aufrecht zu erhalten war, weil sie die einzelnen Fakten nur noch mit zusätzlichen Ad-hoc-Annahmen erklären konnte. Die Vertreter der konstitutionellen Hypothese versuchten, diese Punkte jedenfalls auch zu erklären. Den Anstieg der Lungenkrebsraten führten

sie etwa auf verbesserte Diagnoseverfahren und eine Tendenz der Ärzte zurück, bei starken Rauchern besonders auf Lungenkrebs zu achten. Aus Sicht des abduktiven Schließens, das die Debatte sehr gut beschreibt, gab es aber schon einige Erklärungsanomalien der konstitutionellen Hypothese. So waren z.B. Mormonen, die nicht rauchten, weniger von Lungenkrebs betroffen als andere Menschen. Sollte das durch einen genetischen Common Cause zu erklären sein, müsste der auch für die religiöse Ausrichtung verantwortlich sein. Auch ohne zwingende kausale Argumente hielten die Anhänger der kausalen Hypothese H das für so wenig plausibel, dass sie das nicht als Erklärung akzeptieren konnten. In derartigen Zusammenhängen finden sich wohl auch die gewichtigsten Gründe gegen K.

Die konstitutionelle Hypothese konnte die vorgelegten Fakten nur mit gewissen Ad-hoc-Annahmen erklären und war damit zwar nicht widerlegt, aber bot eindeutig schon in der damaligen Zeit die schlechteren Erklärungen als H. Allerdings gab es auch Studien wie die skandinavische Zwillingsstudie (Cederloff et al. 1977), die keine brauchbaren Ergebnisse erbrachte, so erkrankten 2 Raucher und zwei Nichtraucher von jeweils eineiigen Zwillingen, was nicht für die Hypothese H sprach, aber auch nicht dagegen. Wir haben es in der Praxis leider oft mit »unklaren« Daten zu tun.

Interessant ist sicher noch, dass es auch erste Experimente in dieser Frage gab, die aber ebenfalls praktisch ergebnislos verliefen. So berichteten Rose et al. (1982) über eine zehnjährige randomisierte Interventionsstudie, bei der aus einer Gruppe mittelalter männlicher Raucher eine Versuchsgruppe ausgewählt wurde, deren Mitglieder dann ermutigt wurden, das Rauchen aufzugeben, worüber sie später selbst berichteten. Es ließ sich kein signifikanter Unterschied in den Lungenkrebsraten zu der verbleibenden Rauchergruppe feststellen. Die neue »freiwillige« Nichtrauchergruppe hatte sogar die höheren Mortalitätsraten. Das Problem zeigt sich immer wieder in der Praxis, dass solche randomisierten Studien nicht das Ergebnis liefern, das wir uns von ihnen erhoffen.

Die Frage ist dann jeweils, was die beste Erklärung für diesen Misserfolg ist. Eine Möglichkeit besteht natürlich darin, dass unsere ursprüngliche Hypothese falsch ist. Eine andere muss im Einzelfall gesucht werden. Für die Studie von Rose et al. können wir etwa fragen, ob die Studie

lang genug war, um die positiven Wirkungen des Nichtrauchens schon aufzuzeigen. Und wir müssen fragen, ob die Selbstberichterstattung über die eigenen Erfolge beim Nichtrauchen immer ganz zuverlässig war. So lassen sich im Prinzip aussagekräftigere Studien ausdenken. Allerdings dürfte hier das oben schon erwähnte Problem eines möglichen »Compliance Bias« ernst zu nehmen sein. Nur ein Teil der zufällig ausgewählten Raucher war tatsächlich bereit, auf das Rauchen zu verzichten. Es ist gut nachvollziehbar, dass das diejenigen waren, die sich auch eher in anderen Bereichen dann einer gesünderen Lebensweise zuwendeten. Dann könnte der Effekt des Nichtrauchens überzeichnet sein. Oder sie lebten schon immer gesünder als die anderen Raucher, dann könnte der Effekt des Nichtrauchens geringer geschätzt werden, als er tatsächlich war. Viele solche Beispiele belegen, wie schwierig sich die Suche nach Ursachen in der Praxis gestaltet.

7.3.13 Randomisierte Experimente

Sir Ronald Fisher (1951) setzte jedenfalls ganz auf randomisierte Experimente. Seine Begründung dafür ist gut nachvollziehbar und deckt sich eindeutig mit den Ideen des abduktiven Schließens bzw. der eliminativen Induktion (oder dem Schluss auf die einzig verbliebene Erklärung). Nehmen wir an, wir haben aus unserer Population eine Gruppe von Probanden ausgewählt, um ein Medikament zu testen, so werden wir diese Gruppe per Zufallsauswahl auf die Versuchs- und die Kontrollgruppe aufteilen, weil wir hoffen, damit eine Vielzahl (möglichst alle bis auf eine) von kausalen Hypothesen eliminieren zu können. Sollte sich in der Versuchsgruppe ein bestimmter Effekt Y gegenüber der Kontrollgruppe zeigen, so schließen wir, dass er nicht auf noch unbekannte Faktoren U zurückzuführen sein kann, da die vermutlich in beiden Gruppen durch die Randomisierung ungefähr gleichverteilt sein sollten. Dann wäre es aber unwahrscheinlich, dass Y auf U zurückzuführen wäre, denn U liegt schließlich in ausgeglichener Weise (so der Idealfall) in beiden Gruppen vor. Dann wird ein Schluss auf die Definition von kausaler Wirkung und die Differenzmethode zurückzuführen sein.

Wir haben oben aber schon einige Einwände und Probleme zu dieser Schlussweise genannt. Fisher scheint ihre Möglichkeiten auch für die

Praxis etwas zu überschätzen (das zeigte auch das Raucher-Experiment), aber wir sollten uns genausowenig der Kritik und dem recht negativen Urteil von Howson & Urbach (1993, Kap. 11) gegenüber dem Verfahren anschließen. Allerdings hat Worrall (2007) recht – und belegt das auch in Fallstudien –, dass das Vertrauen in die Randomisierung in der Praxis etwa der Medizin oft so übertrieben wird, dass andere Daten kaum noch zur Kenntnis genommen werden, obwohl sie durchaus ebenfalls etwas über kausale Zusammenhänge aussagen können.

7.4 Kausale Modelle

7.4.1 Die Grundideen der Regressionsrechnung

Viele kennen die Regressionsrechnung aus der Statistik und nicht immer werden dabei die Bezüge zu kausalen Vermutungen deutlich gemacht. Wir werden hier Gleichungen des Typs

(*) $Y = f(X_1,\ldots,X_n)$ (etwa mit einer Funktion $f:\mathbb{R}^n \rightarrow \mathbb{R}$)

zumindest immer so verstehen, dass die erklärenden Variablen $X_1,\ldots,X_n$ auf der rechten Seite der Gleichung *Ursachen* der Zielvariablen Y auf der linken Seite sein sollen. Das ist jedenfalls unsere Zielvorstellung, denn wir suchen letztlich nach entsprechenden kausalen Beziehungen. Deshalb gehört das Thema auch ins Kapitel 7. Dabei sieht die Regressionsrechnung manchmal so aus, als ob man hier ohne weiteres Hintergrundwissen direkt (per Algorithmus) auf bestimmte Kausalhypothesen schließen könne.

Für die Bezeichnungen halte mich hier an die Benennungen in Fahrmeier et al. 2009; die sind jedoch keineswegs einheitlich für den Bereich der Regressionsrechnung. Sie verdeutlichen uns aber, worum es geht. Wir *gehen* von der Zielvariablen Y *zurück* auf die erklärenden Variablen, also normalerweise die Ursachen von Y und versuchen die vorliegende funktionale Abhängigkeit f genauer zu bestimmen bzw. zu schätzen.

Wir können die Gleichung solcher Regressionen nach Freedman (1997) genau genommen auf drei Arten verstehen. Zunächst kann (*) einfach

als bloße Beschreibung bestimmter Daten dienen. So haben wir in der Population bestimmte Objekte $o_1,\ldots,o_n$, an denen wir zwei Größen X und Y gemessen haben und fragen uns, welche Geradengleichung diese Daten am besten systematisiert. Das hat noch nichts mit einer kausalen Interpretation von (*) zu tun. Man kann (*) dann trotzdem nutzen, um damit eine Vorhersage abzugeben. Nehmen wir an, X sei der Grad der Gelbfärbung der Finger und Y das Lungenkrebsrisiko, dann taugt X sehr wohl (wenn wir einmal annehmen, dass es sich um eine relativ stabile Beziehung handelt), um auch neue Werte des Krebsrisikos aus neuen Werten von Gelbfärbung vorherzusagen, obwohl es sich bei (*) danach nicht direkt um eine kausale Gleichung handelt. Allerdings bietet sie nur dann einen guten Grund für eine Vorhersage, wenn es sich zumindest um eine stabile Gleichung handelt, die etwa durch einen Common Cause im Hintergrund bestimmt wird.

Erst die dritte Interpretation von Gleichung (*) durch Freedman stellt einen kausalen Zusammenhang dar (vgl. Woodward 1999, 2003, Kap. 7), und wir schließen auf sie mit Hilfe eines kausalen Schlusses. Danach beschreiben die X_i die direkten Ursachen von Y und (*) gibt uns an, was unter bestimmten Interventionen an den X_i passieren würde. Das ist die Deutung, der auch ich hier folgen werde. Das ist jedenfalls das, wonach wir in solchen Regressionsrechnungen suchen sollten und dann erst befinden wir uns wieder im Gebiet kausalen Schließens.

Wenn wir die qualitative kausale Struktur in einem bestimmten Bereich aufgeklärt haben, können wir dadurch noch einen Schritt weitergehen und versuchen, die genaueren funktionalen Zusammenhänge zwischen den Größen zu beschreiben. Das soll u.a. dazu dienen, das Ausmaß bestimmter Zusammenhänge oder bestimmter Maßnahmen in unserer Population zu quantifizieren. Wissen wir etwa, dass das Rauchen eine Ursache für den Lungenkrebs ist, so bleibt immer noch die Frage offen, welches Ausmaß bestimmte Maßnahmen in unserer Population haben würden, wie etwa die Durchsetzung des Rauchverbots. Gibt es z.B. in der vorliegenden Grundgesamtheit nur 1% Raucher, von denen nur jeder Tausendste einen Lungenkrebs entwickelt, so ist schon einmal klar, dass durch das Rauchverbot höchstens einer von 10 000 Personen gerettet würde. Käme nun noch hinzu, dass nur jeder zweite Lungenkrebspatient aufgrund des Rauchens (eigenes oder Mitrauchen)

erkrankt, könnte das Verbot nur einen von 20 000 Personen retten. Das könnte etwa im Vorfeld wichtig sein, wenn wir über solche Maßnahmen nachdenken. Dazu suchen wir nach quantitativen Zusammenhängen zwischen den Größen etwa im Sinne einer *linearen Gleichung* im Rahmen der Regressionsrechnung:

(LG) $Y = a_0+a_1X_1+\ldots+a_nX_n + U$

Die Gleichung (LG) soll hier so zu lesen sein, dass die erklärenden Größen $X_1,\ldots,X_n$ als Ursachen der Zielgröße Y aufgefasst werden. Die Parameter $a_1,\ldots,a_n$ geben dann darüber Auskunft, wie stark sich Y ändert, wenn sich eines der X_i ändert, wobei hier schon angenommen wird, dass die X_i untereinander nicht interagieren, d.h. keine Kofaktoren voneinander sind, sondern unabhängig voneinander auf Y wirken. Unsere gesuchte Funktion f aus (*) wird hier von Anfang an auf eine bestimmte einfache Form festgelegt. Man sucht nun nach der Gleichung, die die Daten am besten erklären kann. Dafür gibt es das Verfahren der Regressionsrechnung. Dafür kommt noch eine Komplikation ins Spiel, die sich am besten mit einem Schema aus der *Informationstheorie* beschreiben lässt:

Daten = Signal + Rauschen

Das heißt, wir erwarten nicht, dass unsere erklärenden Variablen X_i die Werte der Zielgröße zu einhundert Prozent erklären können, sondern es gibt daneben immer eine Abweichung, die wir als *Rauschen, Störterm, Messfehler* etc. beschreiben können, und die wird hier durch die Zufallsvariable U wiedergeben. Das heißt, wir können das auch so beschreiben:

Messwert = systematischer Wert + Messfehler

In (LG) wird dazu schon angenommen, dass auch U keine Interaktion mit den erklärenden Variablen X_i aufweist. Jedenfalls sollte eine gute Erklärung der Daten diesen »Zufallsfehler« in irgendeiner Form berücksichtigen und ist daher *nicht* darauf angewiesen, dass die lineare Funktion $y = a_0+a_1x_1+\ldots+a_nx_n$ die Daten exakt reproduziert.

Man spricht ebenfalls von einer *Kurvenanpassung*. Zu vorgegeben Daten suchen wir nach einer Kurve (in unserem Spezialfall einer Hyperebene), die nahe an unseren Daten liegt. Der Fehlerterm erlaubt, dass es gewisse Abweichungen zwischen den Daten und dem systematischen Anteil der rechten Seite der Gleichung gibt, ganz nach dem genannten Schema: Daten = Signal+Rauschen. Dieses Rauschen hatten wir oben schon für die Messungen einer so einfache Größe wie dem Gewicht kennengelernt. Die nächsten Fragen werden dann sein: Wie messen wir den Abstand zwischen unserer Kurve und den Daten? Dafür nimmt man heute normalerweise die Summe der Abstandsquadrate, die Frage bleibt aber, warum wir das tun sollten und kein anderes Abstandsmaß verwenden. Zweitens müssen wir wissen, welche Abweichungen wir akzeptieren wollen und was wir dabei gewinnen, wenn wir etwa einen *größeren Fehler* bzw. eine größere Abweichung von der Regressionsfunktion akzeptieren. Meist geht es uns darum, die *einfacheren* Funktionen gegenüber den komplexeren zu bevorzugen (also in unserem Beispiel etwa lineare Funktionen, denn wir haben keine Terme höherer Ordnung X_i^n oder Terme für Interaktionen wie $X_i \cdot X_j$ in unserer Funktionsgleichung).

Das führt schließlich zu den nächsten Fragen, was man genau unter *Einfachheit* verstehen kann und warum wir nach einfacheren Funktionen streben sollten? Geht es nur um unsere mathematische Bequemlichkeit oder bieten die einfacheren Funktionen auch einen Erkenntnisgewinn etwa im Sinne des abduktiven Schließens? Steht vielleicht die metaphysische Spekulation im Hintergrund, dass die Welt im Wesentlichen aus einfachen Zusammenhängen besteht? Das sind schwierige Fragen, auf die es erste Antworten gibt, aber es wird auch klar werden, dass diese Antworten noch nicht das Ende der Debatte darstellen, denn die Antworten sind kaum zwingend, sondern bieten bestenfalls erste dezente Hinweise auf bestimmte Lösungsansätze.

In unserem Raucherbeispiel könnte der lineare Ansatz etwa bedeuten, dass X_1 über die Gesamtmenge an gerauchten Zigaretten Auskunft gibt, und X_2 über die Menge an anderen schädlichen Stoffen (wie Radon oder Asbest) usf. Dann geht man davon aus, dass die Lungenkrebsrate Y linear von beiden Größen abhängt und es nicht zu Synergieeffekten in dem Sinne kommt, dass die Belastungen sich potenzieren würden. Im nächsten Schritt wird man bestimmte Beobachtungsdaten dazu

bestimmen und könnte damit versuchen, die Parameter a_1 und a_2 (und gegebenenfalls weitere) so zu schätzen, dass die resultierende Kurve möglichst nahe an die Daten herankommt. Wenn wir nur X_1 in eine lineare Gleichung einbringen, müssen wir einfach aus allen Geraden die aussuchen, die unseren Daten (x_i, y_i) insgesamt am nächsten kommt.

U beschreibt dann den Fehler, der noch verbleibt, und wird meistens als normalverteilte Zufallsvariable betrachtet. Als Abstandmaß verwenden wir hier wie schon erwähnt die Summe: $\sum_i (y_i - a_0 - a_1 x_1)^2 = \sum_i (d_i)^2$, wobei wir mit d_i nun die senkrechten Abstände zwischen Kurven und Daten bezeichnen wollen. Die Frage war dann, warum wir nicht einfach $\sum_i |d_i|$ oder eine andere Funktion der d_i wählen, um damit die Abstände zwischen der Kurve und unseren Daten (x_i,y_i) zu bestimmen. Darauf kommen wir weiter unten zurück.

Wir nehmen allgemeiner an, dass es jedenfalls irgendeine funktionale Abhängigkeit f der Zielgröße Y von ihren kausalen Eltern PA_Y innerhalb eines kausalen Netzes gibt, die wir durch eine Funktion f beschreiben können, und suchen nach einem geeigneten f. Dann setzen wir ganz allgemein so an:

(FG) $Y = f(PA_Y, U)$

Dabei soll wiederum U als Variable für die noch unbekannten Größen eingeführt werden oder einfach für die zu erwartenden Messfehler bzw. Störterme. Das heißt, U wird eingeführt, um die Differenz zwischen den Werten $f(PA_Y)$ und Y zu erklären.

Für jede Stufe im Kausaldiagramm und jede Wirkung führen wir so eine Gleichung ein, die einen speziellen kausalen Mechanismus in unserem größeren kausalen Netz beschreiben soll. Wenn wir die Eltern von Y genau identifiziert haben, können wir sie explizit in unserer Gleichung auflisten mit $\mathbf{X} = (X_1,\ldots,X_n)$ und die Gleichung neu formulieren:

(GG) $Y = f(\mathbf{X},U)$

Die Zufallsvariable U kann also unterschiedliche Deutungen erfahren. In einem deterministischen Rahmen (den selbst Pearl 2000 in seinem recht probabilistisch geprägten Ansatz annimmt) kann U ein Ausdruck für die

noch *unbekannten Faktoren* sein, die neben den X_i auch auf Y einwirken. Sie werden einfach in einer Zufallsvariablen zusammengefasst.

Eine andere Lesart wäre, es als eine Beschreibung des *Messfehlers* zu deuten, den wir bei jeder Messung berücksichtigen müssen. Im indeterministischen Fall können wir durch U auch einen genuin probabilistischen Anteil zum Ausdruck bringen. Meist nimmt man jedoch an, dass dieser Fehlerteil sich als unabhängig von unseren anderen Ursachenfaktoren einbringen lässt, und notiert die Gleichung daher so:

(EG) $Y = f(\mathbf{X}) + U,$

wobei man meist noch einige Annahmen über U akzeptiert, auf die ich weiter unten zu sprechen komme. Um eine solche Gleichung besser verstehen zu können, starten wir etwa mit einer einfachen linearen Gleichung, die sich etwa im Rahmen einer linearen Regressionsrechnung ergibt:

(LG) $Y = a_0 + a_1X_1 + \ldots + a_nX_n + U$

Komplexere Funktionen lassen sich damit im Prinzip auch modellieren, da wir z.B. die Möglichkeit haben, neue Zufallsvariablen etwa $X^* = \log(X)$ einzuführen und diese dann in einer linearen Regression einzubringen (vgl. Fahrmeier et al. 2009), doch diese ganzen technischen Komplikationen möchte ich nicht weiter verfolgen. Dabei geht man oft davon aus, dass der Fehlerterm oder die Störgröße U eine Reihe von Bedingungen erfüllt (vgl. Woodward 1999, 2003, Kap. 7; Fahrmeier et al. 2009: die 5 Axiome der Testtheorie):

(FA) Typische Fehlerterm-Annahmen

F1 $E(U) = 0$

F2 Homoskedastizität bzw. die Konstanz der Varianz bzgl. aller Werte von **X**.

F3 Statistische Unabhängigkeit bzw. keine Autokorrelation des Fehlerterms für verschiedene Werte von **X**.

F4 Statistische Unabhängigkeit von U und den erklärenden Variablen X_i.

Diese Bedingungen stellen jeweils ganz bestimmte Annahmen dar, die wir mit der Zufallsvariable U verbinden. Die erste Bedingung (F1) besagt, dass U nur einen unsystematischen Fehler beschreibt bzw. die Wirkung der nicht beobachteten Störvariablen sich im Durchschnitt in unserer Population wieder aufhebt. Damit geht für die Beurteilung der Wirkung insgesamt nicht wirklich viel verloren, wenn wir die Variablen in U außen vor lassen. Vereinfachend wird außerdem angenommen (F2), dass die von U hervorgerufenen Abweichungen an jeder Stelle unserer Kurve (LG) den gleichen Umfang annehmen und nicht in einem Bereich besonders groß sind. Das ist oft hilfreich für bestimmte Abschätzungen und kann deshalb zunächst einmal so stehenbleiben. Natürlich können wir solche Annahmen auch wieder aufgeben. In (F3) geht es darum, dass die Fehler an bestimmten Stellen nicht von den Fehlern an anderen Stellen abhängen sollen. Man könnte sagen, das gehört zu unserer Vorstellung von Fehlern, dass sie keine solchen systematischen Zusammenhänge aufweisen, was in (F4) noch weitergetrieben wird, wonach die Fehler auch nicht mit unseren erklärenden Variablen statistisch zusammenhängen. Dabei gehen wir in der Regel davon aus, dass eine kausale Unabhängigkeit auch eine statistische Unabhängigkeit nach sich zieht. Dass das nicht immer zwingend der Fall sein muss, hatten wir bereits im Zusammenhang mit der Graphentreue diskutiert. (Wie Regressionsverfahren ohne solche Annahmen aussehen, finden wir z.B. auch bei Fahrmeier et al. (2009).)

Nehmen wir als einfaches Beispiel im Rahmen unserer Annahmen die Gleichung:

(BG) (Lungenkrebsrate) = a_0 + a_1·(Anzahl gerauchter Zigarretten) + U

Wir nehmen für (BG) also an, dass der durchschnittliche Fehler bzw. die durchschnittliche Abweichung U jeweils Null ist, sonst müssten wir nach einem anderen funktionalen Zusammenhang zwischen unseren beiden Variablen suchen, der diese systematischen Abweichungen berücksichtigt. Eine gewisse Grundquote an Lungenkrebs wird außerdem durch die Konstante a_0 abgedeckt. Die zu erwartenden Abweichungen denken wir uns als dieselben für unterschiedliche Werte der Variablen (Anzahl gerauchter Zigaretten), obwohl das vermutlich kaum sehr realistisch ist. Man könnte etwa annehmen, dass für hohe Anzahlen auch die

Schwankungen etwas größer sind, aber als brauchbare mathematische Idealisierung mag das zunächst durchgehen. Die zufälligen Abweichungen für 10 000 gerauchte Zigaretten sollen zudem nicht von denen für 20 000 abhängen, denn die gesamte Abhängigkeit unserer Variablen steckt bereits im systematischen Teil der Gleichung.

Diese Annahmen sind von Bedeutung, wenn wir unser Verfahren rechtfertigen, etwa im Hinblick auf die Methode der kleinsten Quadrate. Nehmen wir z.B. an, U sei überall normalverteilt mit $U \sim N(0,\sigma^2)$ und einem unbekannten σ (d.h. P(U=u) folgt der Normalverteilung mit Dichtefunktion $\varphi(x) = (2\pi)^{-½} \exp(-½ (x/\sigma)^2)$ und $P(U \leq u) = \int_{x \leq u} \varphi(x)\, dx$, wobei wir für U also kleine Intervalle betrachten sollten). Dann können wir eine bestimmte Rechtfertigung für eine Gerade anhand der Methode der kleinsten Quadrate finden, denn sie ist dieselbe wie der entsprechende Maximum-Likelihood-Schätzer (vgl. Fahrmeier et al. 2009).

Das heißt, wenn wir mit der Methode der kleinsten Quadrate eine Gerade auswählen, ist sie die Gerade g, für die gilt, dass die Wahrscheinlichkeit für die Daten bei dieser Gerade g die höchste Wahrscheinlichkeit im Vergleich aller Geraden g* ist, d.h., für alle g* gilt: $P(\text{Daten}|g) \geq P(\text{Daten}|g^*)$. Das entspricht einer abduktiven Begründung, denn ceteris paribus ist bei dieser Likelihoodungleichung g eine bessere Erklärung für die Daten als andere Geraden g*; und g ist die Gerade, die am ehesten von allen Geraden genau unsere Daten vorhersagen würde. Für diese speziellen Annahmen über U gibt es also zunächst eine Begründung dafür, den Abstand zwischen der Geraden und den Daten anhand der Summe der Quadrate zu bemessen. Allerdings ist die Erklärungsstärke nicht nur durch die entsprechenden Likelihoods bestimmt, sondern auch durch andere Parameter und außerdem wurde der Aspekt, dass die Daten sich aus systematischen und zufälligen Anteilen zusammensetzen, bisher noch nicht genügend berücksichtigt.

Bisher haben wir uns nur auf *Geradengleichungen* beschränkt, aber wenn wir etwa *allgemeinere Polynome* höheren Grades zulassen, können wir damit meistens eine deutlich bessere Anpassung an die Daten erreichen (kleinere Werte der Quadratsummen der Abweichungen), zumal die linearen Funktionen immer noch als Grenzfall mit enthalten sind. Allerdings droht nun eine *Überanpassung*, die mehr die Zufalls-

schwankungen nachzeichnet, als die systematischen Signale zu ermitteln, wenn wir etwa beliebige Polynome zulassen. Darum sehen einige Verfahren Strafterme für eine zu komplexe Funktion vor, mit deren Hilfe die Anpassung vorgenommen werden soll, um der Überanpassung entgegenzuwirken. So werden z.B. im Ansatz von Akaike (1973) den verglichenen Likelihoods jeweils Strafterme für die Komplexität der Funktionen hinzugefügt. Dabei wird die Komplexität etwa durch die Anzahl der freien Parameter eines komplexen Polynoms bestimmt.

7.4.2 Penalisierung und Erklärungen auch für zukünftige Ereignisse

Wie bereits erwähnt, fragen wir uns in der Regressionsrechnung, was die beste Erklärung für unsere Daten ist und suchen nach einer erklärenden Gleichung. Denken wir zunächst an einen Zusammenhang zwischen bloß zwei Faktoren X und Y der Art: Y = f(X)+U. Dabei suchen wir natürlich nach einer insgesamt besten Erklärung, die nicht nur unsere bisherigen Daten gut erklären kann, sondern der entscheidende neue Aspekt ist, dass wir auch die zu erwartenden zukünftigen Daten mit einbeziehen. Selbstverständlich wollen wir eine Erklärung finden, die auch für sie mit zutrifft.

Doch das hilft uns zunächst nicht weiter, da diese Daten schließlich noch unbekannt sind. Sollten wir allerdings Grund zu der Annahme haben, dass sie von einem bestimmten Typ sind, den wir charakterisieren können, so sind sie selbstverständlich mit zu berücksichtigen. Genau das geschieht im Akaike Theorem. Wenn wir annehmen, dass gilt: Messwert = systematischer Wert + Messfehler, so dürfen wir nicht nach einer Kurve suchen, die die Messwerte exakt reproduziert. Wir könnten sonst etwa an ein Polynom höheren Grades (vielleicht 20 oder mehr) denken, dass alle Windungen unserer bisherigen Daten mitmacht.

So eine Kurve würde unsere bisherigen Messfehler exakt nachbilden und dadurch als systematische Anteile darstellen, doch das wäre eine Überanpassung, die neue Daten nicht so gut beschreiben würde, denn der Fehleranteil ist unsystematisch und wird daher nicht in derselben Weise in den neuen noch zu erwartenden Daten wiederzufinden sein. So verfahren wir in der Praxis der Regressionsrechnung deshalb nicht. Wenn

jemand mit einem Polynom 45-ten Grades daherkommt, das unsere Daten exakt reproduziert, werden wir ihn nicht für eine besonders gute Arbeit loben, wenn wir mit einer einfachen Geradengleichung schon eine relativ gute Approximation erzielen.

Das *Theorem von Akaike* weist uns dazu einen Weg, wie wir auf systematische Weise die Komplexität unserer Funktion und die Güte der Anpassung miteinander verrechnen können. Sie stützt sich dazu auf die zu erwartende Prognoseleistung unterschiedlicher Regressionsverfahren.

Eine bessere Prognose liefert statt einer optimalen Anpassung an die Daten (und das ist gerade der Inhalt des Akaike Theorems vgl. Forster & Sober 1994) eine Abtrennung des systematischen Anteils vom unsystematischen oder nicht erfassten Anteil. Das funktioniert in der Praxis so, dass ein Strafterm für die höhere Komplexität der Kurve eingeführt wird, der genau so ausgestaltet wird, dass die resultierenden Kurven die höchsten zu erwartenden Likelihoods für die noch zu erwartenden Daten aufweisen. Das heißt aber gerade, dass wir nach solchen Kurven suchen, die nicht nur die alten Daten, sondern auch noch die zu erwartenden Daten am besten erklären können. Dabei benutzt Akaike die Kullback-Leibler-Distanz für Verteilungen, um die Abstände zwischen Modell und zu erwartenden Daten darzustellen, wobei wir außerdem davon ausgehen, dass die neuen Daten durch dieselbe Verteilung produziert werden, wie die schon gefundenen Daten. So wird zum ersten Mal überzeugend erklärt, inwiefern Einfachheit einen epistemischen Wert darstellt und nicht nur mit unserer Bequemlichkeit zu tun hat (vgl. Forster & Sober 1994).

Das ist eine neue Wendung auch für das abduktive Schließen. Selbstverständlich verlangt ein Schluss auf die beste Erklärung, dass alle relevanten Daten, die es jemals geben wird, möglichst gut durch unsere Hypothesen zu erklären sind. Haben wir also Grund zu der Annahme, bereits bestimmte Dinge über zukünftige Daten zu wissen, so sollten wir das natürlich in unseren heutigen abduktiven Schlüssen berücksichtigen. Nichts anderes geschieht im Akaike Ansatz, der zum Akaike Informationskriterium AIC geführt hat. Das beruht allerdings darauf, dass wir genauer angeben können, welche Daten denn eigentlich zu erwarten sind. Dazu macht Akaike in seinem Theorem zumindest eine plausible Kontinuitätsannahme. Er geht davon aus, dass die zugrundeliegende

Verteilung, die unsere bisherigen Daten hervorgebracht hat, dieselbe ist, die auch die neuen Daten hervorbringen wird.

Diese Annahme scheint eine recht grundlegende Uniformitätsannahme zumindest für viele unserer Induktionsverfahren zu sein. Haben wir keine gegenteiligen Hinweise, ist das sicher unsere nächstliegende Arbeitshypothese. Sie geht im Akaike Ansatz in den Schluss auf die beste Erklärung mit ein und unser allgemeines Modell, wie sich die Daten zusammensetzen, natürlich auch. Die Komplexität unserer Funktionen wird nach Funktionsklassen bestimmt, die durch die Anzahl der freien Parameter gegeben ist, die diese Klassen beschreiben. So wird die Funktionsklasse $f(x) = a_0 + a_1x^1 + a_2x^2 + \ldots + a_nx^n$, durch die n+1 Parameter $a_0, a_1, \ldots, a_n$ beschrieben und erhält einen Strafterm, der von n+1 abhängt und dessen genaue Ausgestaltung durch das Theorem von Akaike bestimmt wird.

Man geht dann so vor, dass man in jeder Funktionsklasse die beste Funktion (etwa mit der maximalen Likelihood L_n relativ zu den Daten d) heraussucht und anschließend die Sieger in ihren Klassen jeweils mit dem Klassenstrafterm versieht und schaut, wer dann den besten Wert (Likelihood minus Strafterm) aufweist. Dabei wird $AIC = -\ln(L_{n+1}) + 2(n+1)$ gewählt und der jeweils kleinste AIC-Wert gewinnt, d.h., die entsprechende Funktion liefert uns das beste Modell. Je mehr Parameter vorkommen, um so schlechter wird also der AIC-Wert, aber für viele Daten fällt der Strafterm nicht mehr so stark ins Gewicht, da L_{n+1} nahe bei Null liegt und damit $\ln(L_{n+1})$ sehr klein und so $-\ln(L_{n+1})$ sehr groß wird. Dann entscheiden eher Unterschiede in den Likelihoods als in der Anzahl der Parameter.

Das so gewählte Modell bietet die insgesamt beste Erklärung für unsere tatsächlichen und noch erwarteten Daten. Das ist für den Durchschnitt der zu erwartenden Daten beweisbar, wenn wir die Annahmen des Akaike Theorems akzeptieren. Allerdings führen schon kleine Änderungen in den Abstandsmaßen zu kleinen Änderungen in den Straftermen und anderen Kriterien als AIC. So gibt es inzwischen viele weitere Kriterien wie das bayesianische BIC Kriterium, das wir bereits in Kapitel 6 erwähnt haben (s.a. Forster 2000).

Schlimmer ist aber noch, dass wir keinen Grund dafür haben anzunehmen, dass unsere Funktion f in (EG) immer ein durchgängiges Polynom

sein muss. Bei Fahrmeier et al. (2009) finden sich Beispiele dafür, dass auch Polynome höheren Grades nicht immer gute Ergebnisse liefern müssen. Warum sollte der liebe Gott die Zusammenhänge zwischen X und Y nur durch ein Polynom gestalten? Er könnte sich ganz andere Zusammenhänge ausdenken. Und auch die Regression wird einfacher und besser, wenn wir etwa nur annehmen, dass die Bereiche stabiler einfacher Zusammenhänge kleiner sind. Dann funktioniert allerdings das oben beschriebene Verfahren nicht mehr.

Wir können unseren Anwendungsbereich in viele kleine Intervalle aufteilen und dann jeweils in diesen kleinen Abschnitten die Daten durch einfache polynomiale Kurven niedrigen Grades darstellen. Fahrmeier et al. (2009) sprechen da von *Polynom-Splines*. Deren Komplexität wird vor allem bestimmt durch die Anzahl der Knoten und die *Glätte* der Funktion an diesen Knoten (man verlangt dort mindestens die Stetigkeit der Gesamtfunktion), die durch den Differenzierbarkeitsgrad an diesen Stellen festgelegt wird. Doch damit bewegen wir uns langsam weg, von unseren klaren und interpretierbaren Vorstellungen von (einfachen) Komplexitätsstraftermen. Die Debatte wird hier sicherlich weiterzuführen sein, aber der große Durchbruch, den Akaike noch für Forster & Sober (1994) zu bringen schien, hat sich z.T. wieder verflüchtigt.

Eine andere Frage, der man sich in der Statistik ebenfalls zuwendet, ist die, welche unterschiedlichen Faktoren wir überhaupt betrachten sollten. Sind Kreativität und Phantasie zwei unterschiedliche Faktoren oder hängen sie doch sehr eng – vielleicht sogar zum Teil analytisch – zusammen? Die *Faktorenanalyse* und die Clusteranalyse sind u.a. Methoden, um solche Faktoren zu ermitteln.

Das allgemeine Ziel der Faktorenanalyse besteht darin, korrelierende Variablen auf höherer Abstraktionsebene zu Faktoren zusammenzufassen. Damit ist die Faktorenanalyse ein datenreduzierendes Verfahren. Musste man im obigen Beispiel zur Charakterisierung einer Person zunächst Messwerte für die sechs Variablen heranziehen, sind es nach der Faktorenanalyse nur noch zwei Werte (sog. Faktorwerte), die praktisch die gesamten Variableninformationen abbilden. Diese datenreduzierende Funktion der Faktorenanalyse ist besonders dann von Nutzen, wenn man mit sehr vielen Variablen arbeitet und es einfach ökonomischer

und übersichtlicher ist, mit Faktorwerten statt mit vielen korrelierten Einzelmessungen zu operieren (vgl. Bortz & Döring 2006 Kap. 6.4).

7.4.3 Zusammenhänge zum abduktiven Schließen

Insgesamt hat sich gezeigt, dass wir auch in den Regressionsverfahren auf eine Vielzahl von Annahmen – etwa über die Form der Funktionen und welche Faktoren kausal relevant sein könnten – angewiesen sind. Wir suchen dann nach der Funktionalgleichung, die unsere Daten am besten erklären kann. Aber auch hier sind die Kriterien für die Qualität der Erklärung nicht genau anzugeben und lassen etwa Spielräume dafür offen, wie die Einfachheit der Funktion zu bestimmen ist und wie diese in die Bewertung eingeht. Jedenfalls stellen die Regressionsverfahren eine Form des abduktiven, kausalen Schließens dar und keine Konkurrenzmethode für den Schluss auf die beste Erklärung.

In seinen unterschiedlichen Varianten werden beim kausalen Schließen unterschiedliche Hinweise abduktiv ausgewertet, um auf die zugrundeliegenden Kausalzusammenhänge zurück zu schließen. Im Falle der minimalen Theorie mit ihrer Annahme deterministischer Zusammenhänge wurden Regelmäßigkeiten (oder Korrelationen) positiv ausgewertet. Allerdings ließen sie sich nur dann als Indikatoren für Kausalität interpretieren, wenn sie auch in speziellen homogenisierten Situationen (also etwa in sogenannten randomisierten Interventionsstudien, aber nicht *nur* dort) auftraten. Dann blieb nur noch eine Erklärung übrig, denn alle möglichen Konkurrenten konnten durch Homogenisierung eliminiert werden. Als Liste der möglichen erklärenden Theorien finden wir hier alle möglichen kausalen Zusammenhänge, die zwischen den Faktoren bestehen können, als Hypothesen. Es handelt sich wiederum um einen Schluss auf die einzig verbliebene Erklärung, deren Grundidee wir schon in der eliminativen Induktion kennengelernt hatten.

Im Falle der bayesschen Netze wurden dagegen die statistischen Unabhängigkeitsbeziehungen ausgewertet. Dazu mussten wir allerdings eine gewisse strukturelle Übereinstimmung zwischen der tatsächlichen Kausalstruktur und den zu beobachtenden Korrelationen (Markov-Bedingung und Graphentreue) voraussetzen. Außerdem haben wir angenommen, die tatsächliche Kausalstruktur sei frei von Zirkeln und

entspreche somit der Struktur eines azyklischen gerichteten Graphen. Gerade der IC-Algorithmus verdeutlichte dann auch hier wieder, wieso das Ganze auf ein Eliminationsverfahren hinausläuft. Es werden zunächst alle möglichen Zusammenhänge zwischen den Variablen zugelassen und dann schrittweise möglichst viele davon eliminiert. Dabei sind zunächst viele Hypothesen im Spiel, die dadurch gegeben sind, dass man zwischen allen beteiligten Variablen eine mögliche Kausalbeziehung annimmt und diese Listen an Hypothesen dann langsam durch Elimination anhand von Unabhängigkeitsbeziehungen reduziert. Erst durch Streichen von Kanten zwischen den Variablen und dann durch Orientieren der Kanten, wodurch jeweils nur noch eine von den zwei noch möglichen Kausalzusammenhängen übrigbleibt.

Wenn wir weitergehende kausale Modelle aufstellen wollen, geht es auch wieder um die jeweilige Erklärungskraft dieser Modelle. Wir vergleichen ein Kontinuum dieser Modelle anhand ihrer Likelihoods, d.h. anhand der Wahrscheinlichkeit, die sie den Daten geben, wobei wir schon annehmen, dass bestimmte kausale Zusammenhänge zwischen den erklärenden Variablen und der Zielvariablen bestehen. Dann ist diese Likelihood sicher ein wesentlicher Aspekt der Erklärungsstärke. Allerdings ist es nicht der einzige und gerade in den betrachteten Fällen müssen wir einen kleinen nicht erklärbaren Zufallsanteil einräumen und in unseren Vergleichen berücksichtigen, so dass die letzten Theorienvergleiche etwas komplizierter ausfallen und schließlich das passiert, was wir schon für das induktive Schließen und das abduktive Schließen im Besonderen eingeräumt hatten, nämlich, dass die Kriterien für die beste Erklärung auch gewisse Spielräume aufweisen und insbesondere die Verrechnung unterschiedlicher Aspekte der Erklärungsstärke nicht immer einheitlich und unumstritten ist.

Vor allem weisen auch alle drei Ansätze (minimale Theorie, probabilistischer und kontrafaktischer Ansatz) gewisse Problemfälle auf. Wenn wir die drei Theorien nur als Ansätze zum kausalen Schließen verstehen und nicht als Explikationen dessen, was Kausalität ausmacht, so sollte das kein Problem darstellen, sondern entspricht nur der Tatsache, dass alle induktiven Schlüsse irrtumsanfällig sind. Die minimale Theorie versagt spätestens auf der singulären Ebene in den Fällen, in denen wir zwei ganz parallele Prozesse finden, wie den der zwei Zündschnüre,

die nebeneinander einen Sylvesterknaller zünden (vgl. Kap. 7.2.4). Da die minimale Theorie zunächst nur nach der kausalen Relevanz auf der generischen Ebene fragt, liefert sie keine Hinweise darauf, welche Zündschnur welchen Knaller entzündet.

Ähnlich ergeht es dem probabilistischen Ansatz im Falle von »Preemption«-Beispielen. Werfen zwei Steinewerfer auf eine Flasche und nur der erste trifft, aber beide erhöhen die Wahrscheinlichkeit für das Zerspringen der Flasche, so wird auch der zweite Wurf für kausal relevant eingestuft. Da der zweite Werfer aber nicht trifft, stellt er keine Ursache des Zerspringens der Flasche dar. Doch die probabilistische Theorie kann diese Prozesse nicht korrekt nachbilden. Selbst wenn wir zu Einzelfallwahrscheinlichkeiten übergingen, würde das Beispiel bestehen bleiben, wenn unsere Welt indeterministisch wäre. Der zweite Werfer könnte die konkrete Zerspringenswahrscheinlichkeit der Flasche tatsächlich erhöhen, ohne eine Ursache ihres Zerspringens zu sein.

Solche Preemption Fälle sind natürlich auch für den kontrafaktischen Ansatz ein ständiges Problem (vgl. Paul & Hall 2013), den wir hier nicht ausführlich behandelt haben. Er trägt zum kausalen Schließen vor allem dadurch bei, dass er die Grundidee für das Design kontrollierter Experimente liefert. Somit kommen alle drei Ansätze zum kausalen Schließen in der Praxis zum Einsatz und geben auch für unterschiedliche Situationen hilfreich Verfahren dafür ab, zeigen aber in bestimmten Fällen auch gewisse Schwächen. Wenn sie jedoch mit Bedacht eingesetzt werden, sollten die Schwächen aber nicht zu kausalen Fehlschlüssen führen. Wir haben somit spezielle Induktionsverfahren zum kausalen Schließen vorgestellt, die für alle anderen Induktionsverfahren eine besondere Bedeutung besitzen, denn wir waren zu Beginn des Buches (in Kapitel 1) schon dafür eingetreten, dass wir beim induktiven Schließen immer schon bestimmte kausale Zusammenhänge mit erschließen müssen.

7.5 Resümee

Das Kapitel zum kausalen Schließen rundet unsere Debatte der Grundlagen des induktiven Schließens ab. Es ergaben sich ähnliche Ergebnisse

wie in den vorhergehenden Kapiteln. Zunächst gibt es keinen einfachen Algorithmus, der von Daten direkt zu kausalen Hypothesen führt. Wir sind bei solchen Schlüssen zumindest immer schon auf gewisse kausale Ausgangsannahmen angewiesen. Das gilt letztlich für alle untersuchten Ansätze. Außerdem ließ sich wieder die Struktur des Schlusses auf die beste Erklärung erkennen. Ursachen sind Unterschiedsmacher und wir suchen speziell nach solchen Unterschieden, die letztlich nur durch eine bestimmte Kausalhypothese erklärbar sind.

Das Verfahren ist dabei als holistisch anzusehen, weil wir immer schon auf weitere kausale Annahmen etwa über die überhaupt relevanten kausalen Faktoren oder die Markov-Eigenschaft und die Graphentreue unserer Netze angewiesen sind. Trotzdem finden wir dann auch Verfahren wie das Tabellenverfahren der minimalen Theorie oder die Schlussverfahren für bayesianische Netze, die jeweils den Charakter lokaler Algorithmen aufweisen, die also in bestimmten Situationen durchaus klare und intuitive Regeln für das induktive Schließen formulieren können. Diese einfachen Schlüsse auf kausale Zusammenhänge sind besonders grundlegend, denn wir haben bereits in den früheren Kapiteln gesehen, dass alle induktiven Schlüsse immer auch auf kausales Hintergrundwissen angewiesen sind.

Einige Notationen

Begriffe der Bestätigung

- P(A|B) bezeichnet die Wahrscheinlichkeit von A unter der Annahme, dass B wahr ist.
- B(H:E) bedeutet, dass die Hypothese H durch das Datum E bestätigt wird.
- B(H:E;K) bedeutet, dass die Hypothese H durch das Datum E bestätigt wird, gegeben ein bestimmtes Hintergrundwissen K.
- B(H:E,E*) bedeutet, dass die Hypothese H durch das Datum E besser bestätigt wird als durch das Datum E* (entsprechend: B(H:E,E*;K))
- P(A) ist die Wahrscheinlichkeit dafür, dass die Aussage A wahr ist
- P(X=x) = P(x) ist die Wahrscheinlichkeit dafür, dass die Zufallsvariable X den Wert x annimmt.

Als logische Zeichen verwende ich:

- & als Konjunktionszeichen »und« und ∨ als einschließendes »oder«
- ∀ (Allquantor) »für alle«
- ∃ (Existenzquantor) »es gibt ein«
- ≡ (logische Äquivalenz) »ist genau dann wahr, wenn«
- ⇒ (semantische Folgerung) »daraus folgt«
- ⇛ soll eine Ursache-Wirkungsbeziehung ausdrücken
- {x; Fx} für die Menge aller x, die die Eigenschaft F erfüllen
- $Ha \equiv \phi a \equiv Fa \rightarrow Ga$ und für abzählbar unendlich viele a_i schreiben wir $\phi a_i \equiv Fa_i \rightarrow Ga_i$ und kürzen dann ab: $\phi_n \equiv Fa_n \rightarrow Ga_n$ und $\phi^n \equiv \forall i(i \in \{1,\ldots,n\} \rightarrow (Fa_n \rightarrow Ga_n)) \equiv \&_{i \leq n}(Fa_i \rightarrow Ga_i)$ oder oft $H \equiv \phi \equiv \forall i(Fa_i \rightarrow Ga_i)$ oder wir schreiben dafür auch $\forall x(Fx \rightarrow Gx)$. Und dazu finden wir einige grundlegende Formeln:
- $P(\phi^n) = P(\phi_1)\ P(\phi_2|\ \phi_1)\ P(\phi_3|\phi_2 \& \phi_1) \ldots P(\phi_n|\phi^{n-1})$ (Kettenregel: folgt aus der allgemeinen Multiplikationsregel)
- Für statistisch unabhängige ϕ_i gilt: $P(\phi^n) = P(\phi_1)\ P(\phi_2) \ldots P(\phi_n)$

- Wir verlangen für Induktionsverfahren die *Induktionseigenschaft*: (IE) $P(\phi_n|\phi^{n-1}) > P(\phi_n)$, die besagt, dass die Wahrscheinlichkeit für ϕa_n wächst, wenn wir schon wissen, dass die vorigen a_i die Eigenschaft ϕ aufweisen.
- Oder: (IE) $P(\phi_n|\phi^{n-1}) \rightarrow 1$ für $n \rightarrow \infty$.
- Argument dazu: Aus $P(\phi)>0$ folgt mit der Multiplikationsformel, dass die $P(\phi_n|\phi^{n-1})$ schnell gegen 1 gehen müssen, da sonst $P(\phi)=0$ folgen würde, denn es gilt $P(\phi) = \lim_{n\rightarrow\infty} P(\phi^n) = \lim_{n\rightarrow\infty} P(\phi_1)\ P(\phi_2|\ \phi_1)\ P(\phi_3|\phi_2\&\phi_1)\ldots P(\phi_n|\phi^{n-1})$ aufgrund der σ-Additivität von P.

Literatur

Achinstein, Peter (1983): *The Nature of Explanation*, New York: Oxford University Press.

Achinstein, Peter (2001): *The Book of Evidence*, New York: Oxford University Press.

Achinstein, Peter (2004): A Challenge to Positive Relevance Theorists: Reply to S. Roush, in: *Philosophy of Science* **71**, 521–524.

Adler, Jonathan (2006): Epistemological Problems of Testimony, *Stanford Encyclopedia of Philosophy*.

Akaike, Hirotugu (1973): Information theory and an extension of the maximum likelihood principle, in: B. N. Petrov (Hrsg.) u.A.: *Proceedings of the Second International Symposium on Information Theory*, Budapest: Akademiai Kiado, 267–281.

Albert, Max (1992): Die Falsifikation statistischer Hypothesen, in: *Journal for General Philosophy of Science* **23**, 1–32.

Aliseda, Atocha (2006): *Abductive Reasoning: Logical Investigations into the Processes of Discovery and Explanation*, Synthese Library. Kluwer Academic Publishers. Dordrecht.

Ambuehl, M. & Baumgartner, M. & Epple, R. & Kauffmann, A. & Thiem, A. (2015): *cna: A package for Coincidence Analysis (CNA)*. http://cran.r-project.org/package=cna, R package version 1.0-2, 2015.

Arntzenius, Frank & Elga, Adam & Hawthorne, John (2004): Bayesianism and Binding, *Mind* **113** No 450, 251–283.

Arntzenius, Frank (2010): Reichenbach's Common Cause Principle, *The Stanford Encyclopedia of Philosophy* (Fall 2010 Edition). URL = <http://plato.stanford.edu/archives/fall2010/entries/physics-Rpcc/>

Balzer, Wolfgang & Moulines, Carlos Ulises & Sneed, John D. (1987): *An Architectonic for Science*, Dordrecht: Reidel.

Bartelborth, Thomas (1987): *Eine logische Rekonstruktion der klassischen Elektrodynamik*, Frankfurt a.M.: Verlag Peter Lang.

Bartelborth, Thomas (1989): Kann es rational sein, eine inkonsistente Theorie zu akzeptieren? in: *Philosophia Naturalis* **26**, 91–120.

Bartelborth, Thomas (1994): Modelle und Wirklichkeitsbezug, in: H.J. Sandkühler (Hrsg.), *Theorien, Modelle und Tatsachen. Konzepte der Philosophie und der Wissenschaften*, Frankfurt a.M.: Peter Lang, 145–154.

Bartelborth, Thomas (1996): *Begründungsstrategien. Ein Weg durch die analytische Erkenntnistheorie*, Akademie Verlag.

Bartelborth, Thomas (1996a): Der Schluß auf die beste Erklärung, in: Hubig, Christoph & Poser, Hans (Hrsg.), *Cognitio Humana – Dynamik des Wissens und der Werte*, Leipziger Universitätsverlag, 552–559.

Bartelborth, Thomas (2001): A-priori-Rechtfertigung und Skeptizismus, in: Thomas Grundmann (Hrsg.), *Erkenntnistheorie*, Paderborn: Mentis, 109–124.

Bartelborth, Thomas (2002): Explanatory Unification, in: *Synthese* **130**: 91–207.

Bartelborth, Thomas (2004): Wofür sprechen die Daten? in: *Journal for General Philosophy of Science* **35**: 13–40.

Bartelborth, Thomas (2005): Is the Best Explaining Theory the Most Probable one?, in: *Grazer Philosophische Studien* **70**, 1–23.

Bartelborth, Thomas (2006): Zum Einsatz formaler Methoden in der analytischen Philosophie, in: Ernst, G.& Niebergall, K.G. (Hrsg.) *Philosophie der Wissenschaft – Wissenschaft der Philosophie* (Festschrift für C.Ulises Moulines zum 60. Geburtstag), Paderborn: Mentis, 13–30.

Bartelborth, Thomas (2007): *Erklären*, Berlin-New York: Walter de Gruyter.

Bartelborth, Thomas (2008): Dimensionen der Erklärungsstärke in modernen Erklärungstheorien, in: *Philosophia Naturalis* **45**, Heft 2, 139–166.

Bartelborth, Thomas (2011): Propensities and Transcendental Assumptions. *Erkenntnis* **74** (3):363–381.

Bartelborth, Thomas (2013): Sollten wir klassische Überzeugungssysteme durch bayesianische ersetzen?, *Logos* **3**, 2–68.

Bartelborth, Thomas (2015): Eine moderate, empirische Kohärenztheorie, *Zeitschrift für philosophische Forschung* **69** (1), 5–25.

Bartelborth, Thomas (2016): How Strong is the Confirmation of a Hypothesis by Significant Data? *Journal for General Philosophy of Science* **47**, 277-291, DOI: 10.1007/s10838-016-9341-0.

Bartels, Andreas & Stöckler, Manfred (Hrsg.) (2009): *Wissenschaftstheorie*, Paderborn: Mentis.

Baumgartner, Michael & Grasshoff, Gerd (2004): *Kausalität und kausales Schliessen. Eine Einführung mit interaktiven Übungen*, Bern Studies in the History and Philosophy of Science, Bern.

Baumgartner, Michael (2007): *Complex Causal Structures. Extensions of a Regularity Theory of Causation*, PhD-thesis, University of Bern.

Baumgartner, Michael (2008a): The Causal Chain Problem, *Erkenntnis*, **69**, 201–226.

Baumgartner, Michael (2008b): Regularity Theories Reassessed, *Philosophia* 36, 327–354.

Baumgartner, Michael (2009a): Inferring Causal Complexity, *Sociological Methods & Research* **38**, 71–101.

Baumgartner, Michael (2009b): Uncovering Deterministic Causal Structures: A Boolean Approach, *Synthese* **170**, 71–96.

Baumgartner, Michael (2009c): An Interventionist Exclusion Argument, *International Studies in the Philosophy of Science*, 23, 161–178.

Baumgartner, Michael (2009d): Interdefining Causation and Intervention, *Dialectica*, 63, 175–194.

Baumgartner, Michael (2013): A Regularity Theoretic Approach to Actual Causation, *Erkenntnis* **78**, 85–109.

Baumgartner, Michael & Epple, Rüdiger (2014), A Coincidence Analysis of a Causal Chain: The Swiss Minaret Vote, *Sociological Methods & Research* **43**, 280–312.

Beck-Bornholdt, Hans Peter & Dubben, Hans Hermann (1998): *Der Hund, der Eier legt. Erkennen von Fehlinformationen durch Querdenken*, Hamburg: Rowohlt.

Beck-Bornholdt, Hans Peter & Dubben, Hans Hermann (2003): *Der Schein der Weisen. Irrtümer und Fehlurteile im täglichen Denken*, Hamburg: Rowohlt.

Beck-Bornholdt, Hans Peter & Dubben, Hans Hermann, (1996): Is the pope an alien, *Nature* **381**, 730.

Beierle, Christoph & Kern-Isberner, Gabriele (2008), *Methoden wissensbasierter Systeme*, Vieweg + Teubner.

Bem, Daryl J. (2011): Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect, *Journal of Personality and Social Psychology*, **100**, 100–119.

Bem, Daryl J. & Utts, J., & Johnson, W. O. (2011): Reply: Must Psychologists Change the Way They Analyze Their Data? *Journal of Personality and Social Psychology*, **100**, 716–719.

Bernardo, J. M. (2009): Statistics: Bayesian methodology in statistics. in: *Comprehensive Chemometrics* (S. Brown, R. Tauler and R. Walczak eds.) Oxford: Elsevier, 213–245

Bird, Alexander (2002): On whether some laws are necessary, in: *Analysis* **62**, 257–270.

Bird, Alexander (2004): Antidotes All the Way Down?, in: *Theoria* **19**, 259–269.

Bird, Alexander (2005): Explanation and Metaphysics , in: *Synthese* **143**, 89–107.

Bird, Alexander (2006): Abductive Knowledge and Holmesian Inference in *Oxford Studies in Epistemology* (eds. Tamar Szabo Gendler and John Hawthorne) Oxford: Oxford University Press, 1–31.

Bird, Alexander (2007): *Nature's Metaphysics: Laws and Properties*, Oxford: Oxford University Press.

Bird, Alexander (2010): Eliminative Abduction—Examples from Medicine, *Studies in the History and Philosophy of Science* **41**, 345–352.

Bird, Alexander (2013): Thomas Kuhn, *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (Hrsg.), URL = <http://plato.stanford.edu/archives/fall2013/entries/thomas-kuhn/>.

Martijn Blaauw (Hrsg.) (2013): *Contrastivism in Philosophy*, London: Routledge.

BonJour, Laurence (1985): *The Structure of Empirical Knowledge*, Harvard University Press.

BonJour, Laurence (1998): *In Defense of Pure Reason*, Cambridge University Press.

Bortz, Jürgen & Döring, Nicola, (2006): *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*, Berlin-Heidelberg: Springer.

Bovens, Luc & Hartmann, Stephan (2006): *Bayesianische Erkenntnistheorie*, Paderborn: Mentis Verlag.

Bradley, Richard (2005): Radical Probabilism and Bayesian Conditioning. *Philosophy of Science* **72**, 342–64.

Brössel, Peter (2014): Assessing Theories: The Coherentist Approach, *Erkenntnis* **79**, 593–623.

Büchter, Andreas & Henn, Hans-Wolfgang (2007): *Elementare Stochastik*, Springer Verlag: Berlin.

Burch, P. (1983): The Surgeon General's Epidemologic Criteria for Causality. A Critique, *Journal of Chronic Diseases* **37**, 148–157.

Bundschuh, M. & Newman, M. C. & Zubrod, J. P. & Seitz, F. & Rosenfeldt, R. R. & Schulz, R. (2013): Misuse of Null Hypothesis Significance Testing: Would Estimation of Positive and Negative Predictive Values Improve Certainty of Chemical Risk Assessment?, *Environmental Science and Pollution Research*, DOI 10.1007/s11356-013-1749-z.

Burnham, Kenneth P. & Anderson, David R. (2002): *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer-Verlag: New York

Burnham, Kenneth P. & Anderson, David R. (2004): Multimodel Inference: Understanding AIC and BIC in Model Selection, in: *Sociological Methods and Research*, **33**, 261–304

Campbell, Scott & Franklin, James (2004): Randomness And the Justification of Induction, *Synthese* **138**, 79–99.

Carnap, Rudolf (1950): *Logical Foundations of Probability*, Chicago: University of Chicago Press.

Carnap Rudolf (1959): *Induktive Logik und Wahrscheinlichkeit*. Wien: Springer.

Carnap, Rudolf (1971): A basic system of inductive logic, part I. In: Carnap and Jeffrey (1971), 33–165.

Carnap, Rudolf (1980): A basic system of inductive logic, part II. In: *Studies in Inductive Logic and Probability*, ed. Richard C. Jeffrey, University of California Press, vol. 2, 7–155.

Carnap, Rudolf and Jeffrey, Richard, eds. 1971. *Studies in Inductive Logic and Probability*, vol. 1. Berkeley: University of California Press.

Carnap, Rudolf, (1968): *Einführung in die Philosophie der Naturwissenschaften*, Ullstein Materialien.

Carrier, Martin (2006): *Wissenschaftstheorie zur Einführung*, Hamburg: Junius Verlag.

Cartwright, Nancy (1989): *Nature's Capacities and their Measurement*, Oxford: Oxford University Press, Oxford.

Cederlof, R., Friberg, L. & Lundman, T. (1972): The interactions of smoking, environment and heredity and their implications for disease etiology, *Acta Med Scand* 612, Suppl.

Chakravartty, Anjan (2007): *A Metaphysics for Scientitic Realism. Knowing the Unobservable*, Cambridge: Cambridge University Press.

Chalmers, Allan F. (1994): *Wege der Wissenschaft*, Berlin: Springer.

Chandler, Jake (2013): Contrastive Support: Some Competing Accounts, **Synthese 190** (1), 129–138.

Childers, Timothy (2013): *Philosophy & Probability*, Oxford: Oxford University Press.

Cheng, Patricia (1997): From Covariation to Causation: A Causal Power Theory, *Psychological Review* 104: 367–405.

Christensen, David (1991): Clever Bookies and Dutch Strategies, *Philosophical Review* **100**, 229–47.

Christensen, David (1999): Measuring confirmation. *Journal of Philosophy* **96**, 437–461.

Collins, John (2006): Counterfactuals, Causation, and Preemption", in Dale Jacquette (ed.) *Handbook of the Philosophy of Logic*, North Holland Press (2006) 1019–1035. Volume 5 in the Handbook of the Philosophy of Science series, edited by Dov Gabbay, Paul Thagard, and John Woods.

Cornfield, J. & Haenszel, W. & Hammond, E. & Lilienfeld, A. & Shimkin, M. & Wynder, E. (1959): Smoking and Lung Cancer: Recent evidence and a discussion of some questions, *Journal of the National Cancer Institut* 22, 173–203.

Cox, Richard T. (1946): "Probability, frequency, and reasonable expectation," *Am. Jour. Phys.* **14**, 1–13.

Craver, Carl F. (2009): Mechanisms and natural kinds, *Philosophical Psychology* **22** (5), 575–594.

Crupi, V. & Tentori, K. & Gonzalez, M. (2007): On Bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science*, **74**, 229–252.

Crupi, Vincenzo & Tentori, Katya (2012): A Second Look at the Logic of Explanatory Power (with Two Novel Representation Theorems), *Philosophy of Science* **79**, 365–385.

de Finetti, Bruno (1937): La Prévision: ses lois logiques, ses sources subjectives, *Annales de l'Institut Henri Poincaré* **7**, 1–68, Übersetzung: Foresight: Its Logical Laws, Its Subjective Sources, in H. E. Kyburg and H. E. Smokler (eds), *Studies in Subjective Probability*, New York: Wiley, 1964.

de Finetti, Bruno (1974): *Theory of Probability*, volume 1. Wiley, New York.

Diaconis, Persi & Zabell, Sandy L. (1982): Updating Subjective Probability, *Journal of the American Statistical Association*, Vol. 77, No. 380. (Dec., 1982), pp. 822–830.

Diez, David & Barr, Christopher & Cetinkaya-Rundel, Christopher (2012): *Openintro Statistics. Second Edition*, Open Textbook Library. (http://www.openintro.org/)

Doll, R. and A.B. Hill, (1950): Smoking and carcinoma of the lung. Preliminary report, *British Medical Journal*, 739–48,

Doll, R. and A.B. Hill, (1952): A study of the aetiology of carcinoma of the lung, *British Medical Journal*, 1271–1286,

Doll, R. and R. Peto, (2003): *Epidemiology of Cancer. Oxford Textbook of Medicine*, ed. D. Warrell, et al. Oxford: Oxford University Press.

Douven, Igor, & Meijs, Wouter (2007): Measuring coherence, *Synthese* **156**, 405–25.

Duhem, P., 1978, *Ziel und Struktur der physikalischen Theorien*, Hamburg: Felix Meiner Verlag.

Eagle, Antony (2004): „Twenty-One Arguments Against Propensity Analyses of Probability", in: *Erkenntnis* 60: 371–416.

Earman, John (1986): *A Primer on Determinism*, Dordrecht: Reidel.

Earman, John (1992): *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, Cambridge, MA: MIT Press.

Edwards, Anthony W.F. (1992) (1. Auflage 1972): *Likelihood*, John Hopkins University Press.

Elga, Adam (2000): Self-locating Belief and the Sleeping Beauty Problem, *Analysis* **60** (2): 143–147.

Ellis, Brian and Lierse, Caroline, (1994): Dispositional Essentialism, *Australasian Journal of Philosophy* **72**: 27–45.

Ellis, Brian, (2001): *Scientific Essentialism,* Cambridge Studies in Philosophy. Cambridge: Cambridge University Press.

Eriksson, Lina & Hájek, Alan (2007): What are degrees of belief? *Studia Logica* **86**, 185–215.

Esfeld, Michael (2007): Kausalität, in Andreas Bartels & Manfred Stöckler (eds.): *Wissenschaftstheorie. Ein Studienbuch*, Paderborn: Mentis, 89–107.

Esfeld, Michael (2008): Die Metaphysik dispositionaler Eigenschaften, *Zeitschrift für philosophische Forschung* **62**, 323–342.

Esfeld, Michael (2008a): *Naturphilosophie als Metaphysik der Natur*. Frankfurt (Main): Suhrkamp.

Esfeld, Michael (2010): Physics and causation, *Foundations of Physics* **40**, 1597–1610.

Esfeld, Michael, (2010a): Humean metaphysics versus a metaphysics of powers, in: Gerhard Ernst & Andreas Hüttemann (eds.): *Time, chance and reduction. Philosophical aspects of statistical mechanics*, Cambridge: Cambridge University Press, 119–135.

Esfeld, Michael & Sachse, Christian (2010): *Kausale Strukturen. Einheit und Vielheit in der Natur und den Naturwissenschaften*, Frankfurt (Main): Suhrkamp.

Esfeld, Michael (2011^2): *Einführung in die Naturphilosophie*, Darmstadt: Wissenschaftliche Buchgesellschaft.

Fahrmeir, Ludwig & Kneib, Thomas & Lang, Stefan, (2009): *Regression. Modelle, Methoden und Anwendungen*, Springer-Verlag Berlin Heidelberg.

Feller, William (1970): *An introduction to probability theory and its applications*: Vol. 1 (2nd ed.). New York: Wiley.

Fishburn, Peter C. (1986): The Axioms of Subjective Probability, *Statistical Science* **1**, 335–358.

Fisher, Ronald Aylmer (1935). The Logic of Inductive Inference, *Journal of the Royal Statistical Society* **98**, 39–82.

Fisher, Ronald Aylmer (1951): *The Design of Experiments*, Edingburgh: Oliver and Boyd.

Fisher, Ronald Aylmer (1958): *Statistical methods for research workers*, (13th ed.), New York: Hafner.

Fisher, Ronald Aylmer (1959): *Smoking. The Cancer Corntroversy*, Edingburgh: Oliver and Boyd.

Fitelson, Branden (1999): The plurality of bayesian measures of confirmation and the problem of measure sensitivity. Philosophy of Science 66 (Proceedings): S362–S378.

Fitelson, Branden (2001): *Studies in Bayesian confirmation Theory*. Ph.D. thesis, University of Wisconsin, Madison. URL http://Fitelson.org/thesis.pdf.

Fitelson, Branden (2002): Putting the Irrelevance Back Into the Problem of Irrelevant Conjunction , *Philosophy of Science* **69** (4), 611–622.

Fitelson, Branden (2003): A probabilistic theory of coherence, *Analysis*, 63 , 194–199.

Fitelson, Branden (2007): Likelihoodism, Bayesianism, and relational confirmation, *Synthese* **156**, 473–489, DOI: 10.1007/s11229-006-9134-9.

Fitelson, Branden, (2013): Contrastive Bayesianism, in: Blaauw (2013), 64-87.

Fitelson, Branden & Hawthorne, James (2010): How Bayesian Confirmation Theory Handles the Paradox of the Ravens, in: Eells and Fetzer (Hrsg.), *The Place of Probability in Science*, Boston Studies in the Philosophy of Science **284**, 247–275.

Fitelson, Branden & Hitchcock, Christopher (2011): Probabilistic Measures of Causal Strength, in: *Causality in the Sciences*, P. Illari, F. Russo and J. Williamson (Hrsg.), Oxford University Press, 600–627.

Forster, Malcolm R. (1988): Unification, Explanation, and the Composition of Causes in Newtonian Mechanics. *Studies in the History and Philosophy of Science* **19**, 55–101.

Forster, Malcolm and Sober, Elliott (1994): "How to Tell when Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions", *British Journal for the Philosophy of Science* **45**: 1–35.

Forster, Malcolm (2000): Key Concepts in Model Selection: Performance and Generalizability, *Journal of Mathematical Psychology* **44**, 205–231.

Franklin, James, (2001): Resurrecting logical probability, *Erkenntnis* **55**, 277–305.

Freedman, D.A. & Pisani, R. & Purves, R. (1983): *Statistics*, W.W. Norton & Company, New York.

Freedman, David A.: (1999): From association to causation: Some remarks on the history of statistics, *Statistical Science* **14**, 243–258.

Fröhlich, Gerhard, (2003): Anonyme Kritik: Peer Review auf dem Prüfstand der Wissenschaftsforschung, *medizin – bibliothek – information* · **3** · Nr 2 · Mai 2003, 33–39.

Gabbay, D., & Woods, J. (2005): *The Reach of Abduction: Insight and Trial* (A Practical Logic of Cognitive Systems Vol. 2), North-Holland, Amsterdam.

Gähde, Ulrich & Stegmüller, Wolfgang (1988): An Argument in favor of the Duhem-Quine-Thesis: from the structuralist point of view', in: *The philosophy of W. V. Quine* / ed. by Lewis Edwin Hahn and Paul Arthur Schilpp. - 3. pr. - La Salle, Ill. : Open Court: 116–136.

Gähde, Ulrich (1983): *T-Theoretizität und Holismus*, Frankfurt a.M.: Peter Lang.

Gaifman, Haim & Snir, Marc (1982). Probabilities Over Rich Languages, Testing and Randomness, *Journal of Symbolic Logic* **47** (3):495–548.

Galak, Jeff & LeBoeuf, Robyn A. & Nelson, Leif D. & Simmons, Joseph P. (2012): Correcting the past: Failures to replicate psi, *Journal of Personality and Social Psychology* **103**, 933–948.

Garber, Daniel (1980): Discussion: Field and Jeffrey conditionalization, *Philosophy of Science* **47**, 142–145.

Gemes, Ken (1998): Hypothetico-Deductivism: The Current State of Play; The Criterion of Empirical Significance Endgame, *Erkenntnis*, **49**, 1–20.

Genschel, Ulrike & Becker, Claudia (2005): *Schließende Statistik*, Berlin-Heidelberg: Springer.

Gigerenzer, Gerd (2002): *Das Einmaleins der Skepsis. Über den richtigen Umgang mit Zahlen und Risiken*, Berlin Verlag.

Gigerenzer, Gerd (2004): Mindless Statistics, *Journal of Behavioral and Experimental Economics*, **33**, 587–606.

Gillies, Donald (2000): Varieties of Propensity, *British Journal for the Philosophy of Science*, **51**, 807–835.

Gilovich, Thomas (1991): *How We Know What Isn´t So. The Fallibility of Human Reason in Everyday Life*, The Free Press New York.

Gintis, Herbert (2009): *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*, Princeton University Press.

Glass, D. H. (2002): Coherence, explanation and Bayesian networks. In M . O 'Neill, & R . F. E Sutcliffe et al. (Eds.), *Artificial intelligence and cognitive science*. Berlin, Heidelberg, New York: Springer-Verlag , Lecture Notes in Artificial Intelligence, **2464**, 177–182.

Glass, D. H. (2007): Coherence Measures and Inference to the Best Explanation. *Synthese* **157**, 275–296.

Goldman, Alvin (1986): *Epistemology and Cognition*, Cambridge MA.: Harvard University Press.

Good, Irving John (1983): *Good Thinking: The Foundations of Probability and Its Applications*. Univ. of Minn. Press, Minneapolis.

Gottwald, Siegfried (1993): *Fuzzy Sets and Fuzzy Logic. Foundations of Application – from a Mathematical Point of View*, Wiesbaden: Vieweg.

Gottwald, Siegfried (2010): Many-Valued Logic, *The Stanford Encyclopedia of Philosophy* (Spring 2010 Edition), Edward N. Zalta (ed.), URL =<http://plato.stanford.edu/archives/spr2010/entries/logic-manyvalued/>

Greaves, H. & Wallace, D. (2006): Justifying conditionalization: Conditionalization maximizes expected epistemic utility, *Mind* **115**, 607–632.

Greco, Daniel (2011): Significance Testing in Theory and Practice, *Brit. J. Phil. Sci.* **62**, 607–637.

Grundmann, Thomas (2008): *Analytische Einführung in die Erkenntnistheorie*, Berlin: de Gruyter.

Haack, Susan (1993): *Evidence and Inquiry. Towards Reconstruction in Epistemology,* Blackwell.

Hagen, R. L. (1997): In Praise of the Null Hypothesis Statistical Test, *American Psychologist*, **52**(1), 15–24.

Hájek, Alan (2003): What conditional probability could not be. *Synthese* **137**, 273–323.

Hájek, Alan, (2009): Fifteen Arguments Against Hypothetical Frequentism, *Erkenntnis* **70**, 211–235.

Hájek, Alan, (2010): Interpretations of Probability, *The Stanford Encyclopedia of Philosophy* (Spring 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2010/entries/ probability-interpret/>.

Hall, Ned (2007): Structural equations and causation, *Philosophical Studies*, **132**, 109–136.

Halliday, David & Resnick, Robert & Walker, Jearl (2003): Physik, Weinheim : Wiley-VCH.

Halpern, Joseph Y. (2001): Conditional plausibility measures and Bayesian networks, *Journal of AI Research* 14, 2001, pp. 359–389.

Halpern, Joseph Y. (2001a): Plausibility Measures: A General Approach For Representing Uncertainty, *Proceedings of the 17th International Joint Conference on AI* (IJCAI 2001), pp. 1474–1483.

Halpern, Joseph Y. (2003): Reasoning About Uncertainty. MIT Press.

Halpern, Joseph Y. & Pearl, Judea (2005): Causes and Explanations: A Structural-Model Approach. Part I: Causes & Part II: Explanations, in: *British Journal for the Philosophy of Science* **56** (4), I: 843–887 & II: 889–911.

Halpern, Joseph Y. & Pucella, Ricardo (2006): A Logic for Reasoning about Evidence, *Journal of Artificial Intelligence Research* **25**.

Hansson, Sven Ove (2006) Logic of Belief Revision, The Stanford Encyclopedia of Philosophy (Spring 2009 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2009/entries/logic-belief-revision/>.

Hawthorne, James (1993): Bayesian Induction Is Eliminative Induction : *Philosophical Topics*, **21**, no. 1, pp. 99–138.

Hawthorne, James (2005), Degree-of-Belief and Degree-of-Support: Why Bayesians Need Both Notions, *Mind* **114** (454), 277–320.

Hawthorne, James, (2009): The Lockean Thesis and the Logic of Belief, in: Franz Huber and Christoph Schmidt-Petri (eds.), *Degrees of Belief*, Synthese Library 342, 49–74.

Hawthorne, James (2011a): Inductive logic. *The Stanford Encyclopedia of Philosophy*, URL http://plato.stanford.edu/entries/logic-inductive/.

Hawthorne, James (2011b): Bayesian Confirmation Theory : in S. French and J. Saatsi (eds.), *The Continuum Companion to the Philosophy of Science*, 2011, London: London: Continuum International Publishing Group, 197–213.

Hawthorne, James (2011c): Confirmation Theory : in Prasanta S. Bandyopadhyay and Malcolm Forster (eds.), *Philosophy of Statistics*: *Handbook of the Philosophy of Science*, Volume 7, Elsevier, 333–389.

Hawthorne, James & Fitelson, Branden (2004): Re-solving Irrelevant Conjunction with Probabilistic Independence, *Philosophy of Science*, **71**, no. 4, 505–514.

Held, Leonhard (2008): *Methoden der statistischen Inferenz. Likelihood und Bayes,* Spektrum Akademischer Verlag.

Hempel, Carl Gustav (1943): A Purely Syntactical Definition of Confirmation. *The Journal of Symbolic Logic* **8**.

Hempel, Carl Gustav (1945): Studies in the Logic of Confirmation, *Mind* **54**, 1–26, 97–121.

Hempel, Carl Gustav (1965): *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: Free Press.

Hitchcock, Christopher (2010): Probabilistic Causation, *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <http://www.science.uva.nl/~seop/entries/causation-probabilistic/>

Hitchcock, Christopher & Woodward, James (2003): Explanatory Generalizations, Part 2: Plumbing Explanatory Depth, in: *Nous* **37**: 181–99.

Hitchcock, Christopher (2001): The intransitivity of causation revealed in equations and graphs, *Journal of Philosophy* **98**, 273–299.

Hitchcock, Christopher (2009): Structural equations and causation: Six counterexamples, *Philosophical Studies*, **144**, 391–401.

Hoefer, Carl (2007): The Third Way on Objective Probability: A Sceptic's Guide to Objective Chance *Mind*, **116(463)**, 549–596.

Howson, Colin & Urbach, Peter (1993): *Scientific Reasoning: The Bayesian Approach*, La Salle, IL: Open Court, 2nd edition.

Howson, Colin & Franklin, Allan (1992): Bayesian Conditionalization and Probability Kinematics, *Brit. J. Phil. Sci.* **45**, 451–466.

Howson, Colin (2000): *Hume's problem: induction and the justification of belief.* Oxford University Press, Oxford.

Huber, Franz (2011): Formal Representations of Belief, *The Stanford Encyclopedia of Philosophy (Fall 2011 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2011/entries/formal-belief/>.

Hume, David (1888): *Treatise of Human Nature,* edited by L. A. Selby Bigge, Oxford, Clarendon Press. Originally published 1739–40.

Humphreys, Paul (1989): *The Chances of Explanation. Causal Explanation in the Social, Medical, and Physical Sciences*, Princeton: Princeton University Press.

Irle, Albrecht (2001): *Wahrscheinlichkeitstheorie und Statistik*, Teubner, Stuttgart.

Ioannidis, John P. A. (2005): Why Most Published Research Findings Are False, *PLoS Medicine*, **2**, 696–701.

Janzing, Dominik (2003): *Kann Statistik Ursachen beweisen?*, Vorlesungsskript (http://iaks-www.ira.uka.de/home/janzing/kausalskriptum03.ps).

Jaynes, E.T. (2003): *Probability Theory: The Logic of Science: Principles and Elementary Applications* Vol 1, Cambridge University Press.

Jeffreys, Harold (1961): *Theory of probability* (3 ed.), Oxford: Oxford University Press.

Joyce, James M. (1998): A Nonpragmatic Vindication of Probabilism. *Philosophy of Science* **65** (4): 575–603.

Joyce, James M. (2009): Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief, in: *Degrees of Belief*, F. Huber and C. Schmidt-Petri (Hrsg.), 263–97. Synthese Library 342. Dordrecht: Springer.

Joyce, James M. (2009a): The Development of Subjective Bayesianism, in: *Handbook of the History of Logic*. Volume 10: *Inductive Logic*. Herausgeber: Dov Gabbay, Stephan Hartmann and John Woods, 415–475.

Kadane, J. & Seidenfeld, T. (1990): Randomization in a Bayesian Perspective, *Journal of Statistical Planning and Inference* **25**, 329–345.

Kahneman, D., Slovic, P. und Tversky, A. (Hrsg.), 1982, *Judgement und Certainty: Heuristics and Biases*, Cambridge University Press.

Kaplan, Mark (2006): *Decision Theory as Philosophy*, Cambridge University Press.

Kelle, Udo (2003): Die Entwicklung kausaler Hypothesen in der qualitativen Sozialforschung. Methodologische Ueberlegungen zu einem häufig vernachlaessigten Aspekt qualitativer Theorie- und Typenbildung. *Zentralblatt für Didaktik der Mathematik*, 35 (2003) **6**, 232–246.

King, Jeffrey C. (2011): Structured Propositions, *The Stanford Encyclopedia of Philosophy (Fall 2011 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2011/entries/propositions-structured/>.

Klärner, Holger (2003): *Der Schluss auf die beste Erklärung*, Berlin: deGruyter.

Klein, Peter & Warfield, Ted (1994): What Price Coherence? *Analysis* **54**, 129–132.

Koehler, J. J. & Chia, A. , Lindsey, J. S. (1995): The Random Match Probability (RMP) in DNA Evidence. Irrelevant and Prejudicial? In: *Jurimetrics Journal* **35**, 201–219.

Korb, Kevin B. & Nicholson, Ann E. (2010): *Bayesian Artificial Intelligence,* Chapman & Hall/CRC Computer Science & Data Analysis.

Koscholke, Jakob (2015): Evaluating Test Cases for Probabilistic Coherence Measures, *Erkenntnis*, **81**(1), 155–181. DOI: 10.1007/s10670-015-9734-1.

Kraft, C. J. Pratt und Seidenberg, A. (1959): Intuitive Probability on Finite Sets, *The Annals of Mathematical Statistics* **30**, 408–419.

Kronz, Frederick M. (1992): Carnap and Achinstein on Evidence . In: *Philosophical Studies* **67**, 151–167.

Kuhn, Thomas S. (1976): *Die Struktur wissenschaftlicher Revolutionen*, Originaltitel: The Structure of Scientific Revolutions, Chicago 1962.

Kuhn, Thomas (1977): *The Essential Tension. Selected Studies in Scientific Tradition and Change*, Chicago: University of Chicago Press.

Kullback, Solomon (1968): *Information Theory and Statistics*, New York: Dover.

Kvart, Igal (1997): Causes and some Positive Causal Impact, *Noûs* Vol. **31**, Supplement: *Philosophical Perspectives* **11**, Mind, Causation, and World, 401–432.

Kvart, Igal (2001): The Counterfactual Analysis of Cause, *Synthese* **127**, 389–427.

Lakatos, I. (1974): Falsifikation und die Methodologie wissenschaftlicher Forschungsprogramme, in: Lakatos/Musgrave (Hrsg.) *Kritik und Erkenntisfortschritt*, 89–189, Braunschweig: Vieweg.

Landes, Jürgen & Williamson, Jon (2013): Objective Bayesianism and the Maximum Entropy Principle, *Entropy* 15(9): 3528–3591.

Laudan, Larry (1981): A confutation of convergent realism, *Philosophy of Science* **48**,19–49.

Lehrer, Keith (1974): *Knowledge*, Oxford Clarendon Press.

Leitgeb, Hannes, & Pettigrew, Richard (2010a): An Objective Justification of Bayesianism I: Measuring Inaccuracy. *Philosophy of Science*, **77** (April 2010) pp. 201–235.

Leitgeb, Hannes, & Pettigrew, Richard (2010b):. An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy. *Philosophy of Science*, **77** (April 2010) 236–272.

Leitgeb, Hannes (2014): The Stability Theory of Belief, *The Philosophical Review* **123** (2), 131–171.

Lenzen, Wolfgang (1974): *Theorien der Bestätigung wissenschaftlicher Hypothesen*, Stuttgart-Bad Cannstatt: Frommann Verlag.

Lewis, C.I. (1946): *An analysis of knowledge and valuation* Chicago: Open Court.

Lewis, David (1973): Causation, *Journal of Philosophy* **70**, 556–567.

Lewis, David (1980): A Subjectivist's Guide to Objective Chance, in Richard C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*, Volume II, Berkeley: University of California Press, 263–293.

Lewis, David (1994): Humean supervenience debugged, in: *Mind* **103**, 473–490.

Lindley, D. V. (1957): A Statistical Paradox, *Biometrika* **44**, 187–192.

Lipton, Peter (2004) (erste Auflage 1991): *Inference to the Best Explanation*, London: Routledge.

Little, Daniel (1998): *Microfoundations, Method, and Causation Essays in the Philosophy of the Social Sciences*, Transaction Publishers.

Löffler, Winfried (2002): Eine vermutlich unerwünschte Konsequenz von Swinburnes probabilistischer Gotteslehre, in: *Argument und Analyse*. Ausgewählte Sektionsvorträge des 4. Kongresses der Gesellschaft für Analytische Philosophie, Bielefeld 2000. Hg. von A. Beckermann und C. Nimtz. Paderborn: mentis 2002 (elektronische Publikation unter http://www.gap-im-netz.de/gap4Konf/Proceedings4/ Proc.htm, 474–484).

Löffler, Winfried (2006): *Einführung in die Religionsphilosophie*, Darmstadt: Wissenschaftliche Buchgesellschaft.

Lowe, E. Jonathan (2006): *The Four-Category Ontology. A Metaphysical Foundation for Natural Science*, Oxford: Clarendon Press.

MacKay, J.C. David (2005): *Information Theory, Inference, and Learning Algorithms,* Cambridge University Press.

Mackie, John Leslie (1974): *The Cement of the Universe*, Oxford University Press.

Maher, Patrick (2002): Joyce's Argument for Probabilism, *Philosophy of Science* 69, 73–81.

Maher, Patrick (2004): Probability captures the logic of scientific confirmation, in *Contemporary Debates in the Philosophy of Science*, ed. Christopher R. Hitchcock, 69–93. Blackwell.

Maher, Patrick (2006): The Concept of Inductive Probability, *Erkenntnis* **65**, 185–206.

Maher, Patrick (2010): Explication of Inductive Probability. *Journal of Philosophical Logic* **39**, 593–616.

Mayo, Deborah (1996): *Error and the Growth of Experimental Knowledge*, Chicago: University of Chicago Press.

Mayo-Wilson, Conor (2014): The Limits of Piecemeal Causal Inference, *British Journal for the Philosophy of Science* **65** (2): 213–249.

Meek, C. (1995): Causal Inference and Causal Explanation with Background Knowledge, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 403–411, Morgan Kaufmann.

Meijs, Wouter (2005): *Probabilistic measures of coherence,* (Ph.D. dissertation, University of Rotterdam).

Meijs, Wouter (2006): Coherence as generalized logical equivalence, *Erkenntnis* **64**, 231–52.

Mill, John Stuart (1865): *System of Logic*, London.

Mikkelson, J. M. (2004): Dissolving the Wine/Water Paradox, *Brit. J. Phil. Sci.* **55**, 137–145.

Moretti, Luca & Piazza, Tommaso (2013): Transmission of Justification and Warrant, The Stanford Encyclopedia of Philosophy (Winter 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2013/entries/transmission-justification-warrant/>.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2015) (im Druck): The fallacy of placing confidence in confidence intervals, *Psychonomic Bulletin & Review*.

Neapolitan, Richard E. (2004): *Learning Bayesian Networks*. Upper Saddle River, NJ: Prentice Hall.

Neapolitan, Richard E. (2009): *Probabilistic Methods for Bioinformatics*, San Francisco: Morgan Kaufmann.

Neta, Ram (2009): Defeating the Dogma of Defeasibility, in *Williamson on Knowledge*, ed. by Greenough and Pritchard (Oxford, 2009): 161–82.

Nida-Rümelin, Julian & Schmidt, Thomas (2000): *Rationalität in der praktischen Philsophie. Eine Einführung*, Akademie Verlag: Berlin.

Niiniluoto, Ilkka (1999): *Critical Scientific Realism*, Oxford University Press.

Nix, Christopher J. und Paris, Jeff B. (2006): A Continuum of Inductive Methods Arising from a Generalised Principle of Instantial Relevance, *Journal of Philosophical Logic* **35**, 83–115.

Nobles, R. & Schiff, D. (2005): Misleading statistics within criminal trials: The Sally-Clark case. *Significance* **2**, 17–9.

Norton, John D. (2003): A Material Theory of Induction, *Philosophy of Science*, **70**, 647–70.

Novack, Gregory (2010): A Defense of the Principle of Indifference, *Journal of Philosophical Logic* **39**, (6), 655–678.

Nozick, Robert (1981): *Philosophical Explanations*. Cambridge: Harvard University Press.

Oddie, Graham (2013): The content, consequence and likeness approaches to verisimilitude: compatibility, trivialization,and underdetermination, *Synthese* **190**, 1647–1687 [published online 28 May 2011, at DOI 10.1007/s11229-011-9930-8].

Oddie, Graham (2014): Truthlikeness, *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2014/entries/truthlikeness/>

Olsson, Erik J. (2002): What is the problem of coherence and truth? *The Journal of Philosophy*, **99**(5), 246–272.

Olsson, Erik J. (2005): *Against coherence*. Oxford: Oxford University Press.

Paul, L.A. & Hall, N. (2013): *Causation: A User's Guide*. Oxford: Oxford University Press.

Pearl, Judea (2000): *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York.

Pearl, Judea (2009): Causal inference in statistics: An overview, *Statistics Surveys* **3**, 96–146.

Peterson, Martin (2009): *An Introduction to Decision Theory*, Cambridge: Cambridge University Press.

Pettigrew, Richard (2011): Epistemic Utility Arguments for Probabilism, *The Stanford Encyclopedia of Philosophy* (Winter 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2011/entries/epistemic-utility/>.

Pettigrew, Richard G. (2013): Epistemic Utility and Norms for Credence, *Philosophy Compass* **8**, 897–908.

Pettigrew, Richard G. (2014): Accuracy, Risk, and the Principle of Indifference, *Philosophy and Phenomenological Research* 89(1): doi: 10.1111/phpr.12097.

Pollock, John (1974): *Knowledge and Justification*. Princeton, NJ: Princeton University Press.

Pollock, John (1994): Justification and Defeat, *Artificial Intelligence* **67**, 377–408.

Ponocny, Ivo & Ponocny-Seliger, Elisabeth (2009): Akte Astrologie Österreich. Vom Schicksal, den Sternen und der Bevölkerungsstatistik, in: *Skeptiker* 4/2009, 176.

Popper, K. R. (1957): The Propensity Interpretation of the Calculus of Probability, and the Quantum Theory', in S. Korner (ed.), *Observation and Interpretation*, Proceedings of the Ninth Symposium of the Colston Research Society, University of Bristol, 65–70, 88–9.

Popper, K. R. (1959): The Propensity Interpretation of Probability', *British Journal for the Philosophy of Science*, **10**, 25–42.

Popper, Karl R. & Miller, David W. (1983): A proof of the impossibility of inductive probability. *Nature* **302**, 687–688.

Popper, Karl Raimund (1984): *Die Logik der Forschung*, Tübingen: Mohr, Original: 1934.

Joel Predd, Robert Seiringer, Elliott H. Lieb, Daniel Osherson, Vincent Poor, and Sanjeev Kulkarni (2009) Probabilistic Coherence and Proper Scoring Rules. *IEEE Transactions of Information Theory*, **55**(10): 4786–4792.

Pritchard, Duncan (2005): *Epistemic Luck*, Oxford: Oxford University Press.

Pritchard, Duncan (2007): *Anti-Luck Epistemology*, Oxford: Oxford University Press.

Pritchard, Duncan (2008):Sensitivity, Safety, and Anti-Luck Epistemology', *The Oxford Handbook of Scepticism*, (ed.) J. Greco, 437–55, (Oxford University Press, 2008).

Putnam, Hilary (1975): *Mathematics, Matter and Method*, Cambridge: Cambridge University Press.

Quine, Willard V. Orman (1969): Natural Kinds, in: *Ontological Relativity and other Essays*, Columbia University Press, 114–138.

Quine, Willard V. Orman (1979): Zwei Dogmen des Empirismus, in: W.v.O. Quine, *Von einem logischen Standpunkt aus*, Frankfurt a.M.: Ullstein.

Quine,W. v. O. (1952): The Problem of Simplifying Truth Functions, The American Mathematical Monthly, **59**, 521–531. (1959), On Cores and Prime Implicants of Truth Functions, *The American Mathematical Monthly*, **66**, 755–760.

Ragin, Charles C. (1987): *The Comparative Method*, Berkeley: University of California Press.

Ragin, Charles C. (2000): *Fuzzy-Set Social Science*, Chicago: University of Chicago Press.

Ragin, Charles C. (2006): Set Relations in Social Research: Evaluating Their Consistency and Coverage, *Political Analysis*, **14**, 291–310.

Ramsey, Frank Plumpton [1926] (1990):. Truth and probability. In D. H. Mellor (ed.) *Philosophical Papers*. Cambridge: Cambridge University Press.

Rasch, Björn & Friese, Malte & Hofmann, Wilhelm, Naumann, Ewald (2010): *Quantitative Methoden. Einführung in die Statistik für Psychologen und Sozialwissenschaftler* (Band 1), Berlin: Springer Verlag.

Reichenbach, Hans (1938): *Experience and Prediction. An Analysis of the Foundations and the Structure of Knowledge*, Univ. of Chicago Press, Chicago. German trans. in Reichenbach (1977a), vol. 4.

Renni, John (2002): 15 Answers to Creationist Nonsense. Opponents of evolution want to make a place for creationism by tearing down real science, but their arguments don't hold up, *Scientific American Magazine* July 2002.

Rose, G. Hamilton, P. Colwell, L. & Shipley, J. (1982): A randomised controlled trial of anti-smoking advice: 10-years results, *Journal of epidemiology and Community Health* **36**, 102–108.

Rosenkrantz, Roger (1994): Bayesian confirmation: Paradise regained, *The British Journal for the Philosophy of Science* **45**, 467–476.

Rosenthal, Jacob (2004): *Wahrscheinlichkeiten als Tendenzen*. Eine Untersuchung objektiver Wahrscheinlichkeitsbegriffe, Paderborn: Mentis.

Rosenthal, Jacob (2012): Probabilities as Ratios of Ranges in Initial-State Spaces, *Journal of Logic, Language and Information* **21**, 217–236.

Roush, Sherrilyn (2004): Discussion Note: Positiv Relevance Defended, *Philosophy of Science* **71**, 110–116.

Royall, Richard (1997): *Statistical Evidence. A Likelihood Paradigm*. Boca Raton: Chapman and Hall.

Sachs, Lothar & Hedderich, Jürgen, (2006): *Angewandte Statistik*, Berlin-Heidelberg: Springer.

Salmon, Wesley C. (1975): Should we Attempt to Justify Induction?, *Philosophical Studies* **8**, 45–47.

Sasco, A. et al. (2004): Tobacco smoking and cancer: a brief review of the epidemiological evidence. *Lung Cancer* **45**, S3–S9.

Savage, Leonard J. (1954): *The Foundations of Statistics*. Wiley Publications in Statistics.

Savage, Leonard J. (1972): *The Foundations of Statistical Inference*, New York: Dover.

Schaffer, Jonathan (2003): Principled Chances, *British Journal for the Philosophy of Science* **54**, 27–41.

Schick, Frederic (1986): Dutch Bookies and Money Pumps, *Journal of Philosophy* **83**, 112–119.

Schippers, Michael (2014): Probabilistic measures of coherence: from adequacy constraints towards pluralism, *Synthese* **191**, 3821–3845, DOI 10.1007/s11229-014-0501-7.

Schoch, Daniel (2000): A fuzzy measure for explanatory coherence, *Synthese* **122**, 291 – 311.

Schupbach, Jonah & Jan Sprenger (2011): The Logic of Explanatory Power, *Philosophy of Science* **78**,105–27.

Schurz, Gerhard (1991): Relevant Deduction. From Solving Paradoxes Towards a General Theory, *Erkenntnis* **35**, 391 – 437.

Schurz, Gerhard (1994): Relevant Deduction and Hypothetico-Deductivism: A Reply to Gemes , *Erkenntnis* **41**, 183 – 188.

Schurz, Gerhard (2004): Normic Laws, Nonmonotonic Reasoning, and the Unity of Science, in: Rahman. S. et al. (eds.), *Logic, Epistemology, and the Unity of Science*, Dordrecht, Kluwer, 181–211.

Schurz, Gerhard (2005): Kuipers' Account to H-D Confirmation and Truthlikeness: From Intuitive Starting Points to Counterintuitive Consequences , in: Festa, R. (eds.), *Logics of Scientific Discovery. Essays in Debate With Theo Kuipers* (Poznan Studies in the Philosophy of Science and the Humanities), Rodopi, Amsterdam 2005, 141–159.

Schurz, Gerhard (2006): *Einführung in die Wissenschaftstheorie*, Darmstadt: Wissenschaftliche Buchgesellschaft.

Schurz, Gerhard (2008): Patterns of Abduction, *Synthese* **164**, 201–234.

Schurz, Gerhard (2008a): The Meta-Inductivist's Winning Strategy in the Prediction Game: A New Approach to Hume's Problem, *Philosophy of Science* **75**, 278–305.

Schurz, Gerhard (2014): *Philosophy of Science - A Unified Approach*, New York: Routledge.

Schurz, Gerhard (2014a): Bayesian Pseudo-Confirmation, Use-Novelty, and Genuine Confirmation, *Studies in History and Philosophy of Science* **45**, 87–96.

Schwarz, Wolfgang (2014): Proving the Principal Principle, in: Alastair Wilson (Hrsg.), Chance and Temporal Asymmetry, Oxford: Oxford University Press, 81–99.

Schweizer, Mark (2006): Intuition, Statistik und Beweiswürdigung, in: *Justice – Justiz – Giustizia* 2006/4
(http://www.decisions.ch/publikationen/intuition_statistik.html).

Scott, Dana (1964): Measurement Structures and Linear Inequalities, *Journal of Mathematical Psychology* **1**, 233–247.

Semmelweis, Ignaz (1983): *The Etiology, Concept, and Prophylaxis of Childbed Fever* (K. Codell Carter trans and ed.). Madison,WI: University of Wisconsin Press.

Shafer, G. (1976): A Mathematical Theory of Evidence. Princeton University Press.

Shogenji, T. (1999): Is Coherence Truth-conducive?, *Analysis* **59**, 338–45.

Siebel, Mark (2004): Der Rabe und der Bayesianist, *Journal for General Philosophy of Science* **35**, 313–329

Siebel, Mark (2004a): On Fitelson's measure of coherence, *Analysis* **64**, 189–190.

Snow, John (1855): *On the Mode of Communication of Cholera*, London: John Churchill.

Snyder, Laura (2006): William Whewell, *The Stanford Encyclopedia of Philosphy*, Edward N. Zalta (ed.),
URL = <http://www.science.uva.nl/~seop/entries/whewell/>

Sober, Elliott (1991): *Reconstructing the Past. Parsimony, Evolution, and Inference*, MIT Press, Cambridge/ Mass.

Sober, Elliott (2001): Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause, *British Journal for the Philosophy of Science* 52, 331–346.

Sörenson, R. (2006): Epistemic Paradoxes , *The Stanford Encyclopedia of Philosophy* (Winter 2006 Edition), Edward N. Zalta (ed.), URL = <http://www.science.uva.nl/~seop/entries/epistemic-paradoxes/>.

Sosa, Ernest (1999): How to Defeat Opposition to Moore, P*hilosophical Perspectives* **13**, 137–49.

Sosa, Ernest (2003): Relevant Alternatives, Contextualism included, *Philosophical Studies* 35–65.

Sosa, Ernest (2007): *A Virtue Epistemology. Apt Belief and Reflective Knowledge*, vol 1, Oxford: Oxford University Press.

Spirtes, P. C. & Glymour, C. & Scheines, R. (2000): *Causation, Prediction and Search* (2nd ed.) Cambridge, MA: MIT Press.

Spohn, Wolfgang (2009): A Survey of Ranking Theory, in: F. Huber & C. Schmidt-Petri (eds.), *Degrees of Belief*, Dordrecht: Springer.

Spohn, Wolfgang (2012): *The Laws of Belief. Ranking Theory & its Philosophical Applications*, Oxford: Oxford University Press.

Sprenger, Jan (2010): Hempel and the Paradoxes of Confirmation, in Dov Gabbay, Stephan Hartmann and John Woods (eds.): *Handbook of the History of Logic*, Volume **10** (Inductive Logic), 231–260. Amsterdam: Elsevier.

Sprenger, Jan (2011): Hypothetico-Deductive Confirmation. *Philosophy Compass* **6**: 497–508.

Sprenger, Jan (2014): A Synthesis of Hempelian and Hypothetico-Deductive Confirmation, *Erkenntnis* **78**: 727–738.

Staley, Kent (2014): *An Introduction to the Philosophy of Science*, Cambridge UP.

Steup, Matthias (2006): Epistemology in the Twentieth Century . Forthcoming in the Routledge *Companion to Twentieth Century Philosophy*, ed. by Dermot Moran.

Strevens, Michael (1999): Objective Probabilities as a Guide to the World. *Philosophical Studies* **95**, 243–75.

Strevens, Michael (2000): Do Large Probabilities Explain Better?, in: *Philosophy of Science* **67**, 366–90.

Strevens, Michael (2003): *Bigger than Chaos: Understanding Complexity through Probability*, Harvard University Press.

Strevens, Michael (2007): Mackie Remixed, in: Joseph Keim Campbell, Michael O'Rourke and Harry S. Silverstein (eds.), *Causation and Explanation, Topics in Contemporary Philosophy*, vol. 4, MIT Press, Cambridge, 93–118.

Suppes, Patrick (1994) Qualitative Theory of Subjective Probability. In G. Wright and P. Ayton (Eds.), *Subjective Probability*. John Wiley & Sons, 17–37.

Swinburne, Richard G. (1970): Choosing between Confirmation Theories . In: *Philosophy of Science* **37**, 602–613.

Swinburne, Richard G., (1987): *Die Existenz Gottes*, Reclam: Stuttgart (Übersetzung der ersten Auflage) (Original: Swinburne, R. (19912, 1979), The Existence of God, Oxford: Clarendon Press).

Tentori, Katya, Crupi, Vincenzo, Bonini, Nicholas & Osherson, Daniel (2007): Comparison of Confirmation Measures, *Cognition* **103**: 107–119.

Thagard, Paul (1978): Why astrology is a pseudoscience (1978) In *PSA* 1978, Volume 1, ed. Asquith PD and Hacking I (East Lansing: Philosophy of Science Association, 1978) 223 ff.

Thagard, Paul (1998): Ulcers and bacteria I: Discovery and acceptance. Studies in History and Philosophy of Science. Part C: *Studies in History and Philosophy of Biology and Biomedical Sciences*, **29**, 107–136.

Thagard, Paul (1998): Ulcers and bacteria II: Instruments, experiments, and social interactions. *Studies in History and Philosophy of Science*. Part C: Studies in History and Philosophy of Biology and Biomedical Sciences.

Thagard, Paul (1999): *How scientists explain disease*. Princeton: Princeton University Press.

Thagard, Paul (2000): *Coherence in thought and action*. Cambridge, MA: MIT Press.

Thagard, Paul (2007): Coherence, Truth, and the Development of Scientific Knowledge, *Philosophy of Science*, **74**, 28–47.

van Fraassen, Bas C. (1980): *The Scientific Image*, Oxford Clarendon Press.

van Fraassen, Bas C. (1984): Belief and the Will, *The Journal of Philosophy*, vol. LXXXI, **5**, 135–56.

von Mises, Richard (1957): *Probability, Statistics and Truth*, revised English edition, New York: Macmillan. (deutsch: von Mises, Richard (1928): *Wahrscheinlichkeit, Statistik und Wahrheit*, Wien: Springer Verlag.)

Vranas, Peter B. M. (2004): Hempel's Raven Paradox: A Lacuna in the Standard Bayesian Solution, *British Journal for the Philosophy of Science* **55** (3), 545–560.

Wacholder, S. & Chanock, S. & Garcia-Closas, M. & El ghormli, L. & Rothman, N. (2004): Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies, *Journal of the National Cancer Institute*, **96**, No. 6, 434–442.

Wagenmakers, E.-J. & Wetzels, R. & Borsboom, D. & Kievit, R. & van der Maas, H. L. J. (2011a): Yes, psychologists must change the way they analyze their data: Clarifications for Bem, Utts, & Johnson (2011) manuscript: http://dl.dropbox.com/u/1018886/Clarifications ForBemUtts Johnson.pdf

Wagenmakers, E.-J. & Wetzels, R. & Borsboom, D. & van der Maas, H. L. J. (2011): Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011), *Journal of Personality and Social Psychology* **100**, 426–432.

Wagenmakers, Eric-Jan (2007): A practical solution to the pervasive problems of p values, *Psychonomic Bulletin & Review* **14**, 779–804.

Weisberg, Jonathan (2011): Varieties of Bayesianism, in: *Handbook of the History of Logic*. Volume **10**: *Inductive Logic*. Herausgeber: Dov Gabbay, Stephan Hartmann and John Woods, 477–551.

Wetzels, R. & Matzke, D. & Lee, M. D. & Rouder, J. N. & Iverson, G. J., & Wagenmakers, E.-J. (2011): Statistical evidence in experimental psychology: An empirical comparison using 855 t tests, *Perspectives on Psychological Science* **6**, 291–298.

Wetzels, R. & Matzke, D. & Lee, M. D. & Rouder, J. N. & Iverson, G. J., & Wagenmakers, Eric-Jan (2011): Statistical evidence in experimental psychology: An empirical comparison using 855 t tests, *Perspectives on Psychological Science* **6**, 291–298.

Wheeler, Gregory & Williamson, Jon (2011): Evidential probability and objective Bayesian epistemology, in P. S. Bandyopadhyay & M. R. Forster (Hg.): Philosophy of Statistics, Handbook of the Philosophy of Science, Bd. 7, Elsevier: 307–331.

Wheeler, Gregory & Scheines, Richard (2013): Coherence and Confirmation Through Causation, *Mind* **122**, 135–70.

Whewell, William (1847): *The Philosophy of the Inductive Sciences*, Founded Upon Their History, 2nd edition, in two volumes, London.

White, Roger (2006): The generalized Sleeping Beauty problem: a challenge for thirders, *Analysis* **66** (2), 114–19.

White, Roger (2010): Evidential Symmetry and Mushy Credence, *Oxford Studies in Epistemology* Vol. 3, Tamar Szabo Gendler & John Hawthorne (Hrsg.), 161–186.

Williamson, Jon (2005): *Bayesian Nets and Causality: Philosophical and Computational Foundations*, Oxford University Press, Oxford.

Williamson, Jon (2007): Inductive influence, *British Journal for the Philosophy of Science* **58**, 689–708.

Williamson, Jon (2007a): Motivating objective Bayesianism: from empirical constraints to objective probabilities, in William L. Harper and Gregory R. Wheeler (eds.): *Probability and Inference: Essays in Honor of Henry E. Kyburg Jr.* London: College Publications, 155–183.

Williamson, Jon (2008): Objective Bayesianism with predicate languages, *Synthese* **163**(3), 341–356.

Williamson, Jon (2010): *In Defence of Objective Bayesianism*, Oxford: Oxford University Press.

Williamson, Jon (2011): Objective Bayesianism, Bayesian Conditionalisation and Voluntarism, *Synthese* 178(1), 67–85.

Williamson, Jon (2011a): An objective Bayesian account of confirmation, in Dennis Dieks, Wenceslao J. Gonzalez, Stephan Hartmann, Thomas Uebel, Marcel Weber (eds), *Explanation, Prediction, and Confirmation. New Trends and Old Ones Reconsidered*, *The philosophy of science in a European perspective* Volume 2, Springer, 53–81.

Williamson, Timothy, (2000): *Knowledge and its Limits*, Oxford: Oxford University Press.

Wilson, Fred, (2007): John Stuart Mill, *The Stanford Encyclopedia of Philosphy*, Edward N. Zalta (ed.) (URL: http://www.science.uva.nl/ ~seop/entries/mill/)

Woodward, James (1999): Causal Interpretation in Systems of Equations, *Synthese* **121**, 199–247.

Woodward, James (2000): Explanation and Invariance in the Special Sciences, in: *The British Journal for the Philosophy of Science* 51: 197–254.

Woodward, James (2003): *Making Things Happen. A Theory of Causal Explanation*, Oxford: Oxford University Press.

Woodward, James & Hitchcock, Christopher, (2003): Explanatory Generalizations, Part 1: A Counterfactual Account, in: *Nous* 37: 1–24.

Worrall, John (2007): Why there's no cause to randomize. The British *Journal for the Philosophy of Science*, 58, 451–488.

Zabell, Sandy L. (2009): Carnap and the Logic of Inductive Inference, in: *Handbook of the History of Logic*. Volume 10: *Inductive Logic*. Herausgeber: Dov Gabbay, Stephan Hartmann and John Woods, 265–309.

Zagzebski, Linda (1994): The Inescapability of Gettier Problems, *Philosophical Quarterly*, **44**, 174 (January 1994), 65–73.

Zankl, Heinrich (2003): *Fälscher, Schwindler, Scharlatane. Betrug in Forschung und Wissenschaft*, Weinheim: VCH-Verlag.

Zankl, Heinrich (2010): *Kampfhähne der Wissenschaft. Kontroversen und Feinschaften*, Weinheim: VCH-Verlag.

Index