Learning, Conditionals, Causation

Mario Konrad Günther

Graduate School of
Systemic Neurosciences

LMU Munich

Learning, Conditionals, Causation

Mario Konrad Günther

Dissertation at the
Graduate School of Systemic Neurosciences
LMU Munich

submitted by
Mario Konrad Günther
born in Laupheim

Munich, October 2, 2018

First Reviewer: Prof. DDr. Hannes Leitgeb
Second Reviewer: Prof. Dr. Stephan Hartmann

Date of Submission: October 2, 2018
Date of Defense: January 24, 2019

First Supervisor:	Prof. DDr. Hannes Leitgeb Chair of Logic and Philosophy of Language Co-Director of the Munich Center for Mathematical Philosophy LMU Munich
Second Supervisor:	Prof. Dr. Stephan Hartmann Chair of Philosophy of Science Co-Director of the Munich Center for Mathematical Philosophy LMU Munich
Third Supervisor:	Prof. Dr. Stephan Sellmaier Head of the Research Center for Neurophilosophy and Ethics of Neurosciences LMU Munich
Fourth Supervisor:	Prof. Dr. Holger Andreas Department of Economics, Philosophy, and Political Science University of British Columbia
First Reviewer:	Prof. DDr. Hannes Leitgeb
Second Reviewer:	Prof. Dr. Stephan Hartmann
Third Examiner:	Prof. Dr. Andreas Herz
Fourth Examiner:	Prof. Dr. Igor Douven

Für Maria und Konrad

Abstract

This dissertation is on conditionals and causation. In particular, we (i) propose a method of how an agent learns conditional information, and (ii) analyse causation in terms of a new type of conditional. Our starting point is Ramsey's (1929/1990) test: accept a conditional when you can infer its consequent upon supposing its antecedent. Inspired by this test, Stalnaker (1968) developed a semantics of conditionals. In Ch. 2, we define and apply our new method of learning conditional information. It says, roughly, that you learn conditional information by updating on the corresponding Stalnaker conditional. By generalising Lewis's (1976) updating rule to *Jeffrey imaging*, our learning method becomes applicable to both certain and uncertain conditional information. The method generates the correct predictions for all of Douven's (2012) benchmark examples and Van Fraassen's (1981) Judy Benjamin Problem. In Ch. 3, we prefix Ramsey's test by suspending judgment on antecedent and consequent. Unlike the Ramsey Test semantics by Stalnaker (1968) and Gärdenfors (1978), our strengthened semantics requires the antecedent to be inferentially relevant for the consequent. We exploit this asymmetric relation of relevance in a semantic analysis of the natural language conjunction 'because'. In Ch. 4, we devise an analysis of actual causation in terms of production, where production is understood along the lines of our strengthened Ramsey Test. Our analysis solves the problems of overdetermination, conjunctive scenarios, early and late preemption, switches, double prevention, and spurious causation – a set of problems that still challenges counterfactual accounts of actual causation in the tradition of Lewis (1973c). In Ch. 5, we translate our analysis of actual causation into Halpern and Pearl's (2005) framework of causal models. As a result, our analysis is considerably simplified on the cost of losing its reductiveness. The upshot is twofold: (i) Jeffrey imaging on Stalnaker conditionals emerges as an alternative to Bayesian accounts of learning conditional information; (ii) the analyses of causation in terms of our strengthened Ramsey Test conditional prove to be worthy rivals to contemporary counterfactual accounts of causation.

Précis

Philosophers have devised various theories of the relation between cause and effect. Few of the theories have occasionally been popular. Today, none is widely agreed upon. So far, any account of causation is plagued by counterexamples. In addition, it is hard to tell whether some account tallies best with our common sense of what causes what. Hence, it is safe to say that no philosophical account of causation has yet succeeded. This is astonishing – given how pervasive and familiar causation is. Causes help us to understand and explain what is going on around us. Causes help us to intervene in the course of events to bring about certain effects, or prevent others from occurring. As thinkers and agents, we are – quite naturally – interested in causal relations. We wonder why the coffee machine is not working, why our colleague seemed so gloomy today, whether a certain diet would support our health. Everyone values knowledge of what causes what. The more surprising that there is no unanimous theory of causation at our disposal.

The importance of causation is not restricted to anyone's everyday life, but extends to the special sciences. The goal of biology, neuroscience, medicine, economics, and history – to name just a few – is to discover the causes of their respective target phenomena. Within philosophy alone, the research on causation in metaphysics and epistemology impacts contemporary debates on mental causation, action theory, decision theory, learning theory, semantics, scientific explanation, and moral and legal responsibility. The ubiquity of causation awards the prospect of a successful account of causation with a broad interest.

Our ordinary concept of causation exhibits two truisms. First, causes produce certain effects. Second, causes make a difference. That is, things would be different if the causes of some effects were absent. This difference-making idea underlies Lewis's (1973c) analysis of causation in terms of counterfactual conditionals. These conditionals are of the form 'if the cause had not been, the effect never had existed'. Lewis says that an event causes another if both events occur, and if the former had not occurred, so would not have the latter. Since Lewis's proposal counterfactual accounts have risen to some prominence in the contemporary debate on causation. Most of these accounts are meant to provide a theory of actual, singular, or token causation. The goal of such a theory is to figure out whether this particular event caused that particular event. In this dissertation, we put forth a novel analysis of actual causation in terms of a conditional that differs in kind from counterfactuals. Thereby, we achieve an analysis of causation in terms of production rather than counterfactual dependence. After all, causes bring about certain effects, they do not only make a difference.

Analyses of causation in terms of certain conditionals reveal that causal relations have a conditional structure. This observation, unfortunately, does not help much to analyse causation. For conditionals are as controversial as causation itself. And for a good reason: conditionals seem to be intimately tied to causation, even if we do not know exactly how. One indication for this conjecture is that conditionals, like causal relations, play a central role in reasoning and learning. However, there is no consensus emerging on what conditionals, and for that matter causal relations, mean. It is not even agreed upon whether conditionals have truth conditions, and thus express propositions at all (see Edg-

ington (1995)). This makes it easy to predict that conditionals will continue to puzzle philosophers, epistemologists, logicians, and cognitive scientists alike.

In spite of the controversies surrounding conditionals, much research has originated from the same source, viz. Ramsey's (1929/1990) test for the acceptance of conditionals. The idea is that you accept a conditional 'if A then C' when you can infer its consequent C upon supposing the antecedent A. Inspired by this test procedure, Stalnaker (1968) has developed a semantics of conditionals by replacing (hypothetical) belief states by a (set of) possible worlds. Roughly, the Stalnaker conditional $A > C$ is true just in case C is true in the possible world that satisfies A and is otherwise most similar to the actual world. In a slogan, $A > C$ is true when the most similar A-world satisfies C. Adams (1975) has developed another semantics of conditionals. He has taken seriously Ramsey's phrase that the evaluation of a conditional requires to fix your degree of belief in the consequent given the antecedent. Accordingly, you accept 'if A then C' when your "degree of belief in C given A" is high. Yet another influential Ramsey Test semantics has been developed within Gärdenfors's (1988) theory of belief revision. You accept a conditional 'if A then C' if a minimal change of your beliefs to accommodate A makes you believe C as well. However, Gärdenfors (1986) showed that this version of the Ramsey test is inconsistent with his theory of belief revision under a mild assumption of non-triviality. It was Hansson (1992) who defended the Ramsey Test against Gärdenfors's inconsistency theorem. In part, Hansson saves the Ramsey Test by an alternative representation of belief states: he has used belief bases rather than belief sets to model the dynamics of belief. In contrast to belief sets, belief bases are in general not closed under logical consequence. Hence, belief bases are a less idealised model of a belief state allowing for a distinction between explicitly held beliefs and merely derived beliefs. Up to now, the Ramsey Test inspires work on conditionals and belief changes (see, e.g., Bradley (2007), Leitgeb (2010), Rott (2011), Rott (2017)).

Following Ramsey (1929/1990), Adams (1975) represents degrees of belief by a probability distribution P. He goes on to stipulate the probability of an indicative conditional as the conditional probability of the consequent given its antecedent, i. e. $P(\text{if } A \text{ then } C) := P(C \mid A)$. Lewis (1976) has shown that this stipulation does not hold for Stalnaker's Ramsey Test conditional $>$. Apart from trivial cases, the probability of a Stalnaker conditional does not equal the corresponding conditional probability, that is $P(A > C) \neq P(C \mid A)$. Lewis's result may be seen as a special case of Gärdenfors's inconsistency theorem, when belief states are modelled by probability functions and belief states change according to the rule $P'(C) = P(C \mid A)$. The setting of this special case describes the two core tenets of orthodox Bayesianism, where P is called the initial or prior degrees of belief and P' the final or posterior degrees of belief, and the rule changing the degrees of belief goes by the name of conditionalization. By conditionalization on A, a Bayesian agent learns a piece of evidence A with certainty. Jeffrey's (1965) generalisation of conditionalization allows a Bayesian agent to update her degrees of belief when the information A is learned with uncertainty. So far so good.

As compared to the learning of factual information, it is less clear how the norms of Bayesian epistemology apply to the learning of indicative conditionals (see Douven (2015)). How should you change your beliefs when you learn 'if A then C'? Virtually all Bayesian accounts agree that learning a conditional imposes some constraint on the posterior degrees of belief for the consequent given the antecedent. Reminiscent of Adams's stipulation, to learn 'if A then C' is often assumed to imply that $P'(C \mid A)$ equals approximately 1 (see, e.g., Evans and Over (2004, Ch. 8) and Oaksford and Chater (2007, p. 118)). In case a Bayesian agent learns uncertain conditional information, the constraint on the posterior takes the form $P'(C \mid A) = a \leq 1$. Then Bayesians usually apply Jeffrey conditionalization. A more sophisticated alternative has been proposed by Van Fraassen (1980a) and Williams (1980). The proposal is to model the learning of conditional information by minimising the Kullback-Leibler (KL) divergence between prior and posterior degrees of belief. When you learn uncertain

conditional information, and so the constraint takes the form $P'(C \mid A) = a$ for $0 < a < 1$, minimising the KL divergence may yield different results from Jeffrey conditionalization. Van Fraassen's approach leads thus beyond the confines of orthodox Bayesianism.

Shortly after, Van Fraassen (1981) and Van Fraassen et al. (1986) have challenged the KL divergence minimizer and the orthodox Bayesian alike. They do so by putting forth a scenario of learning uncertain conditional information, the Judy Benjamin Problem. To say the least, this learning scenario has since proven to be a severe challenge for Bayesians of all varieties. And not only Bayesians are troubled. Douven and Romeijn (2011) and Douven (2012) survey the extant accounts of learning conditional information and observe that there are no satisfactory accounts of learning conditional information. All accounts – ranging from simple conditionalization on the material implication, over possible worlds accounts, to advanced Bayesian perspectives – fail to provide the correct results for Douven's benchmark examples. He concludes that a general account of learning conditional information is yet to be formulated. We aim to remedy this situation.

In light of Lewis's (1976) triviality result, it is dubious whether conditional probabilities, or at least constraints on conditional posterior degrees of belief, are the right tool to model the learning of conditional information. Luckily, an alternative to Bayesian conditionalization is not far to seek. In the same paper, Lewis has found a probabilistic updating rule, which he named imaging. Imaging on A transfers the probability shares associated to $\neg A$-worlds to the respective most similar A-worlds. We may interpret imaging on A as another way to learn A with certainty. Notably, the probability of a Stalnaker conditional equals the probability of its consequent after imaging on the antecedent, that is $P(A > C) = P^A(C)$. By replacing the updating rule of conditionalization with imaging, Adams's stipulation becomes a theorem for Stalnaker's conditional. Unlike standard conditionalization, imaging can be applied when the antecedent has probability 0 and it can be applied to nested conditionals. For instance, the image on the Stalnaker conditional $A > C$, $P^{A>C}(E)$, is well-defined and equals the probability of the nested conditional $(A > C) > E$. Hence, imaging seems to be a promising candidate to provide a rather general account of learning conditional information.

In Chapter 2, we put forth a method of learning conditional information. Roughly, an agent learns 'if A then C' by (Jeffrey) imaging on the Stalnaker conditional $A > C$. Imaging on $A > C$ amounts to (i) learning that the most similar A-world satisfies C; and that (ii) the probability share of each $\neg(A > C)$-world is transferred to its most similar $(A > C)$-world. Jeffrey imaging is our generalisation of Lewis's imaging that mirrors Jeffrey's generalisation of Bayesian conditionalization. In contrast to Lewis's imaging, our generalisation does not transfer the whole probabilistic mass – but only a part thereof – to the respective most similar worlds. Thereby, Jeffrey imaging opens the door to apply our learning method to uncertain information. Unlike extant Bayesian accounts, our method generates the correct predictions for Douven's (2012) benchmark examples and Van Fraassen's (1981) Judy Benjamin Problem. Finally, we adapt our method of learning conditional information to a method of learning causal information. The combination of the two methods provides a unified framework which allows us to clearly distinguish between conditional and causal information.

In Chapter 3, we strengthen Ramsey's test. The idea is to prefix Ramsey's test by a suspension of judgment:

> First, *suspend judgment on the antecedent A and the consequent C*. Second, add A hypothetically to your stock of beliefs. Finally, test whether you can infer C.

The resulting strengthened Ramsey Test semantics of conditionals requires that the antecedent is inferentially relevant for its consequent, unlike the semantics of conditionals due to Stalnaker (1968) and Gärdenfors (1978). Using Hansson's (1999) framework of belief bases, the relevance exhibited by our

strengthened semantics is often asymmetric. No wonder, then, that we can analyse the conjunction 'because' of natural language by our strengthened Ramsey Test semantics.

In Chapter 4, we analyse actual causation in terms of our newly developed strengthened Ramsey Test conditional. The strengthened conditional is meant to express a relation of production. C produces E just in case being agnostic on C and E, you can infer C by supposing E. Roughly, we propose that C is a cause of E if C produces E and $\neg C$ does not also produce E. Hence, we reduce causation to (beliefs about) events (or facts) and generalisations. Our analysis solves the problems of overdetermination, conjunctive scenarios, early and late preemption, switches, double prevention, and spurious causation – a set of problems that still challenges counterfactual accounts of actual causation in the tradition of Lewis (1973c).

In Chapter 5, we carry over our analysis of actual causation to Halpern and Pearl's (2005) framework of causal models. Thereby, our analysis simplifies considerably on the cost of losing its reductiveness. We compare the resulting analysis to the account of Halpern and Pearl (2005) and its modification due to Halpern (2015). Both accounts define actual causation in terms of contingent counterfactual dependence. Roughly, contingent counterfactual dependence says that even if E does not counterfactually depend on C in the actual situation, E counterfactually depends on C under certain contingencies. Their definitions of actual cause still struggle with any set of problems including both overdetermination and conjunctive scenarios, unlike our analysis in the framework of causal models.

In Chapter 6, we refine Halpern and Pearl's (2005) definition of actual causation to allow for disjunctive causes as discovered by Sartorio (2006). She argues, against the verdicts of Lewis (1986b) and Halpern and Pearl (2005), for the existence of disjunctive causes. The switching scenario she considers suggests that a disjunctive fact or event can be a cause without one of its disjuncts being a cause. We show that our refinement of Halpern and Pearl's (2005) definition can take such disjunctive causes into account.

In Chapter 7, we impose properties of causation, as assumed in cognitive neuroscience, upon Woodward's (2005) interventionist account of causation. Within the resulting framework, we investigate to what extent we are justified to derive causal relations between mental properties and properties of the brain, if certain methods are used in the neuroscientific studies. The upshot is that, for methods as diverse as Functional Magnetic Resonance Imaging and Transcranial Magnetic Stimulation, cognitive neuroscientists should dare to interpret their findings as establishing genuine causal relations.

In Appendix A, we apply our method of learning conditional information to Douven and Romeijn's (2011) Jeweller Example. Appendix B contains the proofs of Chapter 3. Appendix C provides the precise definitions for the belief changes underlying the analysis of causation in Chapter 4. Appendix D contains a proof showing that, under certain assumptions, a notion of counterfactual dependence is sufficient for causation according to our definition of Chapter 4. A more detailed summary of each chapter may be found in Section 1.5 of the Introduction.

In sum, we have moved from the Ramsey Test in two directions. First, we have proposed a method of learning conditional information. This method is based on Stalnaker's semantics of conditionals and Jeffrey imaging. As compared to extant Bayesian accounts, our framework does justice to the many facets of learning conditional information. We have gone in a second direction by strengthening the Ramsey Test. This has led us to a new conditional semantics amenable to an analysis of actual causation. In fact, we have analysed actual causation in terms of our strengthened Ramsey Test conditional twice over. The first analysis uses belief bases, while the second is embedded in a framework of causal models. In these guises, our strengthened Ramsey Test conditional gives rise to two analyses of actual causation that do not fall short of contemporary counterfactual accounts. The paths we have gone show the on-going fruitfulness of Ramsey's ideas.

Sources. Chapter 2 of this dissertation builds on the publications Günther (2018) and Günther (2017a). Chapters 3, 4, and 5 build on the publications Andreas and Günther (2018), Andreas and Günther (2019), and Andreas and Günther (2018), respectively. Chapter 6 builds on Günther (2017b).

Acknowledgements

First and foremost, I would like to thank Hannes Leitgeb. He gave me all the freedom I could wish for to pursue my research interests. At the same time, he provided excellent guidance whenever I faced a difficulty. His advice has been invaluable. He knows, in particular, what is yet to be improved upon. His proposals for clarification and relevant literature have always been dead-on. Besides the persistent support he gave me, his profound knowledge makes him an admirable supervisor. For the future, I hope he continues to be there when I need advice.

I would like to express my gratitude to Stephan Hartmann. It was his class back in the Masters program of the Munich Center for Mathematical Philosophy (MCMP) that made me familiar with the problem of learning conditional information from a Bayesian point of view. Many discussions later, I finally outlined my own non-Bayesian method of learning conditional information. This has led directly to my first and second publication. Moreover, he introduced me to a great network of academics and provided many platforms to give talks. He gave me, for instance, the responsibility to organise the ECAP9 Symposium on "Current Trends in Neurophilosophy", and even to hold his lecture on Bayesian networks for the MCMP Master students. It was more than a pleasure to have him as a supervisor.

I am grateful to Stephan Sellmaier. He was always ready to provide advice concerning both, academia and life. Throughout my PhD program, I felt strong support from his side and the environment he provided. In particular at the beginning, he accelerated my personal and academic development considerably. He always had an ear for all the PhD students in his Research Center for Neurophilosophy and Ethics of Neurosciences, which is part of the Neurophilosophy division of the Graduate School of Systemic Neurosciences (GSN). There, he created a space where we could try out new ideas in front of critical but friendly colleagues. I profited from this environment a lot.

My appreciation is due to Holger Andreas. I learned a lot from the frequent and ample discussions we had. We are in the middle of developing a research project on the Ramsey Test and causation. The first fruits of our collaboration can be found in chapters 3, 4, and 5. Over the years, we exchanged numerous drafts commenting on each other's contributions. I am grateful for his thoroughness and patience with me. It is no overstatement that the dissertation in this form would not have been possible without him.

I am indebted to Igor Douven. He proposed a – broadly speaking – Bayesian answer to the question what we learn when we receive conditional information. Without any formal commitment, he commented extensively on the earliest draft I produced on said question. Despite my non-Bayesian tendencies in approaching the question, he and Stephan Hartmann valued my proposal and encouraged me to submit it. Please excuse me for being an enfant terrible as regards Bayesianism.

Special thanks go to Andreas Herz, director of the Bernstein Center for Computational Neuroscience (BCCN) in Munich. He, the BCCN, the GSN, and the MCMP strongly supported a conference I organised together with Holger and Kay Thurley on "Causation, Explanation, Conditionals". Andreas was eager to establish an interdisciplinary exchange between neuroscientists and philosophers,

for which I am very grateful.

I owe a great many thanks to Hans Rott. He asked crucial questions at a talk I gave on learning conditional information and took plenty of time to carefully read my drafts on the Ramsey Test. His comments pointed to gaps, followed by proposals to overcome these and to work out interesting comparisons with extant literature. As a result, I could always improve upon the respective manuscript. His comments helped me, in particular, to rebut objections a reader might have before these were actually levelled. It is noteworthy that some core ideas of this dissertation originate more or less directly from Hans's work on belief revision and the Ramsey Test. Together with Holger, I attempted to carry Hans's ideas further and push them into the arena of causation. Hans contributed significantly and in ingenious ways at each important step during my PhD studies – and all of that without any formal commitment.

I would like to explicitly thank both, the Graduate School of Systemic Neurosciences and the Munich Center for Mathematical Philosophy, in particular the institutional heads Benedikt Grothe, and Hannes and Stephan. All of them are working hard to create a small heaven for PhD students. The GSN made it possible that I could organise the above mentioned conference, for which I feel honoured and am more than grateful. As regards financial support, academic events, soft-skill courses, management, and social activities, the GSN sets a bar so high that no wishes remain unfulfilled. Combined with the likewise rich offer of the MCMP, the opportunities were so numerous that I could not even take up half of the highlights I would have had liked to. It is hard to imagine a better and more vibrant environment for personal and academic development. I am happy to be part of the connection between the GSN and the MCMP.

I would like to thank Olivier Roy, Julian Nida-Rümelin and Fiorella Battaglia for continuous help and support on various occasions.

Finally, I would like to express my gratitude to Atoosa Kasirzadeh, Cameron Beebe, and Gary Mullen for reading my all too early drafts and for an enjoyable time. The same applies to my GSN fellows, especially the Neurophilosophy group, and my former MCMP classmates. I apologise to the many people that helped me over the last couple of years, but which I forgot to mention. I dedicate this work to my parents, Maria and Konrad Günther.

Contents

Chapter 1

Introduction

This is an investigation between the learning of conditionals and conditional analyses of causation. We have three principal aims. First, we aim to provide a method of learning conditional information. This method is based on Stalnaker's Ramsey Test semantics for conditionals. Second, we formally elaborate a strengthened Ramsey Test semantics for conditionals, which we then use for an analysis of 'because'. Third, we aim to analyse causation in terms of our strengthened Ramsey Test conditional. We see already that conditionals, in particular Ramsey Test conditionals, will take center stage in this investigation. A conditional can be expressed in natural language by "If A, (then) C", where A and C are place-holders for two clauses, the antecedent A and consequent C. An example of a natural language conditional is "If Paris shoots this fatal arrow, Achilles dies young." Our strengthened Ramsey Test semantics will allow us to sensibly evaluate such conditionals, and sentences like "Because Paris shot this fatal arrow, Achilles died young." Observe that the *conditional* and the *because sentence* can be uttered to say: Paris's shooting the fatal arrow was a *cause* of Achilles's early decease. This suggests that a semantics of conditionals can be used in an analysis of causation.

Throughout the investigation, we will test whether our theoretical proposals fit our common-sense judgments about different examples. That is, we agree with the methodology laid out by Lewis (1986b, p. 194):

> When common sense delivers a firm and uncontroversial answer about a not-too-far-fetched case, theory had better agree. If an analysis [...] does not deliver the common-sense answer, that is bad trouble.

Hence, we will test our analyses against not-too-far-fetched examples. Sometimes, for instance when we develop an analysis of 'because' based on our strengthened Ramsey Test, certain examples will even play a guiding role in constructing the analysis. In the case of learning conditional information, our method will be tested against a set of examples which has – as of yet – not been captured by any other method.

This introduction is meant to set the stage for the investigations in the following chapters. We review previous work on conditionals and causation, and point out what we aim to contribute. We begin with some remarks by Hume on causation. Our perhaps idiosyncratic interpretation of Hume's remarks is very much in line with Ramsey's regularity analysis of causation. This analysis is given in terms of a certain kind of conditional. For these 'variable hypotheticals' Ramsey provides a test according to which we accept 'If p, then q' just in case assuming p makes us infer q. Early research on Ramsey's test has been carried out by Stalnaker (1968) and Gärdenfors (1988). The former used the idea of the test to develop a possible worlds semantics of conditionals. The latter linked the acceptance

of conditionals more explicitly to belief changes. A brief overview of the accounts of Stalnaker and Gärdenfors shall suffice as a background for the following chapters. It should be noted, though, that the Ramsey Test continues to form a lively area of research.[1]

In another Humean spirit, Lewis (1973c) establishes an analysis of causation in terms of so-called counterfactual conditionals.[2] Counterfactuals, for short, are conditionals in the subjunctive mood, where it is somehow known, assumed, or implicated that the antecedent is contrary to the facts. Let us assume that Mary broke a window by throwing a rock. Then we can sensibly assert the counterfactual "If Mary had not thrown the rock, the window would not have broken." If Mary actually broke the window by throwing a rock, the truth of this counterfactual implies, on Lewis's analysis, that Mary's throwing a rock through the window is an actual cause of the window shattering. His idea that a cause makes a difference as regards the effect is intuitively quite appealing. However, as we will see, there are some examples (such as late preemption and overdetermination) that resist to be correctly analysed by Lewis's (1973c) counterfactual analysis. To capture these recalcitrant examples, Lewis modified his analysis twice. A problem remains, though, no Lewisian analysis can solve at the same time both, cases of late preemption and cases that became to be known as 'double prevention'.

Towards the end of this introduction, we contrast Lewis's (1973c) counterfactual analysis of actual causation with regularity analyses. Roughly, the latter say that effects regularly follow their causes. While counterfactual dependence seems to be sufficient but not necessary for actual causation, regularities seem to be necessary but not sufficient. Of course, a regularity analysis can be amended in several ways, in particular by a so-called best system analysis of regularities, as we will see. We include a few remarks on how our final analysis of causation will relate to both, Lewis's counterfactual analysis and typical regularity analyses of causation. We end the Introduction by a brief preview of what is to come.

1.1 Hume and Ramsey on Causation

> We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed.

This is Hume's (1748) definition of causation. Before the 'other words', his stipulation sounds like a regularity analysis that attempts to define causation in terms of regularities, that is invariable patterns of succession. The core idea is that causes are regularly followed by their effects. In fact, Hume (1739/1978, p. 170) says more specifically:

> A cause is an object precedent and contiguous to another, and so united with it, that the idea of the one determines the mind to form the idea of the other, and the impression of the one to form a more lively idea of the other.

Here, Hume says that the connection between cause and effect is determined by the mind. Only because the mind forms the ideas, the objects in the world are *thought* to be connected. To the mind, this causal connection appears to be 'necessary'. Hume (1739/1978, p. 165) clearly states that "[n]ecessity is something that exists in the minds, not in objects". Consequently, there be no necessary connection in the world between a cause and its effect which would go beyond their regular

[1] See, e.g., Bradley (2007), Leitgeb (2010), Rott (2011), Hill (2012), Bradley (2012), Chandler (2017), Bradley (2017), Rott (2017).

[2] The term 'counterfactual conditional' was coined by Goodman (1947).

association. Rather causation as imposed by our mind, so Hume (1739/1978, p. 73), can "produce a connexion" between objects that goes "beyond what is immediately present to the senses". In this sense, causal reasoning may extend our knowledge – based on our mind's habitual regularities – beyond our experience.[3] On this picture, causation is an epistemic relation imposed on two objects that allows the mind to 'determine' the effect when perceiving its cause. Put differently, the mind has a habit to expect an event after observing another, and we call this expectation causation. Our memory provides us with finitely many instances of objects that occur jointly, or in 'constant conjunction', and from which our mind forms inferential habits. These inferential habits make our mind to expect the effect upon observing its cause. Hence, the epistemic relation of causation is primarily inferential, as Hume (1739/1978, p. 159) suggests:

> [A]fter the discovery of the constant conjunction of any objects, we always draw an inference from one object to another, [...]. Perhaps 'twill appear in the end, that the necessary connexion depends on the inference, instead of the inference's depending on the necessary connexion.

Agreeing that the 'necessary connection' depends on the inferential habits, Ramsey (1929/1990) sketches a regularity analysis in a Humean spirit. Both claim that there is no causal relation in the world that goes beyond regularities. In particular, there be no causal necessity in the world. It is the mind that imputes regularity to the world, and only for this very mind the imputed regularity might appear of causal 'necessity'. For Ramsey, a regularity is expressed by a generalisation such as 'All men are mortal' or 'Arsenic is poisonous'. These generalisations, so Ramsey, express not merely a summary of certain conjunctive facts, but also an expectation for the future. His analysis of causation thus follows Hume in not being purely descriptive. Rather his analysis may be called a forecasting analysis of causation. The idea is the same, viz. that, upon repeated observation of an invariable pattern of succession, the mind forms a habit to expect certain effects following certain causes. Importantly, these expectations apply also to future, as of yet unobserved instances of the regularity.

Ramsey names the generalisations expressing a mind's habitual regularities 'variable hypotheticals'. These variable hypotheticals are 'rules of judgment' that guide a mind's or, in more modern terms, an agent's inferences. Suppose an agent endorses the variable hypothetical 'All men are mortal'. Then, if the agent encounters a man, she expects him to be mortal. In general, an agent's variable hypothetical encodes an expectation: if she encounters an instance of the antecedent, she expects the consequent. Hence, variable hypotheticals have a conditional structure.

The endorsed set of variable hypotheticals are "the system with which the speaker meets the future" (Ramsey (1929/1990, p. 241)). Hence, the set of variable hypotheticals corresponds to the rules according to which an agent changes her epistemic states. Suppose an agent believes a set of variable hypotheticals and a set of facts. If the antecedent of a variable hypothetical is in the set of believed facts, the agent believes its consequent as well.

The set of variable hypotheticals is divided in causal laws and accidental generalisations. Causal laws support counterfactual conditionals whose consequent does not occur earlier than its antecedent.

[3] We should acknowledge that this is our interpretation of Hume which emphasizes the mind-dependency of causation. Hume's regularity theory of *causation* is typically (or at least has often been) interpreted to be a realist theory. The reason is that there really are regularities between events. That is, there really are jointly and repeatedly occurring events that are spatially contiguous and some temporally precede others. The regularities are then mind-independent in the sense that they would still be there, even if nobody observed them. However, this realist interpretation of regularities does not seem to apply to causation. Rather the role of the 'non-realist' mind has not been considered. On the one hand there are mind-independent regularities; on the other the mind infers causal relations from these regularities. But then, should Hume's analysis not reduce causation to regularities *and* the mind (or its inferences)? We think this is about right, but will not try to resolve this tension here.

Hence, Ramsey imposes a Humean constraint of temporal precedence on causal laws. In contrast, the variable hypotheticals that are not causal laws are just accidental generalisations. Before we move on to Lewis's analysis of causation in terms of counterfactual conditionals, we will have a closer look at Ramsey's variable hypotheticals.

1.2 The Ramsey Test and Stalnaker's Conditional

Suppose an agent believes the conditional 'If she eats the cake, she will have a stomach ache', and so does not eat the piece of cake in front of her. We might disagree with the agent by believing that if she had eaten the cake, she would *not* have had a stomach ache. Our difference is then that she believes if she had eaten the cake, she would have had a stomach ache. These "assertions about unfulfilled conditions", so Ramsey (1929/1990, p. 247) about the two conditionals, do not make a difference as to the facts we believe. We both agree that she did no eat the cake. The basis of the dispute is rather that each of us is guided by different variable hypotheticals. Ramsey (1929/1990, p. 247, fn. 1) says about such situations of a hypothetical argument:

> If two people are arguing 'If p will q?' and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q; so that in a sense 'If p, q' and 'If p, $\overline{q}$' are contradictories.[4]

This procedure confers conditionals a role in the practice of hypothetical reasoning: although uncertain whether p is the case, they suppose p in order to argue about q. In such a Ramsey Test situation, hypothesizing p opens up a context which serves as basis for their reasoning and discussing about q. Note that the sense in which 'If p, q' and 'If p, $\neg q$' are contradictories is according to Ramsey not truth-conditional. Ramsey does not think that conditionals have truth values, but that there are rational conditions for accepting and rejecting conditionals, and the notion of acceptance cannot be reduced to belief in the truth of any proposition. The cited procedure is meant to provide a test for such acceptance conditions.

Let us consider another example of a Ramsey Test situation. Two people have doubts whether God exists, and consider the question 'If God exists, is Jesus His son?'. Although they have both doubts whether God exists, they may come to different conclusions when reasoning on the hypothetical basis that God exists. One of them could argue as follows: Suppose God exists. Then we have reason to believe the Bible's story about Virgin Mary, and thus Jesus is the son of God. The other could argue along the following lines: Suppose God exists. Then, since the Bible is not written by God, we have no reason to believe the miraculous story about Mary's immaculate conception, and thus Jesus is not the son of God. In the context they enter by the God hypothesis the first person could respond that, indeed, the Bible is written by men, but still these men were inspired by God, and so on. Again, the two may fiercely disagree on what follows from the God hypothesis because their inferences are lead by different variable hypotheticals. And they can do so even if they believe the same class of facts.

Stalnaker (1968, p. 101) provides a simple paraphrase of Ramsey's test procedure: "add the antecedent (hypothetically) to your stock of knowledge (or beliefs), and then consider whether or not the consequent is true." He goes on to extend his paraphrase in order to cover the cases where an

[4] $\overline{q}$ is Ramsey's notation for the negation of q, which we denote by $\neg q$. The quote continues: "We can say they are fixing their degrees of belief in q given p." Adams (1965, 1966, 1970, 1975) takes the continuation serious and develops a logic of conditionals in terms of conditional probabilities. The basic idea is that a conditional If p, then q should be accepted just in case the conditional probability $P(q \mid p)$ is high. Adams's criterion for his logic is that the improbability of its conclusion cannot exceed the sum of the improbabilities of the premises.

agent is not in doubt as regards the antecedent but believes it to be true or false, respectively. If the agent already believes the antecedent to be true, she does not need to change her beliefs. If the agent believes the antecedent to be false, she is forced to change her beliefs which contradict the antecedent. Stalnaker (1968, p. 102) thus proposes to evaluate a conditional as follows:

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true.

From here and based on Kripke's (1963) modal-logical framework, Stalnaker (1968) develops a possible worlds semantics for conditionals.[5] The idea is that adding the antecedent to your stock of beliefs moves you to a possible world, in which the antecedent is true and otherwise differs minimally from your actual stock of beliefs. Then you can check whether the consequent is satisfied in this most similar world, where the antecedent is true. Thereby, a possible world models your hypothetical beliefs. 'True' in possible worlds roughly corresponds to 'believed to be true' by the agent. In this sense, "a possible world", so Stalnaker (1968, p. 102), "is the ontological analogue of a stock of hypothetical beliefs." As a consequence, Stalnaker uses the concept 'possible world' not so much to 'transition' from acceptance or belief conditions to truth conditions, and thus does not propose a "transition from epistemology to metaphysics" as claimed by Arlo-Costa and Egré (2016).[6] Rather he *models* belief or acceptance conditions for conditionals by truth conditions in a possible worlds framework.

Admittedly, the concept of a possible world sounds metaphysically laden. However, possible worlds might just be regarded as a useful technical tool to model possibilities. This aligns with the view of Stalnaker (1984, p. 57):

> Possible worlds are primitive notions of the theory, not because of their ontological status, but because it is useful to theorize at a certain level of abstraction, a level that brings out what is common in a certain range of otherwise diverse activities. The concept of possible worlds that I am defending is not a metaphysical conception [...]. The concept is a formal or functional notion, like the notion of an individual presupposed by the semantics for extensional quantificational theory. [...] The theory leaves the nature of possible worlds as open as extensional semantics leaves the nature of individuals. A possible world is what truth is relative to, what people distinguish between in their rational activities.

Possible worlds thus understood are simply a means to model ways the world might/could be, or might/could have been. Possible worlds represent alternative scenarios according to which things might happen or might have happened, even if they actually did not happen.[7] In this sense, if we are about to eat a cake and wonder whether we will get a stomach ache, we may consider two possible worlds: one in which we eat the cake, and another in which we do not. If we did not eat the cake, we can still entertain the thought that we could have eaten the cake. The possibility that we could have eaten the cake is true in a possible world. In this way, we can use a possible world to model the

[5] Indeed, Stalnaker (1968) and Stalnaker and Thomason (1970) put forth a logic of conditionals consisting of a semantics and a corresponding proof theory. They have shown that Stalnaker's (1968) semantics of conditionals is sound and complete with the axiom system **C2**. Interestingly, the inference rules of **C2** agree with those of Adams's logic for conditionals. See previous footnote.

[6] This being said, it seems in many text passages following Stalnaker (1968, p. 102) that he actually makes a transition from epistemology to the pragmatics and semantics of natural language. There is, of course, a relation between our cognitive habits and natural language. Here is not the place to discuss it, but in Chapter 3 we show how the Ramsey Test may bear on an analysis of one natural language usage of the connective 'because'.

[7] For more details on Stalnaker's view what possible worlds are, see Stalnaker (2003).

thought about a possibility. Hence, the concept of a possible world is a useful theoretical device that allows us to model thinking about various possibilities, including models of conditional or hypothetical beliefs. For this reason, a possible worlds semantics seems well-suited for analysing the meaning of conditionals.

Stalnaker (1968) proposes roughly the following semantics for conditionals. A conditional is true just in case its consequent is true in the possible world that satisfies the antecedent and is otherwise most similar to the actual world. Apparently, he thinks that this semantics fits both indicative and subjunctive conditionals. In other words, he assumes that a conditional's mood does not affect its propositional content (ibid., p. 99, fn. 3).[8] Adams (1970) argues for the opposite, viz. that a conditional's mood makes a difference to its semantic characterisation. To bolster his claim, he presents the following pair of conditionals:

(1) If Oswald didn't kill Kennedy, someone else did.

(2) If Oswald hadn't killed Kennedy, someone else would have.

Suppose you are convinced but not entirely certain that Oswald actually killed Kennedy. At the same time, you are sure that Kennedy has been assassinated. Then you accept the indicative conditional (1): under the assumption that it was not Oswald someone else did it, because you are still convinced that Kennedy has been killed. In contrast, if you suppose that Oswald had not killed Kennedy, it seems to be an open question whether or not someone else would have killed Kennedy. Your actual conviction that Kennedy has been assassinated does not determine that the consequent of the subjunctive conditional (2) is true under the supposition of its antecedent. If you utter (2) you make the dubious claim that Kennedy's assassination was inevitable, perhaps because you think that Kennedy's actions and policies provoked an assassination. But then you think that Kennedy did not act alone, or you even adhere to some conspiracy theory. Hence, if you accept (1), you might still reject (2).

In the indicative case, it seems that the most similar possible world in which Oswald did not kill Kennedy must still conform to the proposition that Kennedy has been assassinated. Now, Kennedy has been assassinated, only if Oswald did it or someone else. Hence, we are either in a world where Oswald killed Kennedy or where someone else did it. If Oswald did not do it, someone else did. In other words, in the indicative case, we are not free to move to a most similar world where neither Oswald nor someone else killed Kennedy, because we firmly believe that *here* in the actual world Kennedy has indeed been assassinated. This constraint does not seem to be imposed on its subjunctive version (2). Moving to the most similar world from the actual where Oswald had not killed Kennedy may well be a world where Kennedy has not been assassinated. This possibility illustrates that the evaluation of subjunctives is somewhat more independent from the actual context than the evaluation of indicatives. So, even if you are certain that Oswald did in fact kill Kennedy, you can believe to be true that if Oswald had not, nobody else would have. But then you do *not* adhere to a conspiracy theory.

Now, if Stalnaker's possible-world semantics captures both indicatives and subjunctives, we need to amend it in order to explain why two different possible worlds can be, respectively, most similar from the actual, depending on the mood of the conditional. We will provide a sketchy answer in Chapter 2 to the question how the similarity order between worlds might be constrained so that the most similar world of an indicative conditional and the one of the corresponding subjunctive conditional come apart.

[8] As we will see in Chapter 2, the propositional content of a Stalnaker conditional is identical to the set of possible worlds in which the conditional is true.

In general, the question "How to specify a similarity order between possible worlds?" is tricky. A good answer still seems to be elusive. Even Lewis (1979), the realist about possible worlds, accepted that a measure of similarity is context-sensitive in the sense that there is no unique true measure of similarity between possible worlds. Nevertheless, he gave a rough recipe for how to determine a default measure of similarity that we use unless there is a reason to avoid it. Up to this day, however, most authors agree that semantics based on a similarity order remain a quite schematic framework, to be filled in with contextually relevant considerations for assessing similarities. Many authors such as Fine (1975) challenged the idea of similarity with respect to possible worlds.[9] To give an example – an unsophisticated one as compared to the mood of a conditional – consider: 'if this stone were a man, it would be mortal.' To figure out whether we should accept this conditional, we move to the most similar world (from our actual world) in which the stone is a man and check whether it is then mortal. Now, you might move to a most similar world where the stone is a man and mortal. However, somebody else might move to a most similar world in which the stone is a 'man' in the sense of a stone sculpture of a man. In this world, the stone is not mortal. Between different agents, the constraints on similarity may well differ. Worlds which are similar for her may not be similar to you. Such a difference may lead to a Ramsey Test disagreement.

The situation as regards the similarity order is even worse. Even one agent may have no clear-cut judgment on what is more similar to what. Is Tom's right thumb more similar to the agent's left thumb than his father's right thumb? Is a quiche more like a pizza or like a cake? Many of us do not have a clear answer, but are prompted to ponder when asked such questions. The problem is that similarity is multi-dimensional. If two worlds are each similar to the actual world in a different respect, which is most similar to the actual world all things considered?

Relatedly, what if there are two possibilities that appear to be equally similar? Can we always decide which world is *the* most similar world to the actual? Formally, Stalnaker makes this uniqueness assumption: for any non-contradictory antecedent and any world, there will be a unique most similar antecedent world. This uniqueness assumption is challenged by pairs of conditionals such as the following due to Quine (1959, p. 15):

(3) If Verdi and Bizet had been compatriots, Bizet would have been Italian.

(4) If Verdi and Bizet had been compatriots, Verdi would have been French.

Only knowing that Bizet was French and Verdi Italian, it seems that we cannot decide which world should be singled out by the common antecedent.[10] If Stalnaker's uniqueness assumption is in place, exactly one of (3) and (4) would be true according to his semantics. Alas, both seem dubious.

Stalnaker's uniqueness assumption ensures the validity of the principle of Conditional Excluded Middle (CEM). Writing $>$ for Stalnaker's conditional operator, this principle says:

$$(A > C) \vee (A > \neg C) \tag{CEM}$$

In words, either if A, C or if $A, \neg C$. Surprisingly, this validity seems to persist even in the problematic Bizet-Verdi case:

(5) Either if Verdi and Bizet had been compatriots, Bizet would have been Italian, or (if Verdi and Bizet had been compatriots), Verdi would have been French.

[9] For a review of the discussion, see Bennett (2003, chs. 11 and 12).

[10] If you have any reason to break the tie between Verdi and Bizet with respect to their nationality, you can of course decide for one most similar world. If so, just replace the example by one where you cannot break the tie between two (or more) worlds.

Stalnaker (1981) suggests that (3) and (4) are semantically indeterminate, because resolving the similarity tie between the two candidate worlds be context-dependent. (5) is true nevertheless because it would be true under any reasonable resolution of the similarity indeterminacy.[11]

Like Stalnaker (1968), Lewis (1973a,b) provides a semantics for conditionals based on the concept of a possible world and a similarity order. However, Lewis's semantics, meant for counterfactual conditionals only, does not satisfy Stalnaker's uniqueness assumption. For any two propositions A and C, so claims Lewis, the proposition 'if A were true, then C would be true' can be evaluated. Let us call a possible world that satisfies A an A-world. Then, the Lewis counterfactual $A \boxrightarrow C$ is true (at a world w) iff either there are no possible A-worlds, or some A-world that satisfies C is more similar (to w) than any A-world that does not satisfy C. If there are no possible A-worlds, then the counterfactual is vacuously true. A counterfactual is non-vacuously true just in case "it takes less departure from actuality to make the consequent true along with the antecedent than it does to make the antecedent true without the consequent" (Lewis (1973c), p. 560).

Lewis's semantics allows for a set of equally similar possible worlds. In the Bizet-Verdi case, for example, there are two equally similar possible worlds, one in which they are both French, and another in which they are both Italian. Accordingly, on Lewis's semantics, both (3) and (4) come out false, because there are two equally similar worlds in which Verdi and Bizet are compatriots *and* in one it is true that they are Italian (not French), in the other that they are French (not Italian). Interestingly, (5) is true in Lewis's semantics when understood as saying $A \boxrightarrow (C \vee \neg C)$, and false when understood as taking the form of the Conditional Excluded Middle principle, i. e. $(A \boxrightarrow C) \vee (A \boxrightarrow \neg C)$. Hence, $\boxrightarrow$ does not validate (CEM).

In Chapter 2, our goal is not to solve the problem of how one can plausibly specify a similarity order between possible worlds. Thinking one could offer a unique solution that works for all judgments of similarity seems to be either overconfident or even presumptuous. Instead, we aim to propose a method of how a rational agent should learn conditional information. We exploit the possible-worlds framework as established by Kripke and Stalnaker to model examples of learning conditional information. In particular, possible worlds will be used as Stalnakerian means to model an agent's epistemic state. The conditional information to be learned will be encoded by a certain meaning of Stalnaker conditionals. Thereby, our learning method is based on the meaning of Stalnaker conditionals and thus seems to run into the problem of how to specify the similarity order between possible worlds.

Fortunately, there is a decisive difference between the meaning of a (conditional) proposition and the learning of this meaning: while the explicit meaning presupposes an antecedently fixed similarity order, the learning of a (conditional) proposition may specify or adjust the similarity order. To be clear, the explicit evaluation of a Stalnaker conditional's meaning requires an agent to have an already fixed (at least partially) similarity order; otherwise the agent does not know to which most similar antecedent world she is supposed to move in order to check whether there the consequent is true. In contrast, the learning of a Stalnaker conditional's meaning provides the agent with information of how to specify or adjust the similarity order; the agent *learns* that the most *similar* antecedent world is a consequent world.[12] Hence, to learn a Stalnaker conditional is, in part, to obtain some information

[11] Stalnaker (1981) acknowledges indeterminacy as an immanent element of natural language and captures it by adopting Fraassen's (1966) super-evaluations.

[12] Note that, with respect to learning, the difference between Lewis's and Stalnaker's semantics is not very big. Whether you learn that in all most similar worlds, or the most similar world, the consequent is true makes no big difference. If you learn conditional (3), you come to know that both Bizet and Verdi are Italian in the/all the most similar world(s), where they are compatriots. Only if you learn the negation of a conditional, that is a conditional of the form $\neg(A > C)$, the semantics result in different propositions to be learned. In brief, Stalnaker's semantics gives you, by (CEM), the proposition expressed by $A > \neg C$. Lewis's semantic is more problematic with respect to learning. $\neg(A \boxrightarrow C)$ is non-vacuously true at a world w iff for any $(A \wedge C)$-world there is at least one $(A \wedge \neg C)$-world that is equally similar to w. By learning $\neg(A \boxrightarrow C)$ you

about a similarity order. The obtained information can then be used to specify (or adjust) the as of yet un- or under-specified (or even differently specified) similarity order. In this way, we turn the disadvantage of how to specify a 'correct' similarity order in the advantage that a similarity order can be made explicit within the learning process.

The difference between the meaning and the learning of a conditional within a possible-worlds framework has been widely overlooked so far. Douven (2012, pp. 8-9) observed that nobody employed a possible-worlds semantics to model the learning of conditionals. Rather broadly Bayesian approaches would be promising candidates. The idea in these approaches is that learning the unnested conditional "If A, then C" implies that an agent's credence function is constrained by the conditional probability $P(C \mid A)$, often with the further constraint that $P(C \mid A) \approx 1$. In light of Lewis's (1976) results, both the disinterest in a possible-worlds framework to model the learning of conditional information and the proclivity for Bayesian approaches seems to be somewhat odd. After all, Lewis has shown that there is an updating method, called 'imaging', according to which the probability of a Stalnaker conditional corresponds to the probability of its consequent upon updating on its antecedent. Moreover, Lewis has shown that, in general, the conditional probability $P(C \mid A)$ is not equal to the probability of the Stalnaker conditional $P(A > C)$. Hence, if the Stalnaker conditional is a good approximation of the conditional information we learn, we should employ Lewis's imaging rather than Bayesian conditionalization, at least when learning conditional information.

According to our learning method, imaging on a certain proposition expressed by the Stalnaker conditional $A > C$ is tantamount to learning conditional information. Imaging has the advantage over conditionalization that it can handle arbitrary compositions of propositions, for example any nested conditionals. Imaging can be used to transfer the probability mass of each possible world to its most similar possible world that satisfies $A > C$, even if $P(A) = 0$. In order to capture the learning of uncertain conditional information, such as $P(A > C) = k$ for $k \in [0, 1]$, we will generalize Lewis's imaging to what we call Jeffrey imaging. Equipped with this generalisation, our method generates the correct predictions for all of Douven's (2012) benchmark examples and Van Fraassen's (1981) Judy Benjamin Problem. Both are unresolved challenges for broadly speaking Bayesian accounts like Woodward's (2005) and Hartmann and Rad's (2017).

Moving back to Stalnaker's (and Lewis's) semantics, there is another problem: any two propositions A, C that are true in the actual world make a true conditional $A > C$ (and $A \boxrightarrow C$). This is because the most similar possible world that satisfies the antecedent is here the actual world itself. Hence, Stalnaker's (and Lewis's) logic for conditionals validates the following principle:

$$A \wedge C \vdash A > C \tag{1.1}$$

This means, for example, that if you believe 'Lund is a town in Sweden' and 'Munich is a town in Germany', then you should also accept 'If Lund is a town in Sweden, then Munich is a town in Germany' (and also the converse conditional). This conditional seems to be odd, because the consequent seems to be independent of the antecedent. To assume the antecedent seems to have no bearing whatsoever on the acceptance of the consequent. However, Stalnaker (1968, p. 101) has a rationale why we should accept these odd conditionals:

> if you already believe the consequent (and if you also believe it to be causally independent of the antecedent), then it will remain a part of your stock of beliefs when you add the antecedent, since the rational man does not change his beliefs without reason.

obtain only the information that the set of equally most similar A-worlds are not all C-worlds. This is not much to learn.

Here, the change of belief and the plausible acceptance of a conditional come apart. While it is reasonable to retain the belief in the consequent when you add the belief in the antecedent, it seems to be no sufficient ground to accept the conditional. Intuitively, some relation of relevance between the antecedent and the consequent seems to be necessary to plausibly accept a conditional.

The same problem haunts Gärdenfors's (1988) Ramsey Test defined within his theory of belief revision. On p. 147, he paraphrases Ramsey's test maybe more accurately than Stalnaker, but on pain of brevity:

> In order to find out whether a conditional sentence is acceptable in a given state of belief, one first adds the antecedent of the conditional hypothetically to the given stock of beliefs. Second, if the antecedent together with the formerly accepted sentences leads to a contradiction, then one makes some adjustments, which are as small as possible without modifying the hypothetical belief in the antecedent, such that consistency is maintained. Finally, one considers whether or not the consequent of the conditional is then accepted in this adjusted state of belief.

Gärdenfors's variant says: accept the conditional "if A then C" in an epistemic state just in case the minimal change of this state necessary to accept A also requires one to accept C. The changes, which are 'as small as possible', correspond to Stalnaker's moving to the 'most similar' possible world.[13] More forcefully than Stalnaker, Gärdenfors insists that conditionals are accepted or rejected only relative to an agent's epistemic state.

Now, Gärdenfors's Ramsey Test also validates a version of the principle (1.1). Suppose an agent already believes A and C. Then the minimal change to accept A corresponds to no change of the agent's epistemic state. But then the agent still believes C. Hence, the agent is committed to accept 'if A then C'.

In Chapter 3, we tackle the problem that extant formalisations of Ramsey's test lack a relation of relevance that is sufficient for the plausible acceptance of conditionals. Inspired by Rott (1986), we search for plausible ways to invalidate the principle (1.1) within AGM style belief revision theory. When we use Hansson's (1999) belief bases, we find the following strengthening of the Ramsey Test that invalidates the 'and-to-if' principle (1.1). In brief, accept a conditional $A \gg C$ just in case, after suspending judgment on A and C, you can infer C by assuming A. The strengthening consists thus in prefixing the ordinary Ramsey Test by a suspension of judgment. To put it in a Stalnakerian paraphrase:

> First, *suspend judgment on the antecedent A and the consequent C*. Second, add A hypothetically to your stock of beliefs. Finally, test whether you can infer C.

This strengthened Ramsey Test captures a wide range of asymmetric relevance relations in the following sense: to assume the antecedent in the context of a certain kind of belief base allows the agent to infer the consequent, but not vice versa. We will illustrate the asymmetry of our strengthened Ramsey Test conditional by a generalised tower-shadow example. Famously, a shadow is cast because of the tower, together with the sun, whereas the tower is intuitively not there because of the shadow and the sun. The asymmetry and inferential structure of our strengthened Ramsey Test conditional make it apt to analyse the conjunction 'because' of natural language which is involved in such tower-shadow explanations.

[13] Like Stalnaker, Gärdenfors (1988) develops a Ramsey Test semantics for conditionals. After formalizing the Ramsey Test within his belief revision theory, he goes on to show that his semantics is sound and complete with respect to Lewis's axiom system **VC** for his semantics of counterfactuals. Hence, the semantics of Gärdenfors and Lewis agree, at least on their common domain.

Let us relate Stalnaker's conditional back to Ramsey's variable hypotheticals. Stalnaker's semantics for conditionals can be regarded as contributing to the understanding between Ramsey's variable hypotheticals and subjunctives. Variable hypotheticals support subjunctives because the former entail the latter. A variable hypothetical such as 'All men are mortal' does – understood as a Stalnaker conditional – not just state that every actual man is mortal, but also that if something were a man, it would be mortal. Hence, 'if this stone were a man, it would be mortal' can be regarded as an instantiation of the variable hypothetical 'All men are mortal'.

Ramsey (1929/1990) claims that variable hypotheticals and counterfactuals have no truth value. To bolster his claim, he provides the following example. Suppose humans always assumed, for no reason of course, that strawberries would give them stomach ache and thus never ate them. They all accept the variable hypothetical "if I eat strawberries I shall have a stomach ache". As a consequence of the variable hypothetical, they endorse the counterfactual 'if I were to eat strawberries, I would come to have a stomach ache' (and, by the way, also the corresponding indicative conditional). But is it not a *fact* that 'if I were to eat strawberries, I would *not* come to have a stomach ache'? Ramsey says no. This 'fact' be just a consequence of our commonly shared variable hypothetical 'if I eat strawberries I will not have a stomach ache'. In contrast, the conjunction that 'I have eaten strawberries and had no pain' be a fact. In Ramsey's words (ibid., p. 253):

> What is a fact is that I have eaten them and not had a pain. If we regarded the unfulfilled conditional as a fact we should have to suppose that any such statement as 'If he had shuffled the cards, he would have dealt himself the ace' has a clear sense true or false, which is absurd. We only regard it as sense if it, or its contradictory, can be deduced from our system. Otherwise we say 'You can't say what would have happened', which sounds like a confession of ignorance, and is so indeed, because it means we can't foretell what *will* happen in a similar case, but not because 'what would have happened' is a reality of which we are ignorant.

Here, Ramsey says it is absurd that counterfactuals are true or false. Subjunctives are more about what we can infer than about matters of fact. If we cannot infer the consequent, or its negation, from the supposition of the antecedent, we lack some variable hypotheticals. Without them we cannot predict what will happen in similar circumstances. The view that counterfactuals lack truth values stands in a stark contrast with the view to which we turn next.

1.3 Lewis's Counterfactual Analysis of Causation

Lewis (1986d) claims that counterfactuals do have truth values. He states on p. 22: "If some (A-and-C)-world is closer to our world than any (A-and-C)-world, that's what makes the counterfactual true at our world." What allows him to assign proper truth values to counterfactuals is his metaphysical view on possible worlds. Unlike Stalnaker, who uses the concept of a possible world as an instrument to model beliefs, Lewis holds that possible worlds exist; moreover, these worlds could be ordered by similarity in a vague but well-understood way. Hence, it be a factual (although sometimes vague) issue whether or not some $(A \wedge C)$-world is more similar to the actual world than any $(A \wedge \neg C)$-world is. To be clear, he does not understand possible worlds as something that does not exist but could have. Instead, he thinks worlds are ways things could have been and these ways exist. And Lewis (1973a, p. 85) is serious: "When I profess realism about possible worlds, I mean to be taken literally."

So, what is a possible world when it is not a tool to model, for example, information states? For Lewis (1986d, p. 1), the actual world is the concrete physical universe in space and time, in his quite picturesque words:

> The world we live in is a very inclusive thing....There is nothing so far away from us as not to be part of our world. Anything at any distance is to be included. Likewise the world is inclusive in time. No long-gone ancient Romans, no long-gone pterodactyls, no long-gone primordial clouds of plasma are too far in the past, nor are the dead dark stars too far in the future, to be part of this same world....[N]othing is so alien in kind as not to be part of our world, provided only that it does exist at some distance and direction from here, or at some time before or after or simultaneous with now.

The actual world, so Lewis, provides us with the most vivid example of a possible world. Accordingly, possible worlds are of the same kind than the actual world. In this line, Lewis (1973a, p. 85) writes "[o]ur actual world is only one world among others. We call it alone actual not because it differs in kind from all the rest but because it is the world we inhabit". The difference between the actual and the possible worlds be only that the actual world happens to be *our* world. What makes our world actual is thus merely the contingency that we happen to live in this one.

Lewis (1986a) dubs his ontological view about the actual world 'Humean supervenience': there is nothing to our world except the spatio-temporal distribution of local natural properties. On p. ix, he notes:

> Humean supervenience is named in honor of the greater denier of necessary connections. It is the doctrine that all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another. (But it is no part of the thesis that these local matters are mental.)

Lewis has found a way to reconcile the doctrine of Humean supervenience with modalities such as counterfactuals. The solution is, roughly speaking, that possible worlds exist, but no part of one world is a part of another. In any case, Lewis established an interpretation of Hume according to which you can have it both: (i) only spatio-temporal entities exist in our world, and (ii) causal relations can be understood as inter-world, or equivalently modal, relations that have a truth value.

The sentence after the 'other words' in Hume's definition of causation reads like an analysis of causation in terms of counterfactuals. Lewis takes up this idea to analyse causation between events, i. e. natural properties of regions of space-time.[14] Relying on his semantics for counterfactual conditionals, Lewis (1973c) defines actual causation in terms of causal dependence. Let C and E be distinct events.[15] E causally depends on C iff whether or not E occurs counterfactually depends on whether or not C occurs.[16] Causal dependence is thus satisfied just in case two counterfactuals are true: $C \boxdot\!\!\rightarrow E$ and $\neg C \boxdot\!\!\rightarrow \neg E$. We say that E counterfactually depends on C iff the counterfactual $C \boxdot\!\!\rightarrow E$ is true. If C and E are actual events, that is they occur in the actual world, the first counterfactual is true, as we already observed. Hence, if C and E occur, E causally depends on C iff $\neg C \boxdot\!\!\rightarrow \neg E$. In this sense, Lewis (1973c, p. 557) writes:

> We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, [...] its effects would have been absent as well.

[14] See Lewis (1986c) for a detailed description of what his events are.

[15] The restriction to distinct events was first pointed out by Kim (1973). If A and B are not wholly distinct events, A's happening is caused by any essential part of A, let us say B. This seems to be rather a relation of constitution or supervenience than causation.

[16] Strictly speaking, C and E are propositions that express events. However, we can pair each event to the proposition E that is satisfied at exactly those worlds where the event occurs. The set of E-worlds is thus the proposition that this event occurs. Hence, we may use events and propositions interchangeably if we assume that there are no necessary connections between distinct events.

Lewis takes causation to be the transitive closure of causal dependence. Let $\langle C, B, ..., D, E\rangle$ be a finite sequence of actual events. If B causally depends on C, ..., and E causally depends on D, we call $\langle C, B, ..., D, E\rangle$ a causal chain. Lewis's definition of actual causation reads thus: C is a cause of E just in case there is a causal chain from C to E. Given the distinct events C and E are both actual, it is thus sufficient for C being a cause of E that $\neg E$ counterfactually depends on $\neg C$. If E, D, and C are actual events, and $\neg E$ counterfactually depends on $\neg D$ and $\neg D$ counterfactually depends on $\neg C$, then C is a cause of E. However, it is here still possible that $\neg E$ does not counterfactually depend on $\neg C$. Hence, counterfactual dependence is not necessary for causation. The reason is that counterfactual dependence is not transitive, as we shall see shortly. Consequently, causation is a wider notion than counterfactual dependence.

One important amendment of Lewis's (1973c) analysis of causation is the exclusion of what he calls 'backtracking' counterfactuals. For exemplification, assume that C is the only cause of both E and F. Hence, we may infer that there is a counterfactual dependence between E and F: if E had not occurred, then it would have to have been the case that C did not occur; but in this case, F would not have occurred. This counterfactual tracks the unique cause C from one of the effects. Lewis excludes such backtracking counterfactuals allowing only for non-backtracking counterfactuals to be used in the assessment of causal dependence. In fact, Lewis (1973c, p. 566) stipulates that non-backtracking counterfactuals "typically keep fixed the past up until the time at which the counterfactual antecedent is supposed to obtain". The reason be that this constitutes the "least over-all departure from actuality" (ibid.). Consider the following scenario of early preemption:

Example 1. Early Preemption
Suppose two snipers conspire to assassinate a dictator. Sniper A and sniper B both take aim as the dictator appears. A pulls her trigger and fires a shot which hits and kills the dictator. When B sees A pulling the trigger, B desists to do so. However, had A not pulled the trigger, B would have shot and killed the dictator.

Here, A's taking aim (and also A's pulling the trigger) is an actual cause of the dictator's death, while B's taking aim is a preempted, merely potential cause. Lewis's analysis of causation captures the difference of the preempting actual cause to the preempted potential cause. The reason is that there is a causal chain running from A's taking aim to the dictator's death, whereas there is no such chain for B. There exists an intermediary event occurring between A's taking aim and the dictator's death, let us say the bullet fired by A speeding through the air. Now, the speeding bullet causally depends on A's taking aim and pulling the trigger, and the dictator's death on the speeding bullet. By the time the bullet is speeding through the air B has already refrained from firing. Hence, the dictator would not have died if it were not for A's bullet speeding towards her. Since we have a causal chain, we have causation. However, there is no such intermediary event to be found between B's taking aim and the dictator's death. Consequently, B's taking aim is no actual cause of the dictator's death.

Furthermore, this early preemption scenario illustrates that causal dependence is not transitive. Although the speeding bullet causally depends on A's taking aim and pulling the trigger, and the dictator's death on the speeding bullet, the dictator's death does not causally depend on A's taking aim and pulling the trigger.

Lewis (1986b) distinguishes cases of early and late preemption. In early preemption cases the process from the preempted potential cause (B's taking aim) is cut off before the process from the preempting cause (A's taking aim) has gone to completion (dictator's death). In cases of late preemption, however, the process from the preempted cause is cut off only after the process from the preempting cause has brought about the effect. Consider the following example of late preemption:

Example 2. Late Preemption
Suppose two snipers conspire to assassinate a dictator. Sniper *A* and sniper *B* both take aim as the dictator appears. At the same time, both *A* and *B* pull their triggers and fire each a shot. *A*'s shot hits the dictator just a moment before *B*'s would have. The dictator falls down, *B*'s bullet flies through the air where the dictator was standing, and the dictator dies. Each shot alone would have been sufficient for the dictator's death.

Lewis's analysis cannot capture that *A*'s actions are the actual cause of the dictator's death. There is no causal dependence between *A*'s actions and the death. If *A* had not shot, the dictator would have died nevertheless due to *B*'s shot. Furthermore, there is no chain of stepwise causal dependences running from cause to effect, because there is no intermediate event to be found that links *A*'s shot and the death into a chain of causal dependences. In other words, the death of the dictator does not causally depend on *A*'s bullet being in mid-trajectory, since the dictator would still have died due to *B*'s bullet being in mid-trajectory.

Consider a slight modification of Example 2:

Example 3. Overdetermination
Suppose two snipers conspire to assassinate a dictator. Sniper *A* and sniper *B* both take aim as the dictator appears. At the same time, both *A* and *B* pull their triggers and fire each a shot. The two shots hit the dictator simultaneously and the dictator dies. Each shot alone would have been sufficient for the dictator's death.

Here, the actions of both *A* and *B* should count as actual causes of the dictator's death. However, Lewis's analysis says that none of these actions count. If *A* had not shot, for example, the dictator would still have died, because of *B*'s shot. The strategy to look for intermediary events does not work here. Let us consider, for example, *B*'s bullet speeding through the air. In contrast to the case of early preemption, if *B*'s bullet were not speeding through the air, *A*'s bullet would still speed through the air and kill the dictator. Hence, *B*'s actions are no cause of the dictator's death. Due to the symmetry of the scenario, the same applies to *A*'s actions. The cases of late preemption and overdetermination cannot be handled by the analysis of Lewis (1973c).

In light of recalcitrant examples, in particular late preemption, Lewis changed his analysis of causation twice over. Lewis (1986b) refines his notion of causation by the concept of 'quasi-dependence' to solve late preemption scenarios. Dissatisfied with the quasi-dependence causation, Lewis (2000) defines causation as causal influence. We are not so much concerned with the details of the new accounts of causation. Instead we want to point to a rather general problem of counterfactual accounts of causation: it seems as if they can either solve cases of late preemption or of double prevention, but not both.

Lewis (2000) presents one argument against his earlier solution to late preemption in Lewis (1986b). The problem is that Lewis's (1986b) refined analysis makes causation intrinsic to the pair *C* and *E*, but some cases, especially cases of double prevention, show that causation is extrinsic. Briefly explained, double prevention occurs when an event *C* prevents something that would have prevented *E* from happening. Intuitively, some of the double prevention cases are cases of causation. But that *C* causes *E* does not depend on the intrinsic natures of *C* and *E*. Rather, it depends on there being some threat to *E*, a threat that *C* prevents, and the existence of threats is typically extrinsic to events, or so the argument goes.

Hall (2004) provides an example of double prevention, which intuitively does not count as a case of causation. A person is hiking on a mountain trail, when a boulder high above is dislodged and rolls down the mountain slope toward the hiker. The hiker sees the boulder coming and ducks at

the appropriate time. The rolling boulder causes the walker to duck and this, in turn, enables him to continue the hike. This is a case of double prevention: the duck prevents the collision between hiker and boulder which, had it occurred, would have prevented the hiker's continuation. However, the rolling boulder is the sort of thing that would prevent the walker's continuation and so it seems counterintuitive to say that it causes the continuation of the hike. Hence, this example challenges the transitivity of causation.

Recall that Lewis's counterfactual analysis relies on the transitivity of causation to handle cases of early preemption. This commitment to transitivity is costly when facing cases of double prevention. Lewis (1973c), the origin of most contemporary counterfactual accounts of causation, and his successor theories cannot concede the persuasive counterexamples to transitivity without succumbing to the difficulties posed by early and, of course, late preemption.[17]

In Chapter 4, we devise an analysis of causation in terms of our strengthened Ramsey Test conditional. Instead of a counterfactual assumption, we suspend judgment on the cause and the effect. Thereby, the problem of overdetermination does not arise in the first place, which is a crucial advantage over counterfactual analyses. Moreover, early and late preemption turn out to be the same problem. These advantages give us sufficient space to add an extra condition to capture switches and double prevention without suffering any drawbacks on the other problems. Finally, we will tackle the problem of spurious causation to provide a properly reductive analysis of causation. A cause is spurious, for instance, if there is a common cause of two effects having no other causes and there is a non-zero temporal distance between the two. Hence, there is the danger to mistakenly classify the effect that occurs before the other as the genuine cause of the latter. After all, the earlier effect temporally precedes the latter effect and both occur in constant conjunction (because both occur just in case their common cause occurs). Furthermore, if the earlier effect had not occurred, the later never had existed. In our analysis, the problem of spurious causation boils down to have the 'correct' generalisations that let us infer the effects from the genuine causes, but not from spurious causes. Hence, we complete our analysis by a best system account of generalisations. We present the idea behind a best system account of lawful generalisations in the next section.

More recently, Halpern and Pearl (2005) provide a powerful but non-reductive definition of causation in the tradition of Lewis (1973c). Their account of causation can be regarded to be inspired by an observation of Lewis (1986b, p. 2006):

> Hold fixed the laws but change the surroundings, in any of many ways, and we would have the dependence that my original analysis requires for causation.

Halpern and Pearl (2005) implement the idea to 'change the surroundings, in *almost* any of many ways' in their definition of actual causation. Thereby, they change Lewis's counterfactual dependence to counterfactual dependence under certain contingencies. At the same time, the 'laws', or so-called 'structural equations', remain invariant. However, since these structural equations are supposed to represent causal relations, Halpern and Pearl's (2005) account of causation is not fully reductive: some information about causal relations is antecedently encoded in the structural equations. Although they solve a great number of problematic cases, both Halpern and Pearl (2005) and Halpern (2015) still struggle with a satisfactory solution to the set of problems containing overdetermination cases, also called 'disjunctive' scenarios, and 'conjunctive' scenarios, where two events are necessary for an effect to occur.

In Chapter 5, we translate our analysis of actual causation developed in Chapter 4 into Halpern and Pearl's (2005) framework of causal models. We show that our Ramsey-Test analysis defined

[17] For an insightful discussion of this trade-off problem, see Hitchcock (2001).

on causal models deals satisfactorily with both overdetermination and conjunctive scenarios. By the asymmetry of the structural equations and employing Halpern and Pearl's (2005) definition of intervention, our analysis will be simplified considerably. In fact, we will only need two conditions to capture overdetermination, conjunctive scenarios, and preemption.

Following Lewis (1986b) and Lewis (1986c), Halpern and Pearl (2005) do not allow disjunctions of events to be candidates for actual causes. The reason for this exclusion, so Halpern and Pearl (2005, p. 853), be that if a disjunction is a cause at least one of its disjuncts must be a cause – we just do not know which one. Hence, "there be no truly disjunctive causes once all the relevant facts are known." (ibid.) This is in line with the metaphysical verdict of Lewis (1986b, p. 212), who does not know "how a genuine event could be the disjunction of two events both of which actually occur." Sartorio (2006) has shed doubt on this implication of Lewis's metaphysics. She argues for the existence of disjunctive causes by presenting a particular switching scenario. In the light of Sartorio's Switch, we think Lewis's metaphysical reasons to exclude disjunctive causes might have been somewhat premature.

In Chapter 6, we aim to refine Halpern and Pearl's (2005) definition of actual causation such that it allows for disjunctive causes of the type found in Sartorio's switch. In order to treat such disjunctive causes within Halpern and Pearl's framework of causal models, we first extend their causal model semantics by disjunctive antecedents. Based on the extension, we will show that our refined Halpern-Pearl definition aligns with Sartorio's (2006) observation: a disjunctive cause does not imply that one of its disjuncts *must* also be a cause.

1.4 Regularity Analysis of Causation

Lewis's counterfactual analysis of actual causation can be contrasted with regularity analyses. As we have already seen, Hume (1748) and Ramsey (1929/1990) attempt to provide an analysis of causation in terms of regularities. In general, a typical sketch of a regularity theory goes like this: C is a cause of E just in case C belongs to a minimal set of conditions that are jointly sufficient for E, given the lawful regularities.

If the regularities can be expressed by generalisations, we can state the common ground of most regularity analyses: a cause is any member of any minimal set of actual conditions that entail, in the presence of certain generalisations, the occurrence of the effect. In greater detail:

Definition 1. Common Ground Regularity Analysis
Let C and E be distinct events, $\mathcal{L}$ a non-empty set of generalisations, and $\mathcal{F}$ a set of occurring events (or true facts).[18] C is a cause of E relative to $\mathcal{F}$ and $\mathcal{L}$ iff

(i) C and E occur,

(ii) $\mathcal{L} \wedge \mathcal{F} \models C \rightarrow E$,

(iii) $\mathcal{L} \wedge \mathcal{F} \not\models E$, and

(iv) $\mathcal{F} \not\models C \rightarrow E$.

To sum up, C is a cause of E iff (i+ii) C and E occur, and $\mathcal{L}$ and $\mathcal{F}$ jointly imply the material implication $C \rightarrow E$, but (iii+iv) $\mathcal{L}$ and $\mathcal{F}$ jointly do not imply E, and $\mathcal{F}$ alone does not imply $C \rightarrow E$. There are three important considerations here. First, as already mentioned, a genuine actual cause

[18] Of course, $C, E, \mathcal{L}$, and $\mathcal{F}$ are propositions or sets of propositions that represent or express events (or facts) and regularities.

C is, in the presence of the occurring facts and the generalisations, *sufficient* for the occurrence of the effect E. Second, a genuine actual cause C is also necessary for the occurrence of E: without C the generalisations and facts do not allow to infer E. Thirdly, the generalisations are necessary for C being a sufficient condition and thus a cause for E. Note that the Common Ground Regularity Analysis requires a weak condition of difference-making: without C and/or the set of generalisations $\mathcal{L}$ the effect cannot be derived.

Suppose C is a cause of E in virtue of $\mathcal{L}$ and $\mathcal{F}$ according to the Common Ground Regularity Analysis. Lewis (1973c, p. 565) shows that under additional assumptions C is then also a cause of E according to his counterfactual analysis. The additional assumptions are $\mathcal{L} \wedge \mathcal{F} \models \neg C \rightarrow \neg E$ and that $\mathcal{L}$ and $\mathcal{F}$ are counterfactually independent of C and $\neg C$. The argument goes roughly as follows: $\mathcal{L}$ and $\mathcal{F}$ together with $\neg C$ do not imply E (as we know by (ii) and (iii) that they do imply E only with C). Under the stronger assumption that they do imply $\neg E$, we obtain that $\mathcal{L} \wedge \mathcal{F} \models (C \rightarrow E) \wedge (\neg C \rightarrow \neg E)$ and $\mathcal{F} \not\models (C \rightarrow E) \wedge (\neg C \rightarrow \neg E)$. Let us add the final assumption that $\mathcal{L}$ and $\mathcal{F}$ are counterfactually independent of C and $\neg C$, that is $C \; \Box\!\!\rightarrow \; \mathcal{L} \wedge \mathcal{F}$ and $\neg C \; \Box\!\!\rightarrow \; \mathcal{L} \wedge \mathcal{F}$. From this follows that E causally depends on C, and thus C is a cause of E according to Lewis's analysis. In fact, all of Lewis's analyses imply that C is a cause of E if both occur, and the counterfactual $\neg C \; \Box\!\!\rightarrow \; \neg E$ is true (relative to a similarity order between possible worlds). If you replace the 'if' in the preceding sentence by an 'iff', you obtain what has been dubbed the 'simple counterfactual analysis of causation', which is often taken as sufficient but not necessary for causation. In any case, it serves well as a proxy for causal relations.

Lewis's counterfactual analysis of causation can be regarded as an attempt to overcome several shortcomings of typical regularity analyses of causation. Kyburg (1965), for example, points out that a condition that is invariably followed by some outcome may nonetheless be *irrelevant* to that outcome. Salt that has been hexed by a sorcerer invariably dissolves when placed in water, but hexing does not cause the salt to dissolve. Hexing does not make a difference for dissolution. Lewis's analysis, in contrast, gets this right. If the Salt had not been hexed, it would have dissolved anyways. Hence, the hexing is not a cause of the dissolving in water. Counterfactual accounts seem to obtain better results concerning the relevance of the cause on its effect.

Another problem of regularity analyses relates to the asymmetry of causation: if C causes E, then E will not also cause C. While C might belong to a minimal set of sufficient conditions for E when C is a genuine cause of E, this might also be true when C is an effect of E – an effect which could not have occurred, given the laws and the actual circumstances, except by being caused by E. Or it might be true when C and E are joint effects of a common deterministic cause. Or when C is a preempted potential cause of E – something that did not cause E, but would have done so if the actual cause had been absent. Regularity analyses have troubles to break the symmetry of events occurring in 'constant conjunction'. As a consequence, they cannot properly discern causes from their effects (or capture the asymmetry of causation) without further ado.

As we have already observed, the joint effects of a common cause may be used to illustrate the problem of spurious regularities. Here is such an example due to Jeffrey (1969), where a cause is regularly followed by two effects. Suppose that whenever the barometric pressure in a certain region drops below a certain level, two things happen. First, the height of the column of mercury in a particular barometer drops below a certain level. Shortly afterwards, a storm occurs. Then, it may well also be the case that whenever the column of mercury drops, there will be a storm. If so, a simple regularity theory would seem to rule that the drop of the mercury column causes the storm. In fact, however, the regularity relating these two events is spurious.

In light of the difficulties regularity analyses of causation face, Lewis (1973c, p. 557) writes:

> It remains to be seen whether any regularity analysis can succeed in distinguishing genuine causes from effects, epiphenomena [or joint effects], and preempted potential causes–and whether it can succeed without falling victim to worse problems, without piling on the epicycles, and without departing from the fundamental idea that causation is instantiation of regularities. I have no proof that regularity analyses are beyond repair, nor any space to review the repairs that have been tried. Suffice it to say that the prospects look dark. I think it is time to give up and try something else.

We throw our hat into the ring by proposing a more specific regularity analysis. A first step to improve upon simple regularity analyses may be found in the idea of a best system analysis of regularities. The idea can be traced back at least to Ramsey (1928/1978) and Ramsey (1929/1990, p. 242): "causal laws" are "consequences of those propositions which we should take as axioms if we knew everything and organized it as simply as possible in a deductive system." There is a trade-off between strength and simplicity of a deductive system, as Lewis (1994, p. 231-2) observed:

> Take all deductive systems whose theorems are true. Some are simpler, better systematized than others. Some are stronger, more informative, than others. These virtues compete: an uninformative system can be very simple, an unsystematized compendium of miscellaneous information can be very informative. The best system is the one that strikes as good a balance as truth will allow between simplicity and strength. [...] A regularity is a law iff it is a theorem of the best system.

From this a best system account of natural laws almost falls in place. A "contingent generalization is a *law of nature*", so Lewis (1973a, p. 73), "if and only if it appears as a theorem (or axiom) in each of the deductive systems that achieves a best combination of simplicity and strength." In Chapter 4, we supplement our analysis of causation by a best system account of generalisations.

Our novel analysis of causation can be regarded as a regularity analysis.[19] We define actual causation in terms of our strengthened Ramsey Test conditional $\gg$. The analysis can be stated in a simplified way as follows.

Definition 2. Strengthened Ramsey Test Analysis Simplified
Let C and E denote distinct events.[20] C is an actual cause of E relative to an agent's epistemic state S iff

(1) C and E are believed to obtain,

(2) $C \gg E$ is believed relative to S, and

(3) $\neg C \gg E$ is not believed relative to S.

The epistemic state S represents the believed facts and generalisations (or laws). $C \gg E$ is believed relative to S iff suspending judgment on the disjunction $C \vee E$ and, subsequently, assuming C lets the agent infer E. Let $-$ denote an operator on the epistemic state S that removes beliefs. Furthermore, let Cn and $\vdash$ denote the consequence operator and relation of propositional logic, respectively. Then, $C \gg E$ is believed relative to S iff $Cn(S) - (C \vee E), C \vdash E$.[21] Since S is closed under logical

[19] This being said, we think the best label for our analysis is an inferential analysis of causation based on generalisations.

[20] Again, C and E are propositions that represent or express events (or facts).

[21] Here is another simplification at work for expository reasons. Later, we enhance the epistemic state S with a priority ordering between beliefs.

consequence, subtracting the beliefs requires the agent also to suspend judgment on beliefs that, together with the believed facts and generalisations of S, entail $C \vee E$.[22] Let us call $Cn(S) - (C \vee E)$ the agent's agnostic epistemic state and abbreviate it by S'. The agnostic epistemic state S' represents the believed facts and generalisations after suspending judgment on the candidate cause and effect.

On the surface, our analysis resembles Lewis's (1973c) analysis of causation in terms of a certain kind of conditional. However, unlike Lewis's but like regularity analyses, the major work is done by generalisations, not by a similarity order between possible worlds (which is, as we have already seen tricky in many ways.) Rather, we rely on the basic inferential idea of regularity analyses: the generalisations, background conditions, and the cause entail the effect. The big conceptual difference to regularity analyses is our suspension of judgment. In Chapter 3, we define the suspension of judgment such that, by design, none of $C, \neg C, E, \neg E$ is believed in the agnostic state $S' = Cn(S) - (C \vee E)$.

Let us clear up why our analysis of causation is closely related to regularity analyses. Consider condition (2) of our analysis: $C \gg E$ is believed relative to S iff $Cn(S) - (C \vee E), C \vdash E$. The right-hand side is, by the deduction theorem of propositional logic, equivalent to $Cn(S) - (C \vee E) \vdash C \rightarrow E$. Compare the latter expression to condition (ii) of the Common Ground Regularity Analysis. The agnostic epistemic state S' (representing believed facts and generalisations) plays the same role as $\mathcal{L} \wedge \mathcal{F}$. Together with the respective first conditions of actuality, i. e. (i) and (1), both analyses say that an actual cause is sufficient for its effect. According to our analysis, an actual cause is, relative to your beliefs after suspension (represented by S'), sufficient to infer the effect. Furthermore, the agnostic epistemic state is designed such that the agent cannot infer the effect, that is $Cn(S) - (C \vee E) \nvdash E$. Hence, the actual cause is also, relative to S', necessary to infer the effect. This is a more specific implementation of condition (iii) in the Common Ground Regularity Analysis. Finally, the generalisations in the agnostic epistemic state S' are also necessary to infer the effect: without the generalisations $\mathcal{L}_{S'}$ believed in S', the agent cannot infer the effect E from the cause C (except when E follows logically from C). In symbols, $(Cn(S) - (C \vee E)) \setminus \mathcal{L}_{S'} \nvdash C \rightarrow E$. An agent needs at least *some* generalisations that figure as 'inference tickets' to derive E from C. This is a more specific implementation of condition (iv) in the Common Ground Regularity Analysis. Notice that our conditions (1) and (2) can be seen as a specification of conditions (i)-(iv).

Consider condition (3) of our analysis: $\neg C \gg E$ is not believed relative to S iff $Cn(S) - (C \vee E) \nvdash \neg C \rightarrow E$.[23] The condition requires that the putative effect cannot be inferred from the absence of the presumed cause, once the agent has suspended judgment. Condition (3) thus parallels a condition the Common Ground Regularity Analysis gets for free, that is $\mathcal{L} \wedge \mathcal{F} \nvdash \neg C \rightarrow E$ (follows from (ii) and (iii)).

The difference-making of condition (3) (and its sibling of the Common Ground Regularity Analysis) is weaker than counterfactual dependence. If C is an actual cause of E according to our analysis, then C makes a difference as regards E in the following sense: C lets us infer E, while $\neg C$ does not. It is not required, as counterfactual dependence has it, that if $\neg C$ were the case, $\neg E$ would be the case. It seems to be a conceptual truism about actual causation that not both an event C and its absence $\neg C$ are actual causes of E.[24]

We should note that our analysis of causation is epistemic. The ontic commitment in events (or facts) and regularities is replaced by beliefs in these events (or facts) and generalisations. To use

[22] In Chapter 3, we will argue for the further condition that certain generalisations are not subject to the suspension of judgment.

[23] In Chapter 3, we will see why the agent suspends judgment on $C \vee E$ rather than $\neg C \vee E$. In brief, the reason is that she suspends judgment on the literals she actually believes.

[24] See Sartorio (2005) for a defence of this weak condition of difference-making.

the Kantian metaphor, causality is projected onto the world by an epistemic subject. In the spirit of a regularity analysis, Kant (1781/87, p. A193/B238) maintains, that an effect follows its cause in accordance with a particular empirical law. Kant agrees with 'our' Hume that neither the relation of cause and effect nor the idea of necessary connection is given by our sensory perceptions only; both claim, however, that relation and connection are contributions, at least partly, of our mind.[25] We locate ourselves thus in the tradition of (a certain interpretation of) Hume, Kant, Ramsey (1929/1990), Gärdenfors (1988, Ch. 9), and, more recently, for instance, Spohn (2006). Gärdenfors (1988, p. 194), for example, confesses:

> For me, causality is primarily a *cognitive* concept. My position is Kantian to the extent that I believe it is a category mistake to try to give a 'realistic' or 'objective' interpretation of causality, where the causal relation holds among the real events, independently of minds having beliefs about the events. Thus, I interpret causal claims only *in relation to a given state of belief*.

An epistemic as opposed to realistic interpretation of causal relations strikes many philosophers as utterly implausible. However, we think that an analysis should be evaluated with respect to its achievements. As we will see in Chapter 4, our analysis of causation overcomes the problems of relevance and asymmetry, of joint effects, overdetermination, and conjunctive scenarios, as well as (early and late) preemption, switches, and double prevention. Moreover, we aim to solve the problem of spurious 'regularities' by supplementing our analysis by a best system account of generalisations.

Let us compare briefly and superficially our analysis of causation to one of the most advanced analyses of causation in the counterfactual tradition. Halpern and Pearl (2005) and Halpern (2015) define actual causation relative to a causal model, or equivalently to a model of structural equations. Hence, what counts as an actual causal relation depends on how the scenario under consideration is modelled. Similar to our agent-relativity, if you change the model you might get different causal relations.

The causal relata in Halpern and Pearl's (2005) and Halpern's (2015) accounts of actual causation are value assignments to (random) variables. Let C and E be such value assignments. Very roughly, their accounts say that C is an actual cause of E (relative to a causal model) iff C is *necessary* for E when keeping fixed certain other variables. In other words, E would not be the case if it were not for C (under certain contingencies). Which variables can be held fixed? The idea is that *any* set of variables can be kept fixed so long as holding fixed the other variables at the values they actually take does not make E false.

Our analysis of causation says, roughly, that C is an actual cause of E (relative to an agent) iff C is *sufficient* to infer E after suspending judgment on C and E. The suspension of judgment is tantamount to taking away the beliefs in C and E. In contrast to Halpern and Pearl's accounts, our analysis tells you which beliefs must be kept fixed, viz. the beliefs that do not let you infer C or E. You cannot just intervene on almost any set of variables to test for causal relations. Moreover, while their account allows the variables to be held fixed to take on non-actual (of counterfactual) values, our analysis does not. The beliefs remaining after the suspension of judgment are actual beliefs.

In sum, we put forth an epistemic analysis of causation in terms of a strengthened Ramsey Test conditional. The conditional, in turn, depends on generalisations that may be read as regularities. We amend our analysis by a weak condition of difference-making: if the actual cause were believed not to be the case, you could not infer the effect. Unlike Lewis (1973c), Halpern and Pearl (2005), and Halpern (2015), we do not rely on the stronger condition of difference-making, that is, if the cause had not been, the effect never had existed.

[25] For details, see Kant (1781/87, pp. A91-2/B123-4) and Kant (1783, §§ 27-30).

1.5 Preview

We have collected a number of problems which we aim to resolve in the following chapters. Here is a brief preview.

In Chapter 2, we ask Douven et al.'s question "How do we learn conditional information?" We propose, roughly, that we learn conditional information by updating on Stalnaker conditionals. When you update on a Stalnaker conditional $A > C$, you basically learn that the most similar world, in which A is true, is a world that satisfies C. The disadvantage of possible-worlds semantics relying on a similarity order is that it is far from clear how to specify this order explicitly. Regarding the learning of conditional information, however, this 'disadvantage' turns into the advantage that we can learn, at least in part, how to specify the similarity order between possible worlds upon receiving the conditional information.

As Lewis (1976) has shown, the updating method he names 'imaging' corresponds to the Stalnaker conditional: the probability value of a Stalnaker conditional is the probability value of its consequent upon imaging on its antecedent. According to our method, we learn a piece of conditional information by imaging on a certain proposition expressed by a Stalnaker conditional. We generalise Lewis's imaging to Jeffrey imaging in order to account also for the learning of uncertain conditional information. Thereby, we enter the arena dominated by Bayesian approaches. Unlike these approaches, our method based on Jeffrey imaging correctly predicts the rational learning outcomes in all of Douven's (2012) benchmark examples and Van Fraassen's (1981) Judy Benjamin Problem.

We go on to adapt the method of learning uncertain conditional information to a method of learning uncertain causal information. We do so by applying Lewis's (1973c) notion of causal dependence to Stalnaker's semantics. This move will allow us to implement Douven's (2012) abductive conception with respect to the learning of conditionals: when learning a conditional, the explanatory power of the antecedent with respect to the consequent determines the resulting probability of the antecedent. The combination of the methods provides a unified framework within which we can distinguish between the learning of conditional, causal, and conjunctive information. (We sketch in Section 2.4 how our method could also cover the learning of subjunctive conditionals.)

In Chapter 3, we transition from the learning of conditional information to the semantics of conditionals. We aim to overcome the problem that extant Ramsey Test semantics do not account for a proper relation of relevance between the antecedent and consequent, although such a conditional connection seems to be necessary to plausibly accept a conditional. We establish a relation of relevance by strengthening Ramsey's test. The strengthening is inspired by both Ramsey's (1929/1990) original remarks on conditionals and Rott's (1986) strengthened Ramsey Test. Within the framework of belief bases, our strengthened Ramsey Test will prove to be asymmetric for a wide range of cases. That is, in many cases, the agent can infer the consequent from the antecedent but not the antecedent from the consequent. After illustrating the asymmetry of our strengthened Ramsey Test conditional by a generalised tower-shadow example, we use the conditional to provide an analysis of 'because' as it figures in explanations.

In Chapter 4, we aim to analyse actual causation in terms of our strengthened Ramsey Test, where the strengthened conditional is conceived of as expressing a relation of production. The idea is that the strengthened Ramsey Test allows us to verify or falsify that an event actually brings about another event. Our concept of causation as production is reductive and solves the problems posed by cases of overdetermination, preemption, switches, and double prevention. Moreover, we aim to solve the problem of spurious causation by a best system account of generalisations.

In Chapter 5, we translate the analysis of actual causation provided in the previous chapter into the framework of causal models due to Halpern and Pearl (2005). Although the translation simplifies our

analysis considerably, we lose its reductiveness. A comparison of our analysis to Halpern and Pearl's (2005) and Halpern's (2015) definition of actual cause reveals that they (but not we) still struggle with any set of problems including both overdetermination and conjunctive scenarios.

In Chapter 6, we refine Halpern and Pearl's (2005) definition of actual cause such that disjunctions are admissible candidates for actual causes. This endeavour is motivated by Sartorio's (2006) argument for the existence of disjunctive causes. She puts forth a switching scenario that suggests, against Lewis (1986b) and Halpern and Pearl (2005), the following observation: a disjunctive cause does not imply that one of its disjuncts *must* also be a cause.

In Chapter 7, we leave the analyses of causation behind and turn to a rather epistemological question in the philosophy of cognitive neuroscience. We aim to determine reasonable requirements for cognitive neuroscientists to speak of causation as regards the mind-brain interaction. More specifically, we aim to make the characteristics of causation, as assumed in cognitive neuroscience, explicit. Subsequently, we impose these 'demands of causality' upon the account of interventionism put forth by Woodward (2005) and Woodward (2015). Within the resulting framework, we investigate to what extent we are justified to derive causal relations between mental properties and properties of the brain, depending on which scientific methods are used in the neuroscientific studies.

Chapter 2

Learning Conditional Information by Jeffrey Imaging

In this chapter, we propose a method of learning conditional information. In a nutshell, an agent learns conditional information by Jeffrey imaging on the minimally informative proposition expressed by a Stalnaker conditional. We show that the predictions of the proposed method align with the intuitions in Douven's (2012) benchmark examples. Jeffrey imaging on Stalnaker conditionals can also capture the learning of uncertain conditional information, which we illustrate by generating predictions for the Judy Benjamin Problem.

Subsequently, we adapt the method of learning uncertain conditional information to a method of learning uncertain causal information. The idea behind the adaptation parallels Lewis's (1973c) analysis of causal dependence. The combination of the methods provides a unified account of learning conditional and causal information that manages to clearly distinguish between conditional, causal, and conjunctive information. The ensuing account shows that the learning of uncertain conditional and/or causal information may be modelled by Jeffrey imaging on Stalnaker conditionals.

Sources. This chapter builds on Günther (2018) and Günther (2017a). Substantial content of the first paper is reprinted by permission from Springer Nature: Springer Netherlands, *Journal of Philosophical Logic*, Learning Conditional Information by Jeffrey Imaging on Stalnaker Conditionals, Günther, M., License Number 4324300410472 (2017), advance online publication, 15 November 2017 (https://doi.org/10.1007/s10992-017-9452-z, J Philos Logic). Substantial content from the second paper is reprinted by permission from *Organon F*: Institute of Philosophy of the Slovak Academy of Sciences, Learning Conditional and Causal Information by Jeffrey Imaging on Stalnaker Conditionals, Günther, M. (2017).

2.1 Introduction

"How do we learn conditional information?" Douven et al. present this question for consideration in a series of papers.[1] Douven (2012) contains a survey of the available accounts that model the learning of conditional information. The survey comes to the conclusion that a general account of probabilistic belief updating by learning (uncertain) conditional information is still to be formulated. Douven and Pfeifer (2014) analyses the state of the art even more pessimistically by writing that "no one seems to have an idea of what an even moderately general rule of updating on conditionals might look like", even if we restrict the scope of the account to indicative conditionals.[2]

Douven (2012) dismisses the Stalnaker conditional as a means to model the learning of conditional information. He argues for the dismissal by pointing out that the Stalnaker conditional "makes no predictions at all about any of our examples".[3] Douven provides three possible worlds models for his point. Each model consists of four worlds such that all logical possibilities of two binary variables are covered. He observes that imaging on "If α, then γ" interpreted as a Stalnaker conditional has different effects: in model I the probability of the antecedent α, i. e. $P(\alpha)$ decreases, in model II $P(\alpha)$ remains unchanged, and in model III $P(\alpha)$ increases. According to Douven this flexibility of the class of possible worlds models is a problem rather than an advantage, since there were no rationality constraints to rule out certain models as rational representations of a belief state.

Pace Douven, we show that his dismissal of the Stalnaker conditional is unjustified by proposing an updating method based on the Stalnaker semantics and inspired by Lewis's imaging method. The core idea of the proposed method is that an agent learns conditional information by Jeffrey imaging on the minimally informative meaning of the corresponding Stalnaker conditional. The method succeeds in modeling the three examples Douven takes as benchmark for an account of learning conditional information. In addition, Jeffrey imaging, our generalisation of Lewis's imaging method, accounts for the learning of uncertain conditional information, as we will illustrate by applying our learning method to Van Fraassen's (1981) Judy Benjamin Problem.

In Section 2.2, we propose our probabilistic method of learning indicative conditional information. First, we introduce the concepts of a Stalnaker conditional, Lewis's imaging, and our generalisation thereof. Based on these concepts, we supplement the properties of a Stalnaker model's similarity order by the minimally informative interpretation of a Stalnaker conditional and a default assumption. We justify both by the rationale that belief changes should be as conservative as possible. We show that the supplemented Stalnaker models provide sufficient constraints to model the learning of indicative conditional information by applying the learning method to Douven's examples as well as the Judy Benjamin Problem. Thereby we recover possible worlds approaches from Douven's dismissal.

2.2 A Probabilistic Method of Learning Indicative Conditional Information

The proposed learning method may be summarised as follows. (i) We model an agent's belief state as a Stalnaker model. (ii) The agent learns conditional information by (ii).(a) interpreting the received conditional information as a Stalnaker conditional, (ii).(b) constraining the similarity order by the meaning of the Stalnaker conditional in a minimally informative way and in presence of the

[1] See Douven and Dietz (2011), Douven and Romeijn (2011), Douven (2012), Douven and Pfeifer (2014, especially Section 6).

[2] Douven and Pfeifer (2014, p. 213).

[3] Douven (2012, p. 247).

default assumption, and (ii).(c) updating her degrees of belief by Jeffrey imaging on this Stalnaker conditional (together with further contextual information, if available). (iii) We check whether or not the result of (Jeffrey) imaging complies with the correct intuitions associated with the scenario under consideration.

In Section 2.2.1, we introduce the meaning of a Stalnaker conditional. In Section 2.2.2, we present Lewis's updating method called 'imaging', which relates the probability of a Stalnaker conditional and the probability of the consequent after imaging on the antecedent. In Section 2.2.3, we generalise Lewis's imaging in order to model cases of learning uncertain conditional information. In Section 2.2.4, we describe our method of learning conditional information in more detail, before we apply the method, in Section 2.2.6, to Douven's examples and the Judy Benjamin Problem. In Section 2.2.5, we provide a rationale for the minimally informative interpretation of Stalnaker conditionals and the default assumption for learning conditional information.

2.2.1 The Stalnaker Conditional

The idea behind a Stalnaker conditional may be expressed as follows: a Stalnaker conditional $\alpha > \gamma$ is true at a world w iff γ is true in the most similar possible world w' to w, in which α is true.[4] The evaluation of a Stalnaker conditional requires a model of possible worlds. A model of possible worlds, in turn, requires the specification of a logical language.

Definition 3. Full Conditional Language
Let $Prop$ be the set of atomic propositions. Then $\mathcal{L}$ be a set of formulas such that

(i) for each $p_1, p_2, ... \in Prop$, $p_i \in \mathcal{L}$,

(ii) if $\alpha, \gamma \in \mathcal{L}$, then $\neg\alpha \in \mathcal{L}$ and $\alpha \wedge \gamma \in \mathcal{L}$,

(iii) if $\alpha, \gamma \in \mathcal{L}$, then $\alpha > \gamma \in \mathcal{L}$,

(iv) and no other expressions are in $\mathcal{L}$.

We say that $\mathcal{L}$ is the full conditional language.

The full conditional language $\mathcal{L}$ contains any type of Boolean combination of conditionals, e.g. $(\alpha > \gamma) \wedge \beta \in \mathcal{L}$, and arbitrary nestings of conditionals, e.g. $\alpha > (\gamma \wedge (\beta > \delta))$.

Let w denote a Boolean assignment, or equivalently a possible world. We denote the set of Boolean assignments, or equivalently the set of possible worlds, that satisfy a formula α by $[\alpha]$. We thus identify the set $[\alpha]$ with the proposition expressed by α. In symbols, $[\alpha] = \{w \in W \mid w(\alpha) = 1\}$.

Definition 4. Stalnaker Model
We say that $\mathcal{M}_{St} = \langle W, R, \leq, V\rangle$ is a Stalnaker model iff

(i) W is a non-empty set of possible worlds,

(ii) $R : W \times W$ is a binary accessability relation over worlds such that:

(a) for all $w \in W$: wRw. (Reflexivity)

[4] Cf. Stalnaker (1975). Note that Stalnaker's theory of conditionals aims to account for both indicative and counterfactual conditionals. Besides some cursory remarks in Section 2.4, we set the complicated issue of this distinction aside in this chapter. However, we want to emphasise that Douven's examples and the Judy Benjamin Problem only involve indicative conditionals.

(iii) $\leq$ assigns each $w \in W$ a total order $\leq_w$ such that:

(a) for all $w, w', w'' \in W$: if $w' \leq_w w''$ and wRw'', then wRw'.

(b) for all $w, w' \in W$: $w' \leq_w w$, only if $w' = w$. (Unique Center Assumption)

(c) for all $w, w', w'' \in W$: $w' \leq_w w''$ or $w'' \leq_w w'$. (Connectivity)

(d) for all $w, w' \in W$: if wRw' for some $w' \in [\alpha] \subseteq W$, then there is a $w'' \in [\alpha]$ such that $w'' <_w w'''$ for all $w''' \in [\alpha]$.[5] We say that w'' is the unique α-world minimal under $\leq_w$. In symbols, $w'' = min_{\leq_w}[\alpha]$. (Stalnaker's Uniqueness Assumption)

(iv) V is an evaluation function of the full conditional language $\mathcal{L}$ iff

(a) $\forall p \in Prop, \forall w \in W : V(p, w) = 1$ iff $w(p) = 1$ iff $w \vDash p$,

(b) and $\forall \alpha, \gamma$ and $\forall w \in W$:

i. $w \vDash \neg\alpha$ iff $w \nvDash \alpha$

ii. $w \vDash \alpha \wedge \gamma$ iff $w \vDash \alpha$ and $w \vDash \gamma$

iii. $w \vDash \alpha > \gamma$ iff $min_{\leq_w}[\alpha] \vDash \gamma$ if there is a $min_{\leq_w}[\alpha]$.

We comment on two aspects of Definition 4. (I) We presented an equivalent variant to Stalnaker's original models that emphasizes the similarity order $\leq$ between possible worlds, instead of a world selection function.[6] We interpret the world $w' = min_{\leq_w}[\alpha]$ as the most similar α-world from w. (II) The accessability relation R is connective due to (iii).(a), (iii).(c), and (iii).(d) of Definition 4.

Now, we can state more precisely the meaning of a Stalnaker conditional. "If α, then γ" denotes according to Stalnaker's proposal the set of worlds (or equivalently the proposition) containing each world whose most similar α-world is a world that satisfies γ. In symbols, $[\alpha > \gamma] = \{w \mid w \vDash \alpha > \gamma\} = \{w \mid min_{\leq_w}[\alpha] = \emptyset$ or $min_{\leq_w}[\alpha] \vDash \gamma\}$.

Finally, note that any Stalnaker model validates the principle called 'Conditional Excluded Middle' according to which $(\alpha > \gamma) \vee (\alpha > \neg\gamma)$. The reason is that, for any $w \in W$, the single most similar α-world $min_{\leq_w}[\alpha]$ is either a γ-world, or else a $\neg\gamma$-world. This principle will come in handy when modeling the learning of conditional information with uncertainty. In Section 2.2.6, we will apply our method to the learning of uncertain conditional information. First, however, we introduce Lewis's imaging method and our generalisation thereof.

2.2.2 Lewis's Imaging

Lewis (1976) developed a probabilistic updating method called 'imaging'. In order to present this updating method, we introduce a notational shortcut: for each world w and each (possible) antecedens α, $w_\alpha = min_{\leq_w}[\alpha]$ be the most similar world of w such that $w_\alpha(\alpha) = 1$. Invoking the shortcut, we can then specify the truth conditions for Stalnaker's conditional operator $>$ as follows.

$$w(\alpha > \gamma) = w_\alpha(\gamma), \text{ if } \alpha \text{ is possible.}^7 \tag{2.1}$$

Definition 5. Probability Space over Possible Worlds
We call $\langle W, \wp(W), P\rangle$ a probability space over a finite set of possible worlds W iff

(i) $\wp(W)$ is the set of all subsets of W,

[5] Here as elsewhere in the paper, the strict relation $w' <_w w''$ is defined as $w' \leq_w w''$ and $w'' \not\leq_w w'$.

[6] For Stalnaker's presentation of his semantics see Stalnaker and Thomason (1970).

[7] We assume here that there are only finitely many worlds. Note also that if α is possible, then there exists some w_α.

(ii) and $P : \wp(W) \mapsto [0, 1]$ is a probability measure, i.e.

(a) $P(W) = 1, P(\emptyset) = 0$,

(b) and for all $X, Y \subseteq W$ such that $X \cap Y = \emptyset$, $P(X \cup Y) = P(X) + P(Y)$.

As before, we conceive of the elements of $\wp(W)$ as propositions. We define, for each α, $P(\alpha) = P([\alpha])$. We see that W corresponds to an arbitrary tautology denoted by $\top$ and $\emptyset$ to an arbitrary contradiction denoted by $\bot$. Definition 5 allows us to understand a probability measure P as a probability distribution over worlds such that each w is assigned a probability $P(w) > 0$, and $\sum_w P(w) = 1$. We may determine the probability of a formula α by summing up the probabilities of the worlds at which the formula is true.[8]

$$P(\alpha) = \sum_w P(w) \cdot w(\alpha) \tag{2.2}$$

Now, we are in a position to define Lewis's updating method of imaging.

Definition 6. Imaging (Lewis (1976, p. 310))
For each probability function P, and each possible formula α, there is a probability function P^α such that, for each world w', we have:

$$P^\alpha(w') = \sum_w P(w) \cdot \left\{ \begin{array}{ll} 1 & \text{if } w_\alpha = w' \\ 0 & \text{otherwise} \end{array} \right\} \tag{2.3}$$

We say that we obtain P^α by imaging P on α, and call P^α the image of P on α.

Intuitively, imaging transfers the probability of each world w to the most similar α-world w_α. Importantly, the probabilities are transfered, but in total no probability mass is additionally produced and no probability mass is lost. In formal terms, we have always $\sum_{w'} P^\alpha(w') = 1$. Any α-world w' keeps at least its original probability mass (since then $w_\alpha = w'$), and is possibly transfered additional probability shares of $\neg\alpha$-worlds w iff $min_{\leq_w}[\alpha] = w'$. In other words, each α-world w' receives as its updated probability mass its previous probability mass plus the previous probability shares that were assigned to $\neg\alpha$-worlds w such that $min_{\leq_w}[\alpha] = w'$. In this way, the method of imaging distributes the whole probability onto the α-worlds such that $P^\alpha(\alpha) = \sum_{w(\alpha)=1} P(w(\alpha)) = 1$, and each share remains 'as close as possible' at the world at which it has previously been located. For an illustration see Figure 2.1.

[8] We assume here that each world is distinguishable from any other world, i. e. for two arbitrary worlds, there is always a formula in $\mathcal{L}$ such that the formula is true in one of the worlds, but false in the other. In other words, we consider no copies of worlds.

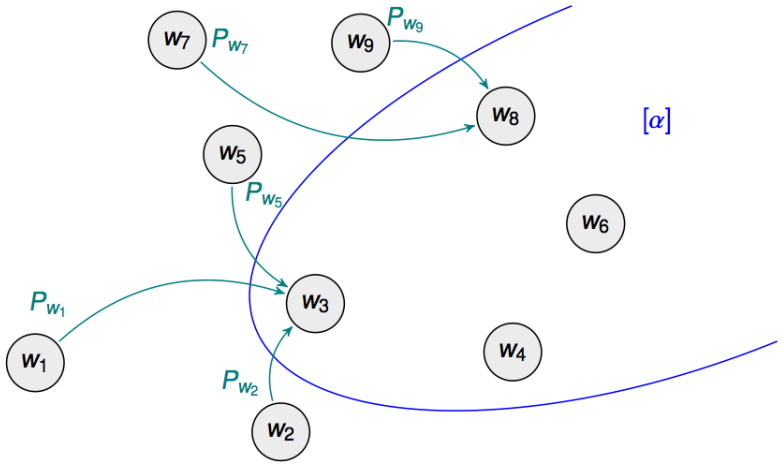

Figure 2.1: A set of possible worlds. The blue area represents the proposition or set of worlds $[\alpha] = \{w_3, w_4, w_6, w_8\}$. The teal arrows represent the transfer of probability shares from the respective $[\neg\alpha]$-worlds to their most similar $[\alpha]$-world. Similarity is graphically represented by topological distance between the worlds such that w_3, for instance, is the most similar or 'closest' $[\alpha]$-world to w_2.

Lewis proved the following theorem, which will be useful to model the respective examples.

Theorem 1. (Lewis (1976, p. 311))
The probability of a Stalnaker conditional equals the probability of the consequent after imaging on the antecedent, i. e. $P(\alpha > \gamma) = P^{\alpha}(\gamma)$, if α is possible.

Proof. Suppose we obtain P^{α} by imaging P on α. Consider some formula $\gamma \in \mathcal{L}$.

$$
\begin{aligned}
P^{\alpha}(\gamma) &= \sum_{w_\alpha} P(w_\alpha) \cdot w_\alpha(\gamma), \quad \text{(2.2) applied on } P^{\alpha} \\
&= \sum_{w_\alpha} \left(\sum_{w} P(w) \cdot \left\{ \begin{array}{ll} 1 & \text{if } w_\alpha = w \\ 0 & \text{otherwise} \end{array} \right\} \right) \cdot w_\alpha(\gamma), \quad \text{Definition 6} \\
&= \sum_{w} P(w) \cdot \left(\sum_{w_\alpha} \left\{ \begin{array}{ll} 1 & \text{if } w_\alpha = w \\ 0 & \text{otherwise} \end{array} \right\} \cdot w_\alpha(\gamma) \right), \quad \text{Algebra} \\
&= \sum_{w} P(w) \cdot w_\alpha(\gamma), \quad \text{Simplification of the inner sum} \\
&= \sum_{w} P(w) \cdot w(\alpha > \gamma), \quad \text{(2.1)} \\
&= P(\alpha > \gamma), \quad \text{(2.2)}
\end{aligned}
$$

□

Note that α in Theorem 1 may itself be of conditional form $\beta > \delta$ for any $\beta, \delta \in \mathcal{L}$.

2.2.3 Jeffrey Imaging

For the case of learning uncertain conditional information, i. e. $P(\alpha > \gamma) = k$ for $k \in [0, 1]$ but unequal to 0 or 1, we need to generalise Lewis's imaging method of Definition 6. In analogy to Jeffrey

conditionalisation, we call the generalised method 'Jeffrey' imaging.[9] Jeffrey imaging is based on Lewis's imaging and the fact that in a Stalnaker model the principle of Conditional Excluded Middle prescribes that $\neg(\alpha > \gamma) \equiv \alpha > \neg\gamma$. We know, for all $w \in W$, presupposed $\alpha > \gamma$ is possible, both (I) that $\sum_w P^{\alpha>\gamma}(w)$ sums up to 1 and (II) that $\sum_w P^{\alpha>\neg\gamma}(w)$ sums up to 1. The idea is that if we form a weighted sum over the terms of (I) and (II) with some parameter $k \in [0, 1]$, then we obtain again a sum of terms $P_k^{\alpha>\gamma}(w)$ such that $\sum_w P_k^{\alpha>\gamma}(w) = 1$. Note, however, that we present the more general case $P_k^{\alpha}(w)$ in the definition below.

Definition 7. Jeffrey Imaging
For each probability function P, each possible formula $\alpha \in \mathcal{L}$ (possibly of conditional form $\beta > \delta$), and some parameter $k \in [0, 1]$, there is a probability function P_k^{α} such that, for each world w' and the two similarity orderings centred on w_α and $w_{\neg\alpha}$, we have:

$$P_k^{\alpha}(w') = \sum_w \left(P(w) \cdot \left\{ \begin{array}{ll} k & \text{if } w_\alpha = w' \\ 0 & \text{otherwise} \end{array} \right\} + P(w) \cdot \left\{ \begin{array}{ll} (1-k) & \text{if } w_{\neg\alpha} = w' \\ 0 & \text{otherwise} \end{array} \right\} \right) \tag{2.4}$$

We say that we obtain P_k^{α} by Jeffrey imaging P on $\alpha \in \mathcal{L}$, and call P_k^{α} the Jeffrey image of P on α. Note that in the case where $k = 1$, Jeffrey imaging reduces to Lewis's imaging.

Theorem 2. Properties of Jeffrey Imaging

(i) $\sum_{w'} P_k^{\alpha}(w') = 1$

(ii) $P_k^{\alpha}(\alpha) = k$

(iii) $P_k^{\alpha}(\neg\alpha) = (1 - k)$

(iv) $P_k^{\alpha}(\gamma) = k \cdot P(\alpha > \gamma)$

Proof. (i)

$$\begin{aligned}
\sum_{w'} P_k^{\alpha}(w') &= \sum_{w'} \sum_w \left(P(w) \cdot \left\{ \begin{array}{ll} k & \text{if } w_\alpha = w' \\ 0 & \text{otherwise} \end{array} \right\} + \right. \\
&\quad \left. P(w) \cdot \left\{ \begin{array}{ll} (1-k) & \text{if } w_{\neg\alpha} = w' \\ 0 & \text{otherwise} \end{array} \right\} \right) && \text{Algebra} \\
&= \sum_{w'} \left(k \cdot \sum_w P(w) \cdot \left\{ \begin{array}{ll} 1 & \text{if } w_\alpha = w' \\ 0 & \text{otherwise} \end{array} \right\} + \right. \\
&\quad \left. (k-1) \cdot \sum_w P(w) \cdot \left\{ \begin{array}{ll} 1 & \text{if } w_{\neg\alpha} = w' \\ 0 & \text{otherwise} \end{array} \right\} \right) && \text{by Def. 6, } \sum_{w'} \sum_w P(w) = 1 \\
&= k + (k-1) = 1 && \text{for all } k \in [0, 1]
\end{aligned} \tag{2.5}$$

[9] Cf. Jeffrey (1965). In personal communication, Benjamin Eva and Stephan Hartmann mentioned that the idea behind Jeffrey imaging is already used in artificial intelligence research to model the retrieval of information. Sebastiani (1998, p. 3) mentions the name 'Jeffrey imaging' without writing down a corresponding formula. Crestani (1998, p. 262) says that Sebastiani (1998) suggested "a new variant of standard imaging called *retrieval by Jeffrey's logical imaging*". However, the formalisation of Jeffrey's idea on p. 263 differs from mine in at least two respects. (i) An additional truth evaluation function occurs in the formalisation for determining whether a formula (i. e. 'query') is true at a world (i. e. 'term'). (ii) Instead of a parameter k locally gouverning the probability kinematics of each possible world, Crestani simply uses a global constraint on the posterior probability distribution.

(ii)

$$
\begin{aligned}
P_k^{\alpha}(\alpha) &= \sum_{w_\alpha} P_k^{\alpha}(w_\alpha) \cdot w_\alpha(\alpha) && \text{Definition 7} \\
&= \sum_{w_\alpha} \Bigg(\sum_{w} P(w) \cdot \left\{ \begin{array}{ll} k & \text{if } w_\alpha = w_\alpha \\ 0 & \text{otherwise} \end{array} \right\} + \\
&\sum_{w} P(w) \cdot \left\{ \begin{array}{ll} (1-k) & \text{if } w_{\neg\alpha} = w_\alpha \\ 0 & \text{otherwise} \end{array} \right\} \Bigg) \cdot w_\alpha(\alpha) && \text{Second term cancels out} \\
&= \sum_{w} P(w) \sum_{w_\alpha} \cdot \left\{ \begin{array}{ll} k & \text{if } w_\alpha = w_\alpha \\ 0 & \text{otherwise} \end{array} \right\} \cdot w_\alpha(\alpha) && \text{Algebra} \\
&= k \cdot \sum_{w} P(w) \cdot w_\alpha(\alpha) && \text{Algebra and (2.2)} \\
&= k \cdot P^{\alpha}(\alpha) = k \cdot P(\alpha > \alpha) = k
\end{aligned}
\tag{2.6}
$$

(iii) Obvious.
(iv) Replace in (ii) the 'consequent' α by γ.

□

We see that in total the revision method of Jeffrey imaging does neither produce additional probability shares, nor destroy any probability shares. In contrast to Lewis's imaging, Jeffrey imaging does not distribute the whole probabilistic mass onto the α-worlds, but only a part thereof that is determined by the parameter k. In particular, as compared to Lewis's imaging, Jeffrey imaging may be understood as implementing a more moderate or balanced movement of probabilistic mass between α- and $\neg\alpha$-worlds. For an illustration see Figure 2.2.

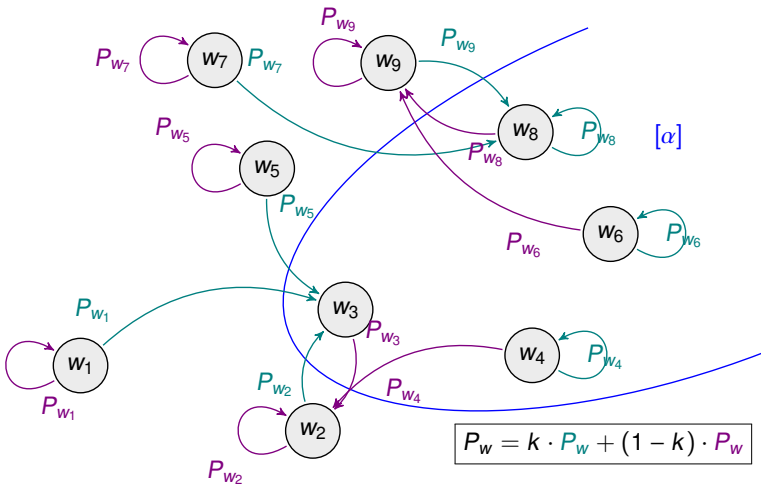

Figure 2.2: An illustration of the probability kinematics of Jeffrey imaging. The Jeffrey image P_k^{α} is characterised by a 'k-inertia' of the probabilistic mass from the respective α-worlds, and a '$(1-k)$-inertia' of the probabilistic mass from the respective $\neg\alpha$-worlds. Each teal arrow represents the transfer of the probability share $k \cdot P(w)$ to the closest α-world from w. Each violet arrow represents the transfer of the probability share $(1-k) \cdot P(w)$ to the closest $\neg\alpha$-world from w.

It is easy to show that P_k^α is a probability function. In a possible worlds framework, such a proof basically amounts to showing that the probability shares of all the worlds sum up to 1 after Jeffrey imaging. Therefore, property (i) of Theorem 2 provides minimal justification for applying Jeffrey imaging to probabilistic belief updating.

2.2.4 A Simple Method of Learning Conditional Information

We outline now a method of learning conditional information in three main steps.

(i) We model an agent's belief state as a Stalnaker model $\mathcal{M}_{St} = \langle W, R, \leq, V\rangle$ such that all and only those logical possibilities are represented as single worlds, which are relevant to the scenario under consideration. For instance, if only a single conditional "If α, then γ" is relevant and nothing else, then W contains exactly four elements as depicted in Figure 2.3.[10]

(ii) An agent learns conditional information "If α, then γ" iff (a) the agent interprets the received conditional information as a Stalnaker conditional $\alpha > \gamma$, (b) changes the similarity order $\leq$ by the meaning of $\alpha > \gamma$ in a minimally informative way and respecting the default assumption, and (c) updates her degrees of belief by Jeffrey imaging on the minimally informative meaning of $\alpha > \gamma$.

(iii) Finally, we check whether or not the result of Jeffrey imaging obtained in step (ii).(c) corresponds to the intuition associated with the respective example.

Step (ii) constitutes the core of the learning method and requires further clarification.

(a) In the agent's belief state, i. e. a Stalnaker model, the received information is interpreted. In the case of conditional information, the received information is interpreted as Stalnaker conditional. Hence, if the agent receives the information "If α, then γ", she interprets the information as meaning that the most similar α-world (from the respective actual world) is a world that satisfies γ (presupposed α is possible). Technically, the interpretation (i. e. the meaning) of $\alpha > \gamma$ relative to the Stalnaker model $\mathcal{M}_{St}$ is the proposition $[\alpha > \gamma] = \{w \in W \mid min_{\leq_w}[\alpha] \in [\gamma]\}$, where w is the respective actual world.

(b) The similarity order(s) is/are changed upon receiving conditional information. The proposition $\{w \in W \mid min_{\leq_w}[\alpha] \in [\gamma]\}$ depends on the similarity order $\leq$. The learning method prescribes that $\leq$ is specified, or adjusted, such that from each world the most similar α-world is a γ-world whenever *possible*. In other words, the method demands a maximally conservative, or equivalently minimally informative, interpretation of the received information. This amounts to specifying or adjusting the orders $\leq_w$ such that *as many worlds as possible* satisfy the received information. On the one hand, we can describe this interpretation as maximally conservative in the sense that no worlds are gratuitously excluded. On the other hand, we may think of possible worlds as information states. Then the exclusion of possible worlds corresponds to a gain of information. If an agent interprets the received information in a maximally conservative way, then as few as possible worlds or information states are excluded. In this sense, her gain of information is minimal. We will also use the abbreviation $[\alpha > \gamma]_{min}$ for the minimally informative proposition expressed by $\alpha > \gamma$.

[10] In other words, we consider "small" possible worlds models and do not allow for copies of worlds, i. e. worlds that satisfy the same formulas.

The learning method prescribes that the agent changes her similarity order respecting a default assumption. This default assumption states that the most similar $\alpha > \gamma$-world from any excluded $\alpha > \neg\gamma$-world is a $\alpha \wedge \gamma$-world, if there is more than one candidate. Formally, this constraint expresses that $min_{\leq_{w(\alpha>\neg\gamma)=1}}[\alpha > \gamma] \vDash \alpha \wedge \gamma$, if $min_{\leq_{w(\alpha>\neg\gamma)=1}}[\alpha > \gamma]$ is underdetermined. ($min_{\leq_{w(\alpha>\neg\gamma)=1}}[\alpha > \gamma]$ denotes the respectively most similar $\alpha > \gamma$-world from any world w such that $w(\alpha > \neg\gamma) = 1$ given the minimally informative similarity order $\leq$.) Notice the interplay between the two constraints for the similarity order: the minimally informative interpretation of the Stalnaker conditional minimises the number of worlds which might have several candidates for the most similar $\alpha > \gamma$-world and, based on this interpretation, the default assumption determines the most similar $[\alpha > \gamma]$-world for each world having more than one candidate world (at least in absence of further information).

(c) Jeffrey imaging is applied on the minimally informative meaning of the Stalnaker conditional $\alpha > \gamma$. The application of Jeffrey imaging determines a probability distribution after learning the (uncertain) conditional information.

We note that the proposed learning method has the following properties resembling *modus ponens* and *modus tollens*. If the agent already knows $[\alpha]$ (or $[\neg\gamma]$ resp.), learning the minimally informative proposition $[\alpha > \gamma]$ implies that the agent also knows $[\gamma]$ (or $[\neg\alpha]$ resp.).

2.2.5 A Rationale for the Minimally Informative Interpretation and the Default Assumption

We aim to justify the minimally informative interpretation of conditional information and the default assumption by the following rationale: an agent should change her belief state as conservatively as possible when learning a proposition. We say that a belief change is as conservative as possible, or equivalently maximally conservative, iff the change is not stronger than necessary to believe the received information.[11]

We argue first that the rationale of maximally conservative belief change warrants the minimally informative interpretation of the received conditional information. The learning of an indicative conditional is constraint by its meaning. By Stalnaker's semantics, an indicative conditional $\alpha > \gamma$ means that γ is the case on the supposition of α. Hence, upon learning the indicative conditional (and having or receiving no more information), the agent does not necessarily learn whether or not α is the case, and *a fortiori* whether or not γ is the case. However, the agent learns at least that the conjunction $\alpha \wedge \neg\gamma$ is not possible. Correspondingly, the minimally informative meaning of $[\alpha > \gamma]$, or equivalently $[\alpha > \gamma]_{min} = [(\alpha \wedge \gamma) \vee (\neg\alpha \wedge \gamma) \vee (\neg\alpha \wedge \neg\gamma)]$, is exhausted by all those possible worlds that verify the material implication.[12] Consequently, $[\neg\alpha] \subset [\alpha > \gamma]_{min}$ and $[\gamma] \subset [\alpha > \gamma]_{min}$, which means that the minimally informative proposition $[\alpha > \gamma]$ is less informative than $[\neg\alpha]$ and $[\gamma]$, respectively. When learning $[\alpha > \gamma]_{min}$ and having or receiving no more information, an agent learns only that the $\alpha \wedge \neg\gamma$-world is excluded, whereas the learning remains silent on the status of the $\neg\alpha$-worlds. Learning only $[\alpha > \gamma]_{min}$ thus qualifies as a maximally conservative belief change: the agent believes the conditional information $[\alpha > \gamma]$ without believing any of the more informative propositions $[\alpha], [\neg\alpha], [\gamma]$ and/or $[\neg\gamma]$.

[11] For proposals and justifications of a similar rationale, see Gärdenfors (1988) and Van Benthem and Smets (2015). For a critical and elucidating discussion of the principle of minimal or conservative belief change, see Rott (2000).

[12] Here the question may arise why we do not simply learn conditional information by Jeffrey imaging on the material implication. A short answer will be provided in Section 2.3.

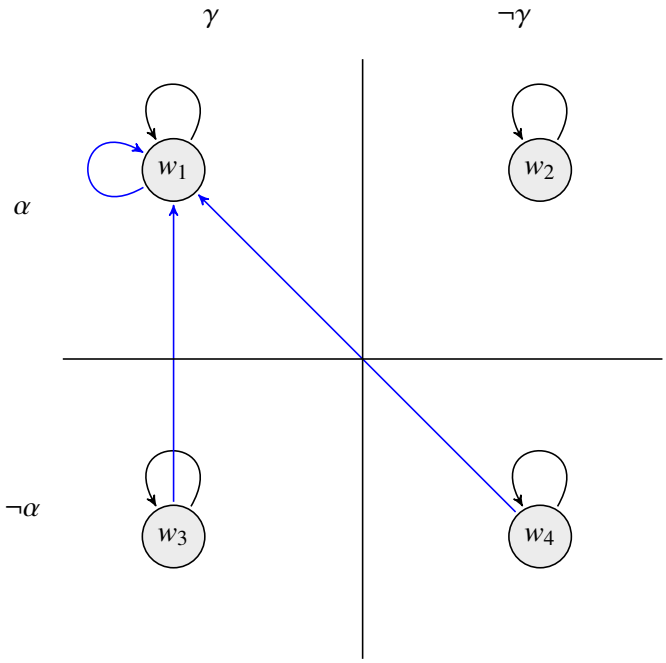

Figure 2.3: A four-worlds Stalnaker model for a case, in which the only received and relevant information is "If α, then γ". The reflexive arrows illustrate that each world w is the most similar to itself under the respective similarity order $\leq_w$. The blue arrows illustrate the change of the similarity order such that the received and interpreted information $[\alpha > \gamma]$ is minimally informative. Here, the minimally informative meaning of $\alpha > \gamma$ is $[\alpha > \gamma] = \{w \in W \mid w \vDash \alpha > \gamma\} = \{w_1, w_3, w_4\}$. Note that according to (iii).(b) and (iii).(d) of Definition 4 world w_2 is its own most similar α-world, but does not satisfy γ, i. e. $min_{\leq_{w_2}}[\alpha] \nvDash \gamma$ and thus $min_{\leq_{w_2}}[\alpha > \gamma] \neq w_2$. Relying on the default assumption of step (ii).(b), $min_{\leq_{w_2}}[\alpha > \gamma] = w(\alpha \wedge \gamma) = w_1$. In words, the method prescribes that w_1 is the most similar $\alpha > \gamma$-world to w_2, which is excluded under the minimally informative meaning of $\alpha > \gamma$. This illustrates that the minimally informative meaning $[\alpha > \gamma]$ implies that $\neg\gamma$ is excluded under the supposition of α. Hence, imaging on the minimally informative meaning of $\alpha > \gamma$ 'probabilistically excludes' w_2 and the probability share of w_2 will be fully transferred to w_1.

There comes a problem of underdetermination when learning the minimally informative proposition $[\alpha > \gamma]$. In general, the most similar $\alpha > \gamma$-world from any excluded $\neg(\alpha > \gamma)$-world is underdetermined, and thus it is not determined whereto the probability shares of the excluded worlds are transferred. In the Stalnaker model depicted in Figure 2.3, for instance, there are three candidate worlds for the most similar $\alpha > \gamma$-world from the excluded w_2. We resort to the default assumption (introduced in Section 2.2.4) to solve the problem of underdetermination.

The default assumption states that the most similar $\alpha > \gamma$-world from any excluded $\alpha > \neg\gamma$-world is a $\alpha \wedge \gamma$-world, if there is more than one candidate. Observe that all worlds included in the minimally informative proposition $[\alpha > \gamma]$ are themselves, respectively, their unique most similar $[\alpha > \gamma]$-world due to the reflexivity of the acquired similarity order. In this way the minimally informative interpretation minimises the number of excluded worlds, for which the default assumption may be needed to overcome the problem of underdetermination. In case only α and γ are relevant and all an agent learns is $[\alpha > \gamma]_{min}$, the default assumption simply states that the $\alpha \wedge \gamma$-world is the most similar world to the excluded $\alpha \wedge \neg\gamma$-world. If the agent obtains additional information over and above $[\alpha > \gamma]_{min}$, the default assumption leads to different outcomes – depending on which additional information is learned. The outcome's non-rigidity or variability with respect to different contextual information is illustrated by the Ski Trip Example and Driving Test Example studied in Section 2.2.6.

There is a link between the default assumption and the probability distribution after learning: in the case where the only received and relevant information is $[\alpha > \gamma]_{min}$, the default assumption is satisfied iff the probability of the antecedent remains unchanged. Assume the default assumption is in place and an agent does not come to believe a more informative proposition than $[\alpha > \gamma]_{min}$. Then our learning method prescribes that the probability share of the $\alpha \wedge \neg\gamma$-world is transferred to the $\alpha \wedge \gamma$-world. Hence, $P^{\alpha>\gamma}(\alpha) = P(\alpha)$. For the converse, assume $P^{\alpha>\gamma}(\alpha) = P(\alpha)$ and that all an agent learns is $[\alpha > \gamma]_{min}$. The only transferred probability share is again $P(\alpha \wedge \neg\gamma)$. Suppose for reductio that $P(\alpha \wedge \neg\gamma)$ is transferred to some $\neg\alpha$-world. Since $P(\alpha \wedge \neg\gamma) > 0$, $P^{\alpha>\gamma}(\neg\alpha)$ would be greater than $P(\neg\alpha)$. By the probability calculus, it would follow that $P^{\alpha>\gamma}(\alpha) < P(\alpha)$, which contradicts the assumption. Hence, $P(\alpha \wedge \neg\gamma)$ is transferred to the $\alpha \wedge \gamma$-world, and thus the default assumption is satisfied.

We argue now that the default assumption is warranted by the rationale of maximally conservative belief change. A consequence of this rationale is that there should be no belief change without any reason. (If there were such a belief change, the change would be stronger than necessary, and thus violating the rationale.) Upon learning the indicative conditional "If α, then γ", there seems to be no reason to change the probability of the antecedent, at least in the absence of additional information.[13] As we have seen in the previous paragraph, if the agent does not possess or receive additional information, the default assumption ensures that the probability of the antecedent remains unchanged when learning "If α, then γ" interpreted as $[\alpha > \gamma]_{min}$. In the absence of further information, the default assumption thus implements a demand of maximal conservativeness, viz. that the probability of the antecedent should remain unchanged.

Let us consider cases where further contextual information is given. Here, the default assumption does not necessarily ensure that the probability of the antecedent remains unchanged. If the contextual information is sufficient to uniquely determine the respective most similar world, we do not need to rely on the default assumption. We will see below that additional (contextual) information may sometimes fully determine the epistemic states under consideration such that we are not in need of the default assumption. If we need to rely on the default assumption, in contrast, we should judge

[13] Notice that the assumption of no additional information literally excludes that there is an epistemic reason, i. e. some belief apart from $[\alpha > \gamma]_{min}$, to change the probability of the antecedent.

the assumption by its predictions for specific scenarios. Unfortunately, there is a myriad of possible scenarios that differ in their contextual information. However, we may refer again to the case studies in Section 2.2.6, in which our learning method generates the intuitively correct results. Hence, the default assumption seems to be justified in absence and presence of further information, at least *prima facie*. This being said, we encourage the search for counterexamples to our learning method, especially counterexamples involving contextual information.

In line with our learning method, Douven and Romeijn (2011) suggest that the probability of the antecedent does not change upon learning an indicative conditional, at least in the absence of further relevant information.[14] They write:

> We are inclined to think that Adams conditioning, or, equivalently, Jeffrey conditioning with the explicit constraint of keeping the antecedent's probability fixed in the update [...] covers most of the cases of learning a conditional. (p. 654)

Since Adams conditioning always keeps the probability of the antecedent fixed, it is no general method for learning conditional information, as sometimes this probability should change. It "may entirely fall upon us", so Douven and Romeijn (2011, p. 660), "to decide, on the basis of contextual information, whether or not [Adams conditioning] applies to the learning of a given conditional"; and further "deciding when Adams conditioning applies, may be an art, or a skill, rather than a matter of calculation or derivation from more fundamental epistemic principles."

In contrast to Douven and Romeijn (2011), our learning method proposes a principled way how learning conditional information should affect the probability of the antecedent. As a consequence of the minimally informative interpretation and the default assumption, learning indicative conditional information does not alter the probability of the antecedent in the absence of further information. If, in addition, contextual information is learned, the default that the probability of the antecedent remains unchanged may be violated. Given the respective contextual information, our learning method automatically calculates a possibly changed probability for the antecedent. No skilful decision about which method should be applied is required.

2.2.6 Douven's Examples and the Judy Benjamin Problem

We apply now our method of learning conditional information to Douven's examples and the Judy Benjamin Problem.

A Possible Worlds Model for the Sundowners Example

Example 4. The Sundowners Example (Douven and Romeijn (2011, pp. 645–646))
Sarah and her sister Marian have arranged to go for sundowners at the Westcliff hotel tomorrow. Sarah feels there is some chance that it will rain, but thinks they can always enjoy the view from inside. To make sure, Marian consults the staff at the Westcliff hotel and finds out that in the event of rain, the inside area will be occupied by a wedding party. So she tells Sarah:

$$\text{If it rains tomorrow, we cannot have sundowners at the Westcliff.} \tag{2.7}$$

Upon learning this conditional, Sarah sets her probability for sundowners and rain to 0, but she does not adapt her probability for rain.

[14] Douven (2012) argues more precisely that the probability of the antecedent should only change if the antecedent is explanatorily relevant for the consequent. It is noteworthy that if the probability of the antecedent should intuitively change in one of Douven's examples, the explanatory relations always involve beliefs in additional propositions (apart from the conditional) given by the example's context description.

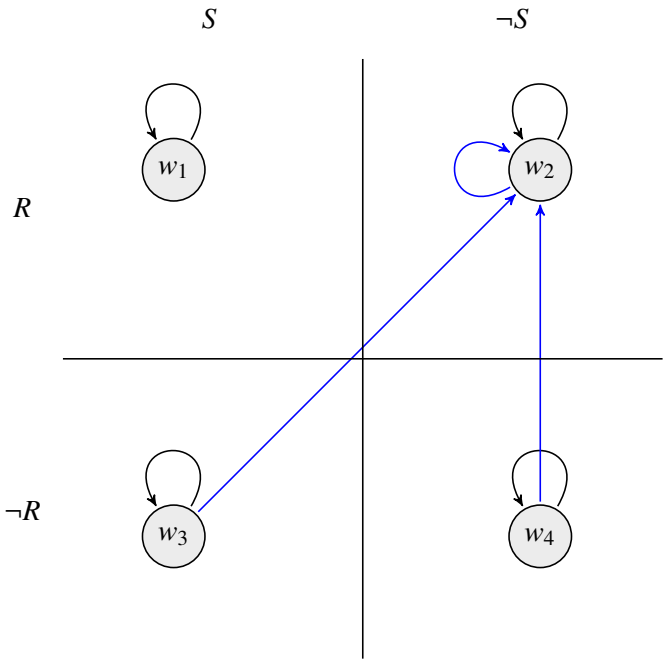

Figure 2.4: A Stalnaker model for Sarah's belief state in the Sundowners Example.

We model Sarah's belief state as the Stalnaker model $\mathcal{M}_{St} = \langle W, R, \leq, V \rangle$ depicted in Figure 2.4. W contains four elements covering the possible events of $R, \neg R, S, \neg S$, where R stands for "it rains tomorrow" and S for "Sarah and Marian can have sundowners at the Westcliff tomorrow".

Sarah interprets the conditional uttered by her sister Marian as saying that the most similar R-world from the respective 'actual' world is a world that satisfies $\neg S$. Note the symmetry to the scheme depicted in Figure 2.3. Critics may find reasons why the default assumption in (ii).(b) of Section 2.2.4 is unjustified, and thus that we encounter here the problem of underdetermination. However, as Douven himself points out, the intuition in the Sundowners Example derives from the verdict that whether or not it rains may affect whether or not they have sundowners, but not the other way around: having sundowners simply has no effect whatsoever on whether or not it rains.[15] Hence, the change of belief between R and $\neg R$ is more far fetched than between S and $\neg S$. In other words, the worlds along the horizontal axis are more similar than the worlds along the vertical axis. Consequently, $min_{\leq_{w_1}}[R > \neg S] = w_2$.

Imaging on the minimally informative proposition $[R > \neg S] = \{w_2, w_3, w_4\}$ results in $P^{R>\neg S}(w') = \sum_w P(w) \cdot \left\{ \begin{array}{ll} 1 & \text{if } w_{R>\neg S} = w' \\ 0 & \text{otherwise} \end{array} \right\}$:

$$\begin{aligned} P^{R>\neg S}(w_1) &= 0 & P^{R>\neg S}(w_2) &= P(w_1) + P(w_2) \\ P^{R>\neg S}(w_3) &= P(w_3) & P^{R>\neg S}(w_4) &= P(w_4) \end{aligned} \tag{2.8}$$

We see immediately that both intuitions associated with the Sundowners Example are satisfied, viz. $P^{R>\neg S}(R) = P(R) = P(w_1) + P(w_2)$ and $P^{R>\neg S}(R \wedge S) = P^{R>\neg S}(w_1) = 0$. We conclude that our method yields the intuitively correct results.

[15] Cf. Douven (2012, p. 8).

Although justified by Douven's remarks, the Sundowners Example demands to invoke the default assumption in order to avoid the problem of underdetermination. We will see in the following examples that more contextual information, in particular additional factual information, may render the default assumption superfluous.[16]

A Possible Worlds Model for the Ski Trip Example

Example 5. The Ski Trip Example (Douven and Dietz (2011, p. 33))
Harry sees his friend Sue buying a skiing outfit. This surprises him a bit, because he did not know of any plans of hers to go on a skiing trip. He knows that she recently had an important exam and thinks it unlikely that she passed. Then he meets Tom, his best friend and also a friend of Sue's, who is just on his way to Sue to hear whether she passed the exam, and who tells him:

If Sue passed the exam, her father will take her on a skiing vacation. (2.9)

Recalling his earlier observation, Harry now comes to find it more likely that Sue passed the exam.

We model Harry's belief state as the Stalnaker model $\mathcal{M}_{St} = \langle W, R, \leq, V\rangle$ depicted in Figure 2.5. W contains eight elements covering the possible events of $E, \neg E, S, \neg S, B, \neg B$, where E stands for "Sue passed the exam", S for "Sue's father takes her on a skiing vacation", and B for "Sue buys a skiing outfit".

Harry interprets the conditional uttered by his friend Tom as saying that the most similar E-world from the actual world is a world that satisfies S. Crucially, Harry observed Sue buying a skiing outfit, and thus has the factual information that B.

In total, Harry learns the minimally informative proposition $[(E > S) \wedge B] = \{w \in W \mid min_{\leq_w}[E] \in [S] \wedge B\} = \{w_1, w_3, w_4\}$. Moreover, the default assumption states that $w_1(E \wedge S \wedge B) = 1$ is the most similar world to all worlds not included in the minimally informative proposition $[(E > S) \wedge B]$ (see the caption of Figure 2.5).

Imaging on the minimally informative proposition $[(E > S) \wedge B] = \{w_1, w_3, w_4\}$ results in $P^{(E>S)\wedge B}(w') = P^*(w') = \sum_w P(w) \cdot \left\{ \begin{array}{ll} 1 & \text{if } w_{(E>S)\wedge B} = w' \\ 0 & \text{otherwise} \end{array} \right\}$:

$$\begin{array}{ll} P^*(w_1) = P(w_1) + P(w_2) + P(w_5) + P(w_6) + P(w_7) + P(w_8) & P^*(w_2) = 0 \\ P^*(w_3) = P(w_3) & P^*(w_4) = P(w_4) \\ P^*(w_5) = 0 & P^*(w_6) = 0 \\ P^*(w_7) = 0 & P^*(w_8) = 0 \end{array} \tag{2.10}$$

Our method yields again the correct result regarding the intuition associated with the Ski Trip Example: $P^*(E) \geq P(E)$, since $P^*(E) = P^*(w_1)$ and $P(E) = P(w_1) + P(w_2) + P(w_5) + P(w_6)$, and $P^*(E) > P(E)$ if $P(w_7) + P(w_8) > 0$.

The just presented model of the Ski Trip Example is surprisingly simple. Critics may say it is too simple: your model should not omit the relation between B and S. Indeed, the first two sentences of the Ski Trip Example suggest that there is some relation between Sue's buying a skiing outfit and her going on a skiing trip (or vacation). Intuitively, (a) going on a skiing trip is a good explanation for buying a skiing outfit ($S > B$), and (b) buying a skiing outfit is a good predictor of going on a skiing

[16]Note that the Sundowners Example seems to be somewhat artificial. It seems plausible that upon hearing her sister's conditional, Sarah would promptly ask "why?" in order to obtain some more contextual information, before setting her probability for sundowners and rain to 0. After all, she "thinks that they can always enjoy the view from inside".

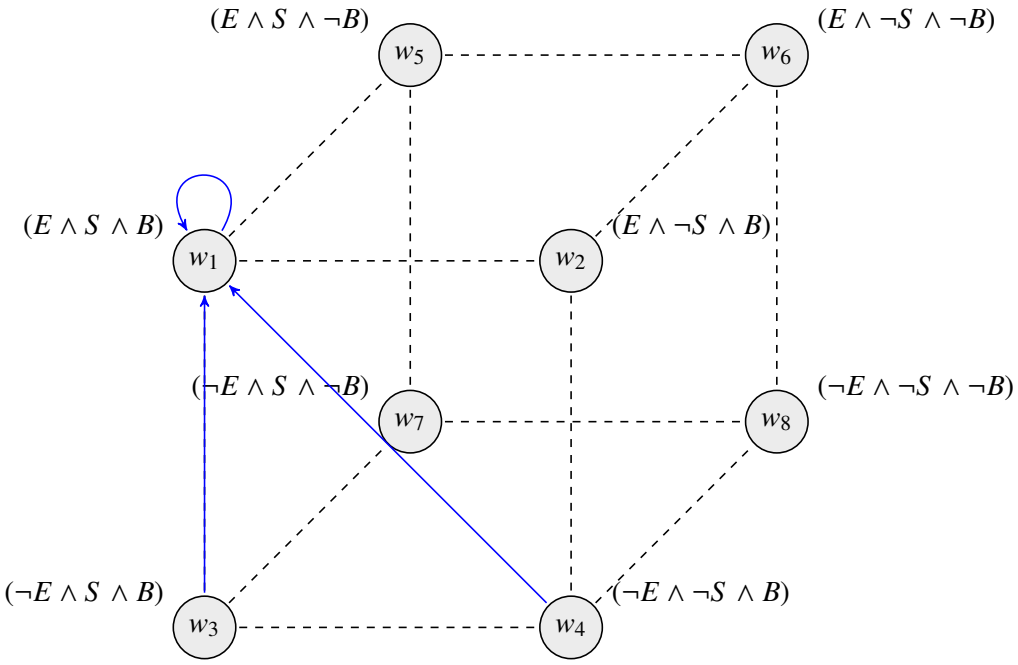

Figure 2.5: An eight-worlds Stalnaker model for Harry's belief state in the Ski Trip Example. The blue arrows illustrate the change of the similarity order such that the received and interpreted information $[(E > S) \wedge B)]$ is minimally informative. By the default assumption, $w_1(E \wedge S \wedge B) = 1$ is the most similar world from any $\neg((E > S) \wedge B)$-world in case of underdetermination. Note that receiving the factual information that B excludes all the worlds on the back of the cube. We see that receiving factual information is very informative, as compared to obtaining conditional information.

trip $(B > S)$. Correspondingly, the critics may say that Harry rather learns $[(E > S) \wedge (S > B) \wedge B]$ or $[(E > S) \wedge (B > S) \wedge B]$.

We will show now that our learning method's result for the Ski Trip Example is robust with respect to (a) and (b), i. e. it is compatible with assuming that Harry assumes, knows, or learns (a) or (b) in addition to $[(E > S) \wedge B]$. (Keep in mind, however, that we have no need to explicitly assume or model any relation between B and S in order to obtain the intuitively correct result.)

Let's assume Harry knows (a) $S > B$. In total, he comes to know the minimally informative proposition $[(E > S) \wedge (S > B) \wedge B] = \{w \in W \mid min_{\leq_w}[E] \in [S] \wedge min_{\leq_w}[S] \in [B] \wedge B\} = \{w_1, w_3, w_4\}$. Since the minimally informative proposition $[(E > S) \wedge (S > B) \wedge B]$ is identical to the minimally informative proposition $[(E > S) \wedge B]$, our method yields the same result. Hartmann and Rad (2017) make and need an assumption similar to (a). "From the story it is clear", they write, that it "is more likely that Sue buys a new ski outfit if her father invites her for a ski trip than if he does not" (p. 11). They represent the relation between B and S as a directed arrow from S to B in a Bayesian network claiming this would "properly represent the causal relation between these variables" (ibid.).

In contrast to the 'causal relation' between the variables, Harry could engage in the predictive inference from Sue buying a skiing outfit to raising the likelihood that she will go on a skiing trip. So, let's assume Harry thinks (b) $B > S$. In total, he comes to know the minimally informative proposition $[(E > S) \wedge (B > S) \wedge B] = \{w_1, w_3\}$. Since the minimally informative proposition $[(E > S) \wedge (B > S) \wedge B]$ is a subset of the minimally informative proposition $[(E > S) \wedge B]$, our method yields an even stronger result: if B is a predictor of S and B is believed, then it is (possibly) even more likely that Sue passed the exam (E). We obtain $P^{(E>S)\wedge(B>S)\wedge B}(E) \geq P^{(E>S)\wedge B}(E) \geq P(E)$. (Notice that our method is also apt to handle cases when Harry assumes, knows, or learns the conditionals (a) or (b) to a certain degree, as we will illustrate in the discussion of the Judy Benjamin Problem.)

Of course, Harry could also be equipped with some other contextual knowledge. Douven (2012, p. 11) himself, for example, provides an alternative picture:

> Sue's having passed the exam would, if true, *explain* why she bought the skiing outfit; that makes her having passed the exam more credible.

Here, Douven seems to propose that Harry has a relation between B and E in mind, viz. $E > B$ (given he knows already $E > S$ and B). On this picture, Harry learns in total the minimally informative proposition $[(E > S) \wedge (E > B) \wedge B] = \{w_1, w_3, w_4\}$. We see that if S and B are related in the sense that they are both a 'consequent' of E, our method yields again the correct result.

We take this as further evidence that our method's result for the Ski Trip Example is largely independent of and compatible with additional plausible assumptions between its variables. If the respective additional assumptions correspond to different (admissible) interpretations, the robustness of the result could explain why it is intuitively so clear that Harry should believe it more likely that Sue passed the exam.[17] In any case, our learning method needs fewer assumptions than the other accounts to obtain the desired result for the Ski Trip Example. At the same time, the result still stands if we adopt the additional assumptions on which Douven (2012) as well as Hartmann and Rad (2017) rely.

A Possible Worlds Model for the Driving Test Example

Example 6. The Driving Test Example (Douven (2012, p. 3))
Betty knows that Kevin, the son of her neighbors, was to take his driving test yesterday. She has no

[17] In Section 2.8, we generalise the proposed method to the learning of causal information, which allows us to define an inference to the best explanation scheme, as Douven envisioned for the Ski Trip Example.

idea whether or not Kevin is a good driver; she deems it about as likely as not that Kevin passed the test. Betty notices that her neighbors have started to spade their garden. Then her mother, who is friends with Kevin's parents, calls her and tells her the following:

If Kevin passed the driving test, his parents will throw a garden party. (2.11)

Betty figures that, given the spading that has just begun, it is doubtful (even if not wholly excluded) that a party can be held in the garden of Kevin's parents in the near future. As a result, Betty lowers her degree of belief for Kevin's having passed the driving test.

We model Betty's belief state as the Stalnaker model $\mathcal{M}_{St} = \langle W, R, \leq, V\rangle$ depicted in Figure 2.6. W contains eight elements covering the possible events of $D, \neg D, G, \neg G, S, \neg S$, where D stands for "Kevin passed the driving test", G for "Kevin's parents will throw a garden party", and S for "Kevin's parents have started to spade their garden".

Betty interprets the conditional uttered by her mother as saying that the most similar D-world from the actual world is a world that satisfies G. Furthermore, Betty infers from her contextual knowledge that if Kevin's parents are spading their garden, then they will not throw a garden party, in symbols $S > \neg G$. Therefore, Betty also obtains the information that the most similar S-world from the actual world is a world that satisfies $\neg G$. Finally, Betty knows that Kevin's parents have started to spade their garden, and thus has the factual information that S.

In total, Betty learns the minimally informative proposition $[(D > G) \wedge (S > \neg G) \wedge S] = \{w \in W \mid min_{\leq_w}[D] \in [G] \wedge min_{\leq_w}[S] \in [\neg G] \wedge S\} = \{w_4\}$. The obtained information, although interpreted in a minimally informative way, is sufficient to identify the actual world. Therefore, the Stalnaker model provides us with a unique most similar world (to any other world) under the changed similarity order (see the caption of Figure 2.6).

Intuitively, Betty learns that she is in a S-world, since she factually obtains the information that S. Hence, the conditional $S > \neg G$ implies that $\neg G$ is true in the actual world. By the conditional $D > G$, we know that G is satisfied in the most similar D-world from the actual world. Since $\neg G$ is true in the actual world, we know that the actual world is not a D-world. But then the actual world is a $\neg D$-world. For, if the actual world w were a D-world, w would satisfy G. To summarise, the actual world satisfies $\neg D, \neg G$, and, obviously, S.

Imaging on the minimally informative proposition $[(D > G) \wedge (S > \neg G) \wedge S] = \{w_4\}$ results in $P^{(D>G)\wedge(S>\neg G)\wedge S}(w') = P^*(w') = \sum_w P(w) \cdot \left\{ \begin{array}{ll} 1 & \text{if } w_{(D>G)\wedge(S>\neg G)\wedge S} = w' \\ 0 & \text{otherwise} \end{array} \right\}$:

$$\begin{array}{ll} P^*(w_1) = 0 & P^*(w_2) = 0 \\ P^*(w_3) = 0 & P^*(w_4) = 1 \\ P^*(w_5) = 0 & P^*(w_6) = 0 \\ P^*(w_7) = 0 & P^*(w_8) = 0 \end{array} \tag{2.12}$$

Our method yields again the correct result regarding the intuition associated with the Driving Test Example: $P^*(D) < P(D)$, since $P^*(D) = 0$ and $P(D) = P(w_1) + P(w_2) + P(w_5) + P(w_6) > 0$. The following Judy Benjamin Problem will show that if Betty thinks that the conditional $D > G$ or $S > \neg G$ (or both) is/are uncertain, then the probability shares for some other worlds will not reduce to zero. This fact fits nicely with the Driving Test Examples's remark that "given the spading that has just begun, it is doubtful [or uncertain] (even if not wholly excluded) that a party can be held in the garden of Kevin's parents". We will treat the learning of uncertain conditional information in the next section.

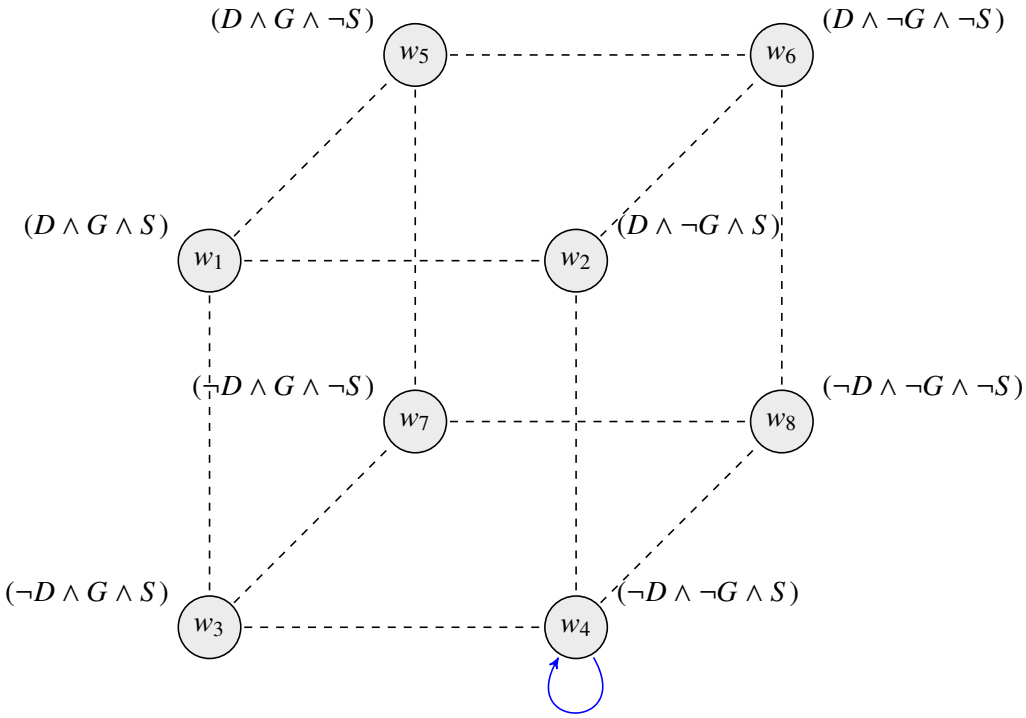

Figure 2.6: An eight-worlds Stalnaker model for Betty's belief state in the Driving Test Example. There is only a single world that satisfies the 'minimally informative' proposition $[(D > G) \wedge (S > \neg G) \wedge S]$. For, $[(D > G) \wedge (S > \neg G)] = \{w \in W \mid min_{\leq_w}[D] \in [G] \wedge min_{\leq_w}[S] \in [\neg G]\} = \{w_4, w_5, w_7, w_8\}$. Of those worlds only w_4 is in the S-plane of the cube and thus the actual world.

A Possible Worlds Model for the Judy Benjamin Problem

We apply now our method of learning conditional information to a case, in which the received conditional information is uncertain. We show thereby that the method may be generalised to cases in which the learned conditional information is uncertain, provided we use Jeffrey imaging. Following the presentation in Hartmann and Rad (2017), we consider Bas Van Fraassen's Judy Benjamin Problem.[18]

Example 7. The Judy Benjamin Problem (Hartmann and Rad (2017, p. 7))
A soldier, Judy Benjamin, is dropped with her platoon in a territory that is divided in two halves, Red territory and Blue territory, respectively, with each territory in turn being divided in equal parts, Second Company area and Headquarters Company area, thus forming four quadrants of roughly equal size. Because the platoon was dropped more or less at the center of the whole territory, Judy Benjamin deems it equally likely that they are in one quadrant as that they are in any of the others. They then receive the following radio message:

$$\begin{aligned}&\text{I can't be sure where you are. If you are in Red Territory,}\\&\text{then the odds are 3 : 1 that you are in Second Company area.}\end{aligned} \tag{2.13}$$

After this, the radio contact breaks down. Supposing that Judy accepts this message, how should she adjust her degrees of belief?

Douven claims that the probability of R should, intuitively, remain unchanged after learning the uncertain conditional information. Furthermore, the probability distribution after hearing the radio message, i. e. P^*, should take the following values:

$$\begin{aligned}P^*(R \wedge S) &= \frac{3}{8} \quad P^*(R \wedge \neg S) = \frac{1}{8}\\ P^*(\neg R \wedge S) &= \frac{1}{4} \quad P^*(\neg R \wedge \neg S) = \frac{1}{4}\end{aligned} \tag{2.14}$$

We model Judy Benjamin's belief state as the Stalnaker model $\mathcal{M}_{St} = \langle W, R, \leq, \leq', V\rangle$ depicted in Figure 2.7. W contains four elements covering the possible events of $R, \neg R, S, \neg S$, where R stands for "Judy Benjamin's platoon is in Red territory", and S for "Judy Benjamin's platoon is in Second Company area". The story prescribes that the probability distribution before learning the uncertain information is given by:

$$P(R \wedge S) = P(R \wedge \neg S) = P(\neg R \wedge S) = P(\neg R \wedge \neg S) = \frac{1}{4} \tag{2.15}$$

In the previous examples, our agents learned a Stalnaker conditional $\alpha > \gamma$ with certainty. According to Theorem 1, this amounts to the constraint that $P(\alpha > \gamma) = P^{\alpha}(\gamma) = 1$ (provided α is not a contradiction). Given this constraint and since P^{α} is a probability distribution, we have $P^{\alpha}(\neg\gamma) = 1 - P^{\alpha}(\gamma) = 0$. This means that we were able to probabilistically exclude any $\neg\gamma$-world under the supposition of α.

Now, our agent Judy Benjamin learns a Stalnaker conditional with uncertainty. According to Theorem 1, this amounts in this case to the constraint that $P(R > S) = P^{R}(S) = \frac{3}{4}$. By the law of total probability, we obtain $P(R > \neg S) = P^{R}(\neg S) = \frac{1}{4}$. In contrast to learning conditional information with certainty, we cannot subtract the whole probabilistic mass from the $\neg S$-worlds under the supposition

[18] Cf. Van Fraassen (1981, pp. 376–379) and Van Fraassen et al. (1986).

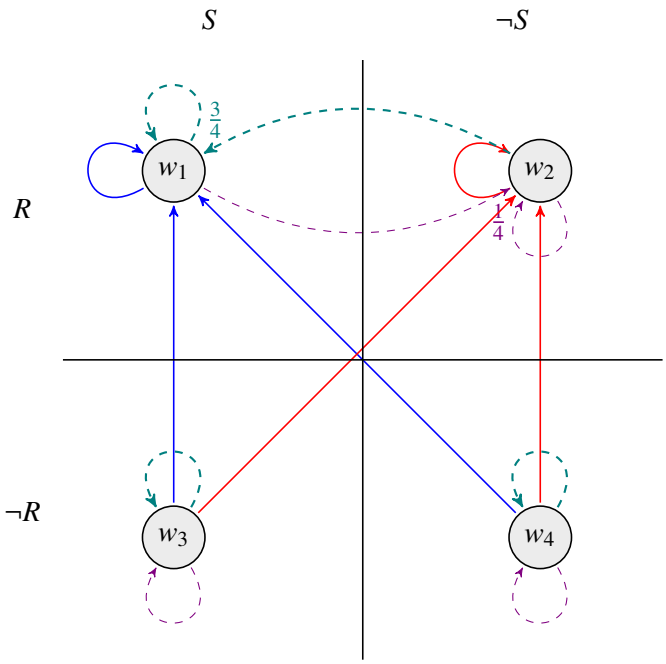

Figure 2.7: A Stalnaker model for Private Benjamin's belief state in the Judy Benjamin Problem. The blue arrows illustrate the specification of a similarity order $\leq'$ such that the received information $[R > S]$ is minimally informative. Note that each world having a blue arrow satisfies $R > S$. The red arrows illustrate the specification of another similarity order $\leq \neq \leq'$ such that the received information $[R > \neg S]$ is minimally informative. Each world having a red arrow satisfies $R > \neg S$. In sum, the similarity orders are specified such that one makes $[R > S]$ a minimally informative proposition and the other makes $[R > \neg S]$ a minimally informative proposition. By the default assumption, we obtain $min_{\leq_{w(R>\neg S)=1}}[R > S] = w_1$ and $min_{\leq'_{w(R>S)=1}}[R > \neg S] = w_2$. Note that under $\leq$, worlds w_3 and w_4 satisfy $R > S$, under $\leq'$ they satisfy $R > \neg S$. The teal arrows represent the transfer of $k \cdot P(w)$, while the violet arrows represent the transfer of $1 - k \cdot P(w)$. The application of Jeffrey imaging on $[R > S]$ with $k = \frac{3}{4}$ leads to the following probability distribution: $P^{R>S}_{\frac{3}{4}}(w_3) = \frac{3}{4} \cdot P(w_3) + \frac{1}{4} \cdot P(w_3)$ and $P^{R>S}_{\frac{3}{4}}(w_4) = \frac{3}{4} \cdot P(w_4) + \frac{1}{4} \cdot P(w_4)$, whereas $P^{R>S}_{\frac{3}{4}}(w_1) = \frac{3}{4} \cdot P(w_1) + \frac{3}{4} \cdot P(w_2)$ and $P^{R>S}_{\frac{3}{4}}(w_2) = \frac{1}{4} \cdot P(w_1) + \frac{1}{4} \cdot P(w_2)$.

of R. However, Judy Benjamin is informed from an external source about the proportion to which she should gradually 'exclude' or downweigh the probability share of $\neg(S > R)$-worlds. Equivalently, we may say that the most similar $S > R$-world (from any $\neg(S > R)$-world) obtains a gradual upweight of probability such that it receives $\frac{3}{4}$ of the probability shares of the $\neg(R > S)$-worlds; in turn, however, this $\neg(S > R)$-world then receives a probability share from its most similar $R > S$-world weighed by $\frac{1}{4}$. Note that in Stalnaker models $\neg(S > R) \equiv S > \neg R$ given a similarity order (and provided S is possible). Judy Bejamin's learning process may thus be modeled by considering the degree of belief in two Stalnaker conditionals that are, under a single similarity order, by the principle of Conditional Excluded Middle, mutually exclusive.

We apply now Jeffrey imaging to the Judy Benjamin Problem, where a source external to Judy provides her with the information that $k = \frac{3}{4}$.

$$P^{R>S}_{\frac{3}{4}}(w') = \sum_{w}(P(w) \cdot \left\{ \begin{array}{ll} k & \text{if } w_{R>S} = w' \\ 0 & \text{otherwise} \end{array} \right\} + P(w) \cdot \left\{ \begin{array}{ll} (1-k) & \text{if } w_{R>\neg S} = w' \\ 0 & \text{otherwise} \end{array} \right\}) \tag{2.16}$$

Given the probability distribution before the learning process in Equation (2.15), Judy obtains the following probability distribution after being informed that $P(R > S) = \frac{3}{4}$:

$$\begin{array}{ll} P^{R>S}_{\frac{3}{4}}(w_1) = P^{R>S}_{\frac{3}{4}}(R \wedge S) = \frac{3}{8} & P^{R>S}_{\frac{3}{4}}(w_2) = P^{R>S}_{\frac{3}{4}}(R \wedge \neg S) = \frac{1}{8} \\ P^{R>S}_{\frac{3}{4}}(w_3) = P^{R>S}_{\frac{3}{4}}(\neg R \wedge S) = \frac{1}{4} & P^{R>S}_{\frac{3}{4}}(w_4) = P^{R>S}_{\frac{3}{4}}(\neg R \wedge \neg S) = \frac{1}{4} \end{array} \tag{2.17}$$

Our learning method using Jeffrey imaging matches the intuitions Douven claims to be correct for the Judy Benjamin Problem. Note, in particular, that $P^{R>S}_{\frac{3}{4}}(R) = P(R) = \frac{1}{2}$.[19]

2.3 Taking Stock

Let us take stock. We have seen that Douven's dismissal of the Stalnaker conditional as a tool to model the learning of conditionals is unjustified. Rather, the learning may be modelled by Jeffrey imaging on the meaning of Stalnaker conditionals under the following two conditions: (i) the similarity order of the Stalnaker model is changed in a way such that the meaning of the conditional is minimally informative, and (ii) the default assumption is in place. The proposed learning method leads to the intuitively correct results in Douven's examples and complies with Douven's intuition for the Judy Benjamin Problem.

The minimally informative meaning of a Stalnaker conditional corresponds to the meaning of the material implication. So why – for the sake of simplicity – do we not propose a method of learning conditional information by Jeffrey imaging on the material implication? There are two reasons. First, the application of Jeffrey imaging is defined with respect to similarity orders independent of using the material implication or the Stalnaker conditional. An equivalent of the default assumption for the material implication would require (an equivalent to) a similarity order as well. Hence, the proposal to Jeffrey image on the material implication is *prima facie* not more simple than our proposal.

Second, it is far from clear how to formulate Jeffrey imaging and the default assumption for the material implication with respect to uncertainty, contextual information and nested conditionals, as we did for the Stalnaker conditional. The material implication and the minimally informative interpreted

[19] Appendix A contains a possible worlds model of Douven and Romeijn's (2011) Jeweller Example. There, we show that our method also applies to examples where uncertain factual information is learned.

Stalnaker conditional come apart when negated. Whereas the negation of the material implication carries the strong information $[\alpha \wedge \neg\gamma]$, the negation of the minimally informative proposition expressed by a Stalnaker conditional, i. e. $[\neg(\alpha > \gamma)]_{min}$, is again a minimally informative proposition, viz. $[\alpha > \neg\gamma]_{min} = [\neg\alpha \vee \neg\gamma]$. As should be clear by now, this difference is crucial for the learning of uncertain conditional information by Jeffrey imaging and shows that the Stalnaker conditional is better suited than the material implication.

Even if we try to repair the material implication account by simply stipulating that $\neg(\alpha \rightarrow \gamma)$ means $\alpha \rightarrow \neg\gamma$ (which it doesn't!), we run into further problems when considering nested conditionals. Our learning method validates the import-export principle for right-nested conditionals, i. e. $[\alpha > (\beta > \gamma)]_{min} = [(\alpha \wedge \beta) > \gamma]_{min}$, and minimally informative interpreted right-nested Stalnaker conditionals correspond to their material counterparts given their usual meaning, i. e. $[\alpha > (\beta > \gamma)]_{min} = [\alpha \rightarrow (\beta \rightarrow \gamma)]$. However, for left-nested conditionals, we obtain $[(\alpha \rightarrow \beta) \rightarrow \gamma] \subset [(\alpha > \beta) > \gamma]_{min} \subset [(\alpha \wedge \beta) > \gamma)]_{min}$, a divergence between the material implication and the Stalnaker conditional that is not easily overcome.

As we have just defended against the material implication, the proposed learning method is based on Stalnaker's semantics for conditionals. Crucially for our learning method, this semantics validates the principle of Conditional Excluded Middle as opposed to, for example, Lewis's (1973a) semantics for counterfactuals.[20] We are not claiming that Stalnaker's semantics are the correct semantics for conditionals (if there is any). Rather the employed semantics seem to be a good approximation. However, we are committed to the following conditional claim: if Stalnaker's semantics for conditionals are considered to be correct, or at least rational (for many cases), then the proposed learning method should be considered correct, or at least rational (for those cases). Next, we touch upon the question whether Stalnaker's semantics apply to subjunctive and counterfactual conditionals, and thus whether our learning method is applicable to subjunctives and counterfactuals.

2.4 Subjunctive Conditional Information

So far, we have applied our method of learning conditional information only to examples involving indicative conditionals. Can our method of learning conditional information also be applied to the learning of subjunctive conditionals? On the face of it, there seems to be no particular problem. From the subjunctive conditional "If Oswald had not killed Kennedy, someone else would have" you learn according to our method that the most similar world in which Oswald did not kill Kennedy, is a world in which someone else did. However, as we have discussed in the Introduction, the subjunctive conditional differs from its corresponding indicative conditional "If Oswald did not kill Kennedy, someone else did" in meaning. Hence, our learning method should be able to learn different propositions depending on the mood of a conditional.

We have already discussed the Oswald-Kennedy example in the Introduction (Chapter 1). In light of the Kennedy-Oswald pair of conditionals, we need to explain how the information to be learned differs depending on the subjunctive and indicative mood. As our method relies on the Stalnaker conditional, we need to say how the most similar antecedent world differs when changing the mood of a conditional. The rough idea to evaluate a Stalnaker conditional is to move to the possible world in which the antecedent obtains and is as much like the actual world as possible both with regard to particular facts about the past as well as generalisations about what might follow from what. In the

[20] See Stalnaker (1981) for a defense of the principle of Conditional Excluded Middle based on superevaluations and cognitive habits.

Introduction, we promised to sketch an amendment of Stalnaker's semantics that allows for the most similar antecedent world to differ depending on the mood of the conditional. Here it comes.

To evaluate a conditional we move to possibly different most similar worlds. In the case of indicative conditionals, the world we move to is just the most similar antecedent world to the actual world. In the case of subjunctive conditionals, the world we move to is the most similar antecedent world to the actual world as it has been immediately before the point of time to which the antecedent refers. The idea is borrowed from Lewis (1973c, p. 566) who claims that "counterfactuals typically keep fixed the past up until the time at which the counterfactual antecedent is supposed to obtain." Hence, when you move to a most similar antecedent world in the subjunctive case, you are not restricted by the facts of the actual world in-between the reference time of the antecedent and the now. A full account of antecedent reference will not be given here.[21] However, we would like to answer two obvious challenges.

First, consider one of Ramsey's 'universal' variable hypotheticals: "If you *ever* took arsenic, you would come to be poisoned." Here it seems that the antecedent's time reference may be represented by a variable bound by a universal quantifier on the whole conditional. Given *any* (also future) time t, if you were to take arsenic, you would be poisoned after t. Hence, to evaluate a variable hypothetical we are free to move to any most similar antecedent world at any time t. In particular, we may just move to the most similar antecedent world here and now: "If you take arsenic, you will be poisoned."

Second, relatedly, there are antecedents that seemingly do not refer to any time. Take, for example, Lewis's humorous conditional "If kangaroos had no tails, they would topple over." Interestingly, if kangaroos have no tails, they topple over as well. For timeless antecedents, the most similar antecedent world for indicatives and subjunctives seems to coincide. This fits nicely to Lewis's remark that subjunctives about the future seem equivalent to their corresponding indicatives. "If she were to eat the cake in 10 minutes from now, she would come to have a stomach ache" indeed seems to be equivalent to "If she eats the cake in 10 minutes from now, she will have a stomach ache."

As is widely acknowledged, let us suppose that our beliefs about the future are based inductively on our beliefs of the past. Then, if the antecedent world refers to the present or future, we are *at present* unable to distinguish the antecedent world most similar to the actual world from the most similar world in 10 minutes. The antecedent world we judge most similar now, or before any time in the future, is just the very antecedent world we judge most similar up to the present.

Metaphorically speaking, indicative conditionals stick closer to the actual world than subjunctive conditionals because indicatives involve no time travel in thoughts. The most similar antecedent world of an indicative can be no further from the actual world than the corresponding subjunctive (at least, as we always tacitly assumed, that time is more fundamental than anything else, including causation.) Except in cases where they coincide, indicative antecedent worlds are more similar to and thus more constrained by the (facts of the) actual world (as it is now). This explains our observation in the Introduction that the evaluation of subjunctives is somewhat more independent from the actual here and now than the evaluation of indicatives.

Let us revisit the Oswald-Kennedy example. Both conditionals are composed out of the following atoms:

O: Oswald killed Kennedy.

S: Someone else killed Kennedy.

[21] Bennett (1974) and Davis (1979) work out full accounts of antecedent reference, and so does Khoo (2017) more recently.

Let us use $>$ for conditionals in the indicative mood, and ${}_t>$ for conditionals in the subjunctive mood. The subscript t reminds us that the evaluation of a subjunctive conditional requires us to move to the antecedent world that is most similar to the actual world as it has been immediately before the point of time to which the antecedent refers. Let $s(A)$ denote the most similar A-world to the actual world, and $s_t(A)$ the most similar A-world to the actual world immediately before the time at which (the event or fact described by) A occurs.

As in the Introduction, let us assume that we believe to be true that Oswald actually killed Kennedy. Whether or not there was a conspiracy to assassinate Kennedy, we believe $\neg O > S$ to be true. Even if Oswald acted alone in killing Kennedy, $\neg O > S$ is believed to be true in $s(\neg O)$ because we believe that Kennedy has been assassinated in $s(\neg O)$. The fact that Kennedy has been assassinated in the actual world makes $s(\neg O)$ more similar to it than a world in which Kennedy has not been assassinated. The crux is that this fact is not part of the actual world restricted in time up to immediately before Oswald assassinated Kennedy. If we judge which world is more similar to the the actual world as it has been before Oswald assassinated Kennedy at time t, then the fact that Kennedy has been killed at time t plays obviously no role. Consequently, this fact is not available to single out $s_t(\neg O)$.

According to our proposed amendment to Stalnaker's semantics, $s_t(\neg O)$ is the world in which Oswald did not kill Kennedy and is otherwise most similar to the actual world as it has been immediately before the point of time at which Oswald killed Kennedy. Now, if Oswald acted alone in killing Kennedy, we believe $\neg O_t{>}S$ to be false; if there was a conspiracy to assassinate Kennedy, we believe $\neg O_t{>}S$ likely to be true. In both cases, we do not consider whether or not Kennedy actually has been assassinated. The amendment explains this: as we move to immediately before the time referred to in the antecedent, we do not know whether or not Kennedy has been assassinated. Hence, we are free to believe either way, probably guided by our context knowledge and system of variable hypotheticals.

Let us indicate how our method would apply to the learning of subjunctive conditionals. Assume our agent to be modelled is firmly convinced that Oswald killed Kennedy, that is O. Furthermore, it seems to her that Oswald having killed Kennedy implies that no one else did. From this follows the material implication $O \rightarrow \neg S$. Hence, the possibility $O \wedge S$ is already excluded.[22] Finally, let us assume that she does not believe in a conspiracy theory according to which many powerful people planned to kill Kennedy. In fact, she beliefs the anti-conspiracy hypothesis "If Oswald had not killed Kennedy, no one would have". To her, this is a counterfactual because she already believes O. We may define the counterfactual (relative to our agent) $\neg O_t^{\neg}{>}\neg S$ by $O \wedge (\neg O_t{>}\neg S)$.[23] If she now learns $\neg O_t{>}S$ (from another person) and accepts it, she learns about a conspiracy theory and accepts it. In reality, of course, the more likely outcome of this situation is a Ramsey Test disagreement, at least if the other person believes the same factual information $O \wedge \neg S$.

Learning (and accepting) $\neg O_t{>}S$ is tantamount to learn (and accept) that the possibility $\neg O \wedge \neg S$ is excluded. By the default assumption, the most similar possible world from the excluded $\neg O \wedge \neg S$-world is the world $s_t(\neg O)$ that satisfies S. If there is still some small probabilities associated to the $\neg O$-worlds, they are transferred to this $\neg O \wedge S$-world. If $\neg O$ is already probabilistically excluded, because the agent firmly believes O in the sense that $P(O) = 1$, no probabilities are transferred. However, note that the agent still learns from $\neg O_t^{\neg}{>}S$ that $\neg S$ is not the case under the supposition of $\neg O$.[24] We conjecture that our method should, in principle, be applicable to model the learning of

[22] It seems intuitively safe to say that a person cannot be killed by two different assassins. But then, there *are* cases of overdetermination. Let us just assume that the agent we model does not consider cases of overdetermination, like the one where two assassins shoot at exactly the same time.

[23] In general, we may define a counterfactual $A_t^{\neg}{>}C$ by $\neg A \wedge (A_t{>}C)$.

[24] Here, one might think about an amendment to our amendment. We might add probabilities to our learning method that

subjunctive and counterfactual information. However, the respective model will be more complicated than the corresponding indicative model, because time-indexed duplicate worlds need to be included. If our conjecture is correct, our amended learning method is even more general than before, in so far it would also capture the learning of subjunctively conditional information.

We turn now to the learning of (uncertain) causal information by adapting the proposed method. The adaptation is inspired by Lewis's (1973c) notion of causal dependence which we apply to Stalnaker conditionals. This move will allow us to formally implement Douven's (2012) idea of how the degree of belief in the antecedent should change as a result of its explanatory status when learning a conditional. The combination of the methods provides a unified framework that manages to clearly discern between the more informative causal and merely conditional interpretation of a conditional.

2.5 Learning Causal Information

In the previous sections, we proposed and tested a method of learning conditional information. We have shown that the predictions of the proposed method align with the intuitions in Douven's (2012) benchmark examples and that it can generate predictions for the Judy Benjamin Problem. Now, we adapt the method of learning conditional information to a method of learning causal information. The adapted method allows us to conceive causally of the information conveyed by the conditionals uttered in Douven's examples and the Judy Benjamin Problem. The combination of the proposed learning method and its adaptation amounts to a unified framework for the learning of (uncertain) conditional and causal information.

It may come as a surprise that we propose an account of learning that involves (Jeffrey) imaging. After all, the standard view on learning that α is Bayesian conditionalization on α, while David Lewis's imaging on α is widely conceived of as modeling the supposition of α. But learning a conditional may – according to the suppositional view on conditionals – be interpreted as learning what is true under a supposition (about which we may be uncertain). In particular, learning the conditional "If α, then γ" is thus equivalent to learning the conditional information that γ is the case under the supposition that α is the case.

Douven aims to provide an account of learning conditional information that is an empirically adequate account of human reasoning. Douven and Verbrugge (2010) submitted the thesis whether the acceptability of an indicative conditional 'goes by' the conditional probability of its consequent given the antecedent to empirical testing, and claim that the experiments speak against the thesis.[25] Their results indicate that conditional probabilities do not correspond to probabilities of conditionals. This was proven by Lewis (1976) when conditionals are to be understood as Stalnaker conditionals. These formal and empirical results obviously provide a severe challenge for Bayesian analyses of learning conditionals, where conditional probabilities usually take center stage.

Moreover, Zhao et al. (2012) obtained empirical results that indicate a fundamental difference between supposing and learning. In particular, supposing a conditional's antecedent α seems to have less impact on the credibility of the consequent γ than learning that α is true. We will provide a framework that allows us both, to distinguish between the learning of 'factual' and conditional information and to generate empirically testable predictions.

---

are not actual, or equivalently counterfactual, in the actual world, but that have been actual immediately before the belief as regards the antecedent was settled. See, for instance, Edgington (2008), Leitgeb (2012a,b), and Hájek (2014) for accounts of such counterfactual probabilities.

[25] The 'goes by' is Lewis formulation that may be found in Lewis (1976, p. 297).

In Section 2.6, we introduce Douven's desideratum for accounts of learning (uncertain) conditional information. His own proposal is based on the explanatory status of the antecedent. In Section 2.7, we sketch his argumentation against the method of imaging on the Stalnaker conditional as an account of learning conditional information. The reason for Douven's dismissal of the method is that the rationality constraints of Stalnaker models are not sufficient to single out a model, which may count as a representation of a belief state.

In Section 2.8, we adapt the method of learning conditional information to a method of learning causal information. The adaptation is inspired by Lewis's notion of causal dependence and replaces the pragmatic assumption by the assumption that the antecedent makes a difference. In Section 2.9, we apply our adapted method of learning causal information to Douven's examples and the Judy Benjamin Problem. In Section 2.10, we formally implement Douven's idea concerning the explanatory status of the antecedent within our framework.

2.6 Douven's Account of Learning Conditional Information via the Explanatory Status of the Antecedent

Douven (2015) propounds a broadly Bayesian model of learning conditional information. As the standard Bayesian view of learning, Douven's account assumes that learning the unnested indicative conditional "If α, then γ" implies that the posterior degree of belief for γ given α is set to approximately 1, i. e. $P^*(\gamma \mid \alpha) \approx 1$. In contrast to standard Bayesian epistemology, explanatory considerations play a major role in his model of updating on conditionals.

Douven proposes a desideratum for any account of learning conditional information, viz. a criterion that determines whether an agent raises, lowers, or leaves unchanged her degree of belief $P(\alpha)$ for the antecedent upon learning a conditional. He even writes that we "should be [...] dissatisfied with an account of updating on conditionals that failed to explain [...] basic and compelling intuitions about such updating, such as, in our examples".[26] Douven's methodology consists in searching for an updating model that accounts for our intuitions with respect to three examples, the Sundowners Example, the Ski Trip Example and the Driving Test Example. The three examples represent the classes of scenarios, in which $P(\alpha)$ should intuitively remain unchanged, be increased and decreased, respectively. He dismisses any method of learning conditional information that cannot account for all of the three examples. He emphasises that no single account of learning uncertain conditional and/or causal information is capable of solving all of his examples. Taking the examples as benchmark, he also dismisses the Stalnaker conditional as a tool to model the learning of conditional information. We have already shown that he errs on both: our account of learning conditional information succeeds in modelling all of his benchmark examples and is based on the Stalnaker conditional.

The core hypothesis of Douven's account is that the change in explanatory quality or 'explanatory status' of the antecedent α during learning the information results in a change of the degree of belief for α. If the explanatory status of α goes up, that is α explains γ well, then the degree of belief after learning the conditional increases, i. e. $P^*(\alpha) > P(\alpha)$; if the explanatory status of α goes down, $P^*(\alpha) < P(\alpha)$; if the explanatory status remains the same, a variant of Jeffrey conditioning is applied that has the property that $P^*(\alpha) = P(\alpha)$. Following Bradley (2005), Douven calls this Jeffrey conditioning over a restricted partition 'Adams conditioning on $P^*(\gamma \mid \alpha) \approx 1$'.[27]

[26] Douven (2012, p. 3).

[27] The partition is restricted according to the odds for the consequent of the learned conditional. For details, see Bradley (2005, pp. 351–52), and Douven and Romeijn (2011, pp. 650–53).

Douven and Romeijn proposed a solution to the Judy Benjamin Problem. The problem indicates that the revision method that minimises the Kullback-Leibler divergence leads to counterintuitive results for learning uncertain conditional information. Their solution uses the variant of Jeffrey conditioning mentioned above. However, their proposed method fails to account for examples where the probability of the antecedent is supposed to change, since it has the invariance property that $P^*(\alpha) = P(\alpha)$, for all α, and thus disqualifies as a general account of learning conditional information.[28]

2.7 Douven's Dismissal of the Stalnaker Conditional

Douven claims that Stalnaker conditionals are not suited to model the learning of conditional information. He argues for this claim by pointing out that a learning method based on the Stalnaker conditional "makes no predictions at all about any of our examples".[29] The cited reason is that we would not be able to exclude certain Stalnaker models as rational representation of a belief state.

Recall from 2.2 that Douven provides three possible worlds models for his point. Each model consists of four worlds such that all logical possibilities of two binary variables are covered. He observes that imaging on "If α, then β" interpreted as a Stalnaker conditional has different effects: in model I the probability of the antecedent α, i. e. $P(\alpha)$ decreases, in model II $P(\alpha)$ remains unchanged, and in model III $P(\alpha)$ increases. According to Douven this flexibility of the class of possible worlds models is a problem rather than an advantage, since there would be no rationality constraints to rule out certain models as rational representations of a belief state.

Consider a scenario of the class, where the antecedent remains unchanged (e.g. the Sundowners Example). The problem is, so Douven, that there are no criteria to exclude models I and III as rational representations of a belief state, in which $P(\alpha)$ should not change. In Douven's words:

> In fact, to the best of my knowledge, nothing said by Stalnaker (or Lewis, or anyone else working on possible worlds semantics) implies that, supposing imaging is the update rule to go with Stalnaker's account, models I and III [...] could not represent the belief state of a rational person; [...] In short, interpreting "If A, B" as the Stalnaker conditional and updating on it [...] by means of imaging offers no guarantee that our intuitions are respected about what should happen – or rather *not* happen – after the update [...]. Naturally, it cannot be excluded that some of these models – and perhaps indeed all on which [...] [the degree of belief in the antecedent] changes as an effect of learning [the conditional] – are to be ruled out on the basis of rationality constraints that I am presently overlooking, perhaps ones still to be uncovered, or at least still to be related to possible worlds semantics as a tool for modelling epistemic states. It is left as a challenge to those attracted to the view considered here to point out such additional constraints.[30]

In the previous sections, we met the challenge Douven mentions in the quote. We discovered two constraints that singled out Stalnaker models that plausibly represent the belief states in Douven's benchmark examples. Imposing the two additional constraints amounts to interpreting the meaning of a Stalnaker conditional in a minimally informative way and supplementing the analysis by a default assumption. However, our account is somewhat incomplete because it cannot model explanatory or causal information conveyed by a conditional.

[28] Cf. Douven and Romeijn (2011, pp. 648–655), Douven (2012, pp. 9–11).
[29] Douven (2012, p. 7).
[30] Douven (2012, pp. 8-9).

2.8 An Adaptation of the Method to the Learning of Causal Information

In Section 2.6, we have seen that Douven invokes explanatory considerations in order to model the learning of conditional information. His account presupposes an explanatory reading of the learned conditional information, which may be of the form 'If α, then γ'. While we are skeptical about the presupposition that *any* conditional can or should be read as (a part of) an explanation or causal dependence, we admit that conditionals often figure in explanations. Hence, the proposed method of learning conditional information should be able to account for the learning of causal information conveyed by conditionals; otherwise, the proposed method suffers a major drawback.

In this section, we sketch how the proposed method may be adapted to a method of learning causal information. The adaptation is inspired by Lewis's analysis of causal dependence in terms of counterfactuals. Douven claims that, in any account of explanation that relies on a Stalnaker model, 'to explain' means to 'provide causal information', where 'causal' refers to a Lewis style analysis.[31]

We write $\alpha \Rightarrow \gamma$ for the causal reading of "If α, then γ". According to Lewis's idea of causal dependence, $\alpha \Rightarrow \gamma$ is satisfied iff $\alpha > \gamma$ and $\neg\alpha > \neg\gamma$. We may apply the proposed method of learning conditional information by taking the minimally informative meaning of $\alpha \Rightarrow \gamma$ into account (instead of the one of $\alpha > \gamma$), if we substitute the default assumption. We call the adaptation the 'method of learning causal information'.

The substitution of the default assumption to what we call 'causal difference assumption' runs as follows. Assume we have no further contextual knowledge. Then, the most similar $\alpha \Rightarrow \gamma$-world from any excluded $\alpha \Rightarrow \neg\gamma$-world is a $(\alpha \wedge \gamma)$-world, if the excluded $\alpha \Rightarrow \neg\gamma$-world satisfies α. Furthermore, the most similar $\alpha \Rightarrow \gamma$-world from any excluded $\alpha \Rightarrow \neg\gamma$-world is a $(\neg\alpha \wedge \neg\gamma)$-world, if the excluded $\alpha \Rightarrow \neg\gamma$-world satisfies $\neg\alpha$. In symbols,

$$min_{\leq_{w_\alpha \Rightarrow \neg\gamma}}[\alpha \Rightarrow \gamma] = \left\{ \begin{array}{ll} w_{\alpha\wedge\gamma} & \text{if } w_{\alpha\Rightarrow\neg\gamma}(\alpha) = 1 \\ w_{\neg\alpha\wedge\neg\gamma} & \text{if } w_{\alpha\Rightarrow\neg\gamma}(\alpha) = 0 \end{array} \right\} \tag{2.18}$$

The causal difference assumption is justified, if we understand causal dependence as difference making à la Lewis.[32] The antecedent α makes the difference as to whether γ or $\neg\gamma$. Hence, $\alpha \Rightarrow \gamma$ means that worlds in which α obtains are worlds in which γ obtains, and accordingly that worlds in which $\neg\alpha$ obtains are worlds in which $\neg\gamma$ obtains. It is built in the analysis of causal dependence, so to speak, that the difference making factors (α and $\neg\alpha$) are more dissimilar than the ensuing effects.

Note that causal dependence is more informative than conditional dependence. For, the minimally informative meaning of $[\alpha \Rightarrow \gamma]$ is in the absence of further information always a strict subset of the minimally informative meaning of $[\alpha > \gamma]$. The reason is that causal dependence, by definition, conveys in addition to the indicative conditional information also the information $[\neg\alpha > \neg\gamma]$. In brief, if an agent learns $\alpha \Rightarrow \gamma$, our adapted method prescribes that the $\alpha \wedge \neg\gamma$-worlds transfer their

[31] Cf. Douven (2012, pp. 8-9, especially footnote 7), Lewis (1973c). Furthermore, Douven claims that Lewis's and Stalnaker's semantics for conditionals are "exactly the same" (Douven (2012, p. 8, lines 8–11)). However, there is a difference between the Stalnaker and Lewis semantics. In a Stalnaker model, there is always a single most similar world (or no world) to the actual world, whereas Lewis's semantics allows for a *set* of worlds (or no world) whose elements are equally similar to the actual world. A consequence of the difference is that Lewis's 'official' semantics for conditionals, i. e. the system **VC**, does not validate the principle of Conditional Excluded Middle, whereas Stalnaker's logic **C2** for conditionals does. In Lewis's nomenclature, system **C2** is labelled by **VCS**. (Cf. Lewis (1973b), Lewis (1973a), and, for details, Unterhuber (2013, especially Chapter 3.2 and Chapters 3.3.3 and 3.3.4)) The non-identity of Lewis's and Stalnaker's semantics implies that our method of learning causal information is not equivalent to Lewis's notion of causal dependence. While the method relies on Lewis's *idea*, we stick to Stalnaker's semantics in this paper.

[32] Cf. Lewis (1973c).

probability shares to the most similar $\alpha \wedge \gamma$-world, and the $\neg\alpha \wedge \gamma$-worlds transfer their probability shares to the most similar $\neg\alpha \wedge \neg\gamma$-world. In other words, if the antecedent α is a difference maker, then the probability mass of those worlds w that do not satisfy $\alpha \Rightarrow \gamma$ is shifted to the most similar $\alpha \Rightarrow \gamma$-world w' that agrees with the Boolean evaluation for α, i. e. $w(\alpha) = w'(\alpha)$.

2.9 Douven's Examples and the Judy Benjamin Problem as Causal Scenarios

We apply now our adapted method of learning causal information to Douven's examples and the Judy Benjamin Problem.

Recall Example 4, the Sundowners Example, from Section 2.2.6. We model Sarah's belief state as the Stalnaker model depicted in Figure 2.8. W contains four elements covering the possible events of $R, \neg R, S, \neg S$, where R stands for "it rains tomorrow" and S for "Sarah and Marian can have sundowners at the Westcliff tomorrow".

Let us assume that Sarah interprets the conditional uttered by her sister Marian as conveying the causal information $R \Rightarrow \neg S$. As Douven himself points out, the intuition in the Sundowners Example derives from the verdict that whether or not it rains makes the difference as to whether or not they have sundowners, but not the other way around: having sundowners simply has no effect whatsoever on whether or not it rains.[33] Hence, the change of belief between R and $\neg R$ is more far fetched than between S and $\neg S$. In other words, the worlds along the horizontal axis are more similar than the worlds along the vertical axis. Since $R \Rightarrow \neg S \equiv (R > \neg S) \wedge (\neg R > S)$, $R \Rightarrow \neg S$ expresses both that S is excluded under the supposition of R and $\neg S$ is excluded under the supposition of $\neg R$. By the causal difference assumption, we obtain $min_{\leq_{w_1}}[R > \neg S] = w_2$ and $min_{\leq_{w_4}}[\neg R > S] = w_3$. Lewis's imaging method results in a shift of probability shares along the horizontal axis of Figure 2.8.

Imaging on the minimally informative proposition $[R \Rightarrow \neg S] = \{w_2, w_3\}$ results in $P^{R \Rightarrow \neg S}(w') = \sum_w P(w) \cdot \left\{ \begin{array}{ll} 1 & \text{if } w_{R \Rightarrow \neg S} = w' \\ 0 & \text{otherwise} \end{array} \right\}$:

$$\begin{array}{ll} P^{R \Rightarrow \neg S}(w_1) = 0 & P^{R \Rightarrow \neg S}(w_2) = P(w_1) + P(w_2) \\ P^{R \Rightarrow \neg S}(w_3) = P(w_3) + P(w_4) & P^{R \Rightarrow \neg S}(w_4) = 0 \end{array} \tag{2.19}$$

We see immediately that both intuitions associated with the Sundowners Example are satisfied, viz. $P^{R \Rightarrow \neg S}(R) = P(R) = P(w_1) + P(w_2)$ and $P^{R \Rightarrow \neg S}(R \wedge S) = P(w_1) = 0$. We conclude that the method of learning causal information yields the intuitively correct results. Our account, thus, allows to conceive of the conditional learned in the Sundowners Example as conveying causal information.

Recall Example 5, the Ski Trip Example, from Section 2.2.6. We model Harry's belief state as the Stalnaker model depicted in Figure 2.9. W contains eight elements covering the possible events of $E, \neg E, S, \neg S, B, \neg B$, where E stands for "Sue passed the exam", S for "Sue's father takes her on a skiing vacation", and B for "Sue buys a skiing outfit".

We assume that Harry interprets the conditional uttered by his friend Tom as conveying the causal information $E \Rightarrow S$. Furthermore, The Ski Trip Example assumes that Harry is equipped with the following contextual knowledge: Sue buying a skiing outfit may causally depend on the invitation of Sue's father to a skiing vacation, in symbols $S \Rightarrow B$. Finally, Harry observed Sue buying a skiing outfit, and thus has the factual information that B.

[33] Cf. Douven (2012, p. 8).

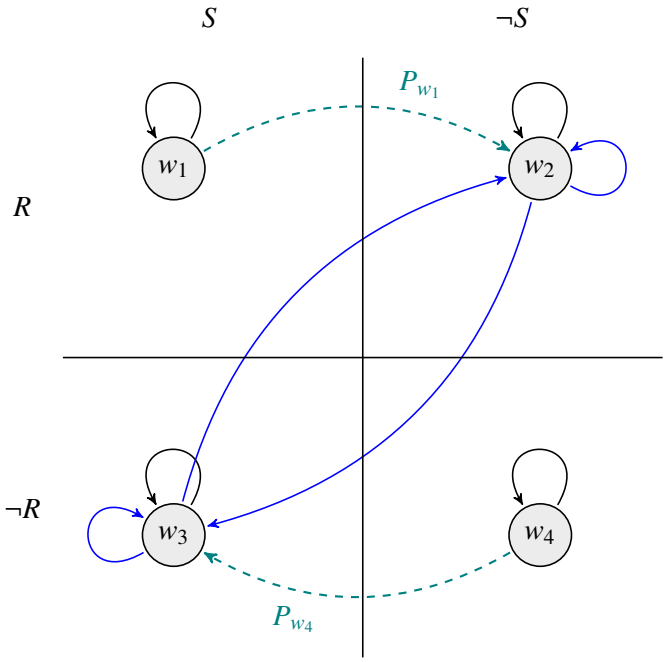

Figure 2.8: A Stalnaker model for Sarah's belief state in the Sundowners Example. The blue arrows illustrate the change of the similarity order such that the received information, causally understood as $R \Rightarrow \neg S$, is minimally informative. Here, the minimally informative meaning of $R \Rightarrow \neg S$ is $[R \Rightarrow \neg S] = [R > \neg S] \cap [\neg R > S] = \{w_2, w_3\}$.

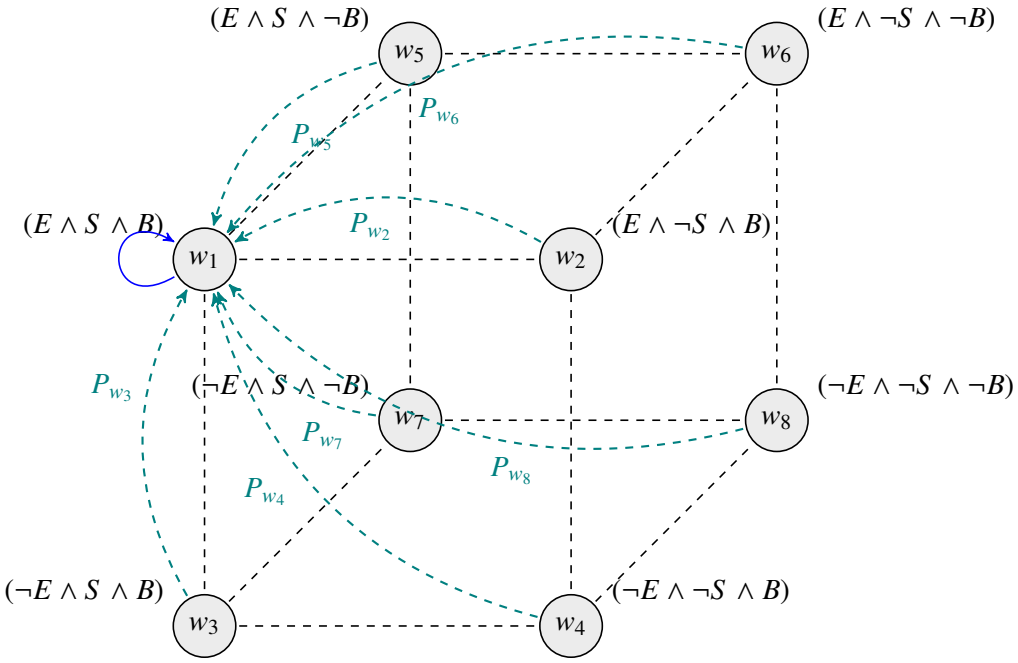

Figure 2.9: An eight-worlds Stalnaker model for Harry's belief state in the Ski Trip Example. Harry learns the minimally informative proposition $[(E \Rightarrow S) \wedge (S \Rightarrow B)] = \{w \in W \mid (min_{w_\leq}[E] \in [S]) \wedge (min_{w_\leq}[\neg E] \in [\neg S]) \wedge (min_{w_\leq}[S] \in [B]) \wedge (min_{w_\leq}[\neg S] \in [\neg B])\} = \{w_1, w_8\}$. Since Harry also obtains the factual information B, we can also exclude the $\neg B$-world w_8.

In total, Harry learns the minimally informative proposition $[(E \Rightarrow S) \wedge (S \Rightarrow B) \wedge B] = \{w_1\}$. Since w_1 is the only world that is not probabilistically excluded, we do not need to appeal to the causal difference assumption in this example.

Imaging on the minimally informative proposition $[(E \Rightarrow S) \wedge (S \Rightarrow B) \wedge B] = \{w_1\}$ results in the following probability distribution, where we do not display the vanishing probabilities:

$$P^{(E \Rightarrow S) \wedge (S \Rightarrow B) \wedge B}(w') = P^*(w') = \sum_w P(w) \cdot \left\{ \begin{array}{ll} 1 & \text{if } w_{(E \Rightarrow S) \wedge (S \Rightarrow B) \wedge B} = w' \\ 0 & \text{otherwise} \end{array} \right\}:$$

$$P^*(w_1) = 1 \tag{2.20}$$

The result meets the intuition associated with the Ski Trip Example: $P^*(E) > P(E)$, since $P^*(E) = P^*(w_1)$ and $P(E) = P(w_1) + P(w_2) + P(w_5) + P(w_6)$. Later on, we will see that the probabilities of the worlds w_2, w_3, w_4 would not have vanished entirely, if either $E \Rightarrow S$ or $S \Rightarrow B$ (or both) had conveyed only uncertain information.

In Section 2.2.6, when we interpreted the uttered conditional as merely conveying conditional information, we needed to rely on the default assumption to model the Ski Trip Example. If we appeal to the causal interpretation in the Ski Trip Example, we need neither the default assumption nor the causal difference assumption any more. Again, our framework allows to conceive of the conditional learned in the Ski Trip Example as conveying causal information.

Recall Example 6, the Driving Test Example, from Section 2.2.6. We model Betty's belief state as the Stalnaker model depicted in Figure 2.10. W contains eight elements covering the possible events of $D, \neg D, G, \neg G, S, \neg S$, where D stands for "Kevin passed the driving test", G for "Kevin's parents

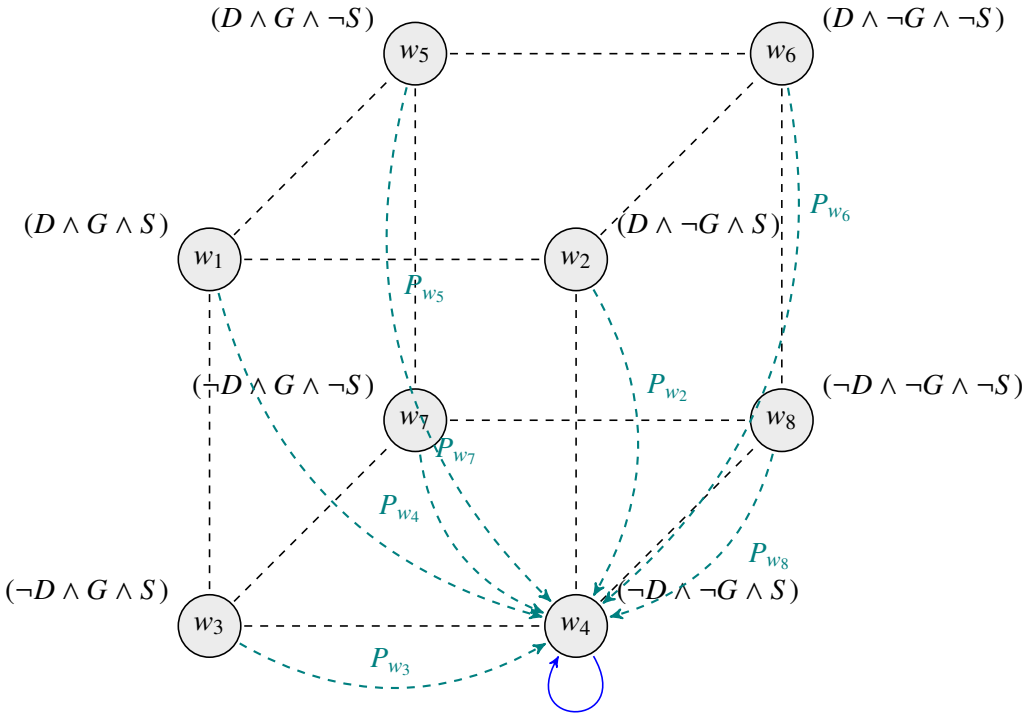

Figure 2.10: An eight-worlds Stalnaker model for Betty's belief state in the Driving Test Example.

will throw a garden party", and S for "Kevin's parents have started to spade their garden".

Assume Betty interprets the conditional uttered by her mother as the causal information $D \Rightarrow G$. Furthermore, Betty infers from her contextual knowledge that because Kevin's parents are spading their garden, they will not throw a garden party, in symbols $S \Rightarrow \neg G$. Finally, Betty knows that Kevin's parents have started to spade their garden, and thus has the factual information that S.

In total, Betty learns the minimally informative proposition $[(D \Rightarrow G) \wedge (S \Rightarrow \neg G) \wedge S] = \{w_4\}$. In Figure 2.10, we see that the Driving Test Example is structurally similar to the Ski Trip Example.

Imaging on the minimally informative proposition $[(D \Rightarrow G) \wedge (S \Rightarrow \neg G) \wedge S] = \{w_4\}$ results in the following probability distribution, where we do not display the vanishing probabilities:

$$P^{(D \Rightarrow G) \wedge (S \Rightarrow \neg G) \wedge S}(w') = P^*(w') = \sum_w P(w) \cdot \left\{ \begin{array}{ll} 1 & \text{if } w_{(D \Rightarrow G) \wedge (S \Rightarrow \neg G) \wedge S} = w' \\ 0 & \text{otherwise} \end{array} \right\}:$$

$$P^*(w_4) = 1 \tag{2.21}$$

Our method yields again the correct result regarding the intuition associated with the Driving Test Example: $P^*(D) < P(D)$, since $P^*(D) = 0$ and $P(D) = P(w_1) + P(w_2) + P(w_5) + P(w_6) > 0$. Again, our framework allows to conceive of the conditional learned in the Driving Test Example as conveying causal information.

The following Judy Benjamin Problem will illustrate that if Betty thinks that the conditionals $D \Rightarrow G$ or $S \Rightarrow \neg G$ (or both) convey uncertain information, then the probability shares for some other worlds will not reduce to zero. This fact fits nicely with the Driving Test Examples's remark that "given the spading that has just begun, it is doubtful [or uncertain] (even if not wholly excluded) that a party can be held in the garden of Kevin's parents". We exemplify the application of our method to

the learning of uncertain causal information now.

Recall Example 7, the Judy Benjamin Problem, from Section 2.2.6. We apply now our method of learning causal information to this case, in which the received information is uncertain. We show thereby that the method may be generalised to cases in which the learned information is uncertain, provided we use Jeffrey imaging.

We model Judy Benjamin's belief state as the Stalnaker model depicted in Figure 2.11. W contains four elements covering the possible events of $R, \neg R, S, \neg S$, where R stands for "Judy Benjamin's platoon is in Red territory", and S for "Judy Benjamin's platoon is in Second Company area". The story prescribes that the probability distribution before learning the uncertain information is given by:

$$P(R \wedge S) = P(R \wedge \neg S) = P(\neg R \wedge S) = P(\neg R \wedge \neg S) = \frac{1}{4} \tag{2.22}$$

In the previous examples, our agents implicitly learned Stalnaker conditionals of the form $\alpha > \gamma$ with certainty. According to Theorem 1, this amounts to the constraint that $P(\alpha > \gamma) = P^{\alpha}(\gamma) = 1$ (provided α is not a contradiction). Given this constraint and since P^{α} is a probability distribution, we have $P^{\alpha}(\neg\gamma) = 1 - P^{\alpha}(\gamma) = 0$. This means that we were able to probabilistically exclude any $\neg\gamma$-world under the supposition of α.

Now, our agent Judy Benjamin learns uncertain causal information, i. e. she implicitly learns Stalnaker conditionals with uncertainty. According to Theorem 1 and since $R \Rightarrow S$ is equivalent to $(R > S) \wedge (\neg R > \neg S)$, this amounts in the Judy Benjamin Problem to the constraint that $P(R \Rightarrow S) = \frac{3}{4}$. By our method, we obtain $P(R \Rightarrow \neg S) = \frac{1}{4}$. In contrast to learning causal information with certainty, we cannot subtract the whole probabilistic mass from the $\neg S$-worlds under the supposition of R, and accordingly from the S-worlds under the supposition of $\neg R$. However, Judy Benjamin is informed from an external source about the proportion to which she should gradually 'exclude' or downweigh the probability share of $R \Rightarrow \neg S$-worlds. Equivalently, we may say that the most similar $R \Rightarrow S$-world (from any $R \Rightarrow \neg S$)-world) obtains a gradual upweight of probability such that it receives $\frac{3}{4}$ of the probability shares of the $R \Rightarrow \neg S$-worlds; in turn, however, this $R \Rightarrow \neg S$-world then receives a probability share from the $R \Rightarrow S$-world weighed by $\frac{1}{4}$. Note that in Stalnaker models $R \Rightarrow \neg S$ is equivalent to $(R > \neg S) \wedge (\neg R > S)$.

We apply now Jeffrey imaging to the Judy Benjamin Problem, where a source external to Judy provides her with the information that $k = \frac{3}{4}$.

$$P_k^{R \Rightarrow S}(w') = \sum_w \left(P(w) \cdot \left\{ \begin{array}{ll} k & \text{if } w_{R \Rightarrow S} = w' \\ 0 & \text{otherwise} \end{array} \right\} + P(w) \cdot \left\{ \begin{array}{ll} (1-k) & \text{if } w_{R \Rightarrow \neg S} = w' \\ 0 & \text{otherwise} \end{array} \right\} \right) \tag{2.23}$$

Given the probability distribution before the learning process in Equation (2.22), Judy obtains the following probability distribution after being informed that $P(R \Rightarrow S) = \frac{3}{4}$:

$$\begin{array}{ll} P_{\frac{3}{4}}^{R \Rightarrow S}(w_1) = P_{\frac{3}{4}}^{R \Rightarrow S}(R \wedge S) = \frac{3}{8} & P_{\frac{3}{4}}^{R \Rightarrow S}(w_2) = P_{\frac{3}{4}}^{R \Rightarrow S}(R \wedge \neg S) = \frac{1}{8} \\ P_{\frac{3}{4}}^{R \Rightarrow S}(w_3) = P_{\frac{3}{4}}^{R \Rightarrow S}(\neg R \wedge S) = \frac{1}{8} & P_{\frac{3}{4}}^{R \Rightarrow S}(w_4) = P_{\frac{3}{4}}^{R \Rightarrow S}(\neg R \wedge \neg S) = \frac{3}{8} \end{array} \tag{2.24}$$

The probability distribution of (2.24) does not conform to Douven's intuitively correct distribution of (2.14), while the desideratum $P_{\frac{3}{4}}^{R > S}(R) = P(R) = \frac{1}{2}$ is met. Note that the learning of causal information results in $P_{\frac{3}{4}}^{R \Rightarrow S}(\neg R \wedge \neg S) = \frac{3}{8}$, which may be plausible for cases of causal dependence. However, we do not think that the conditional of the Judy Benjamin Problem is meant to express a

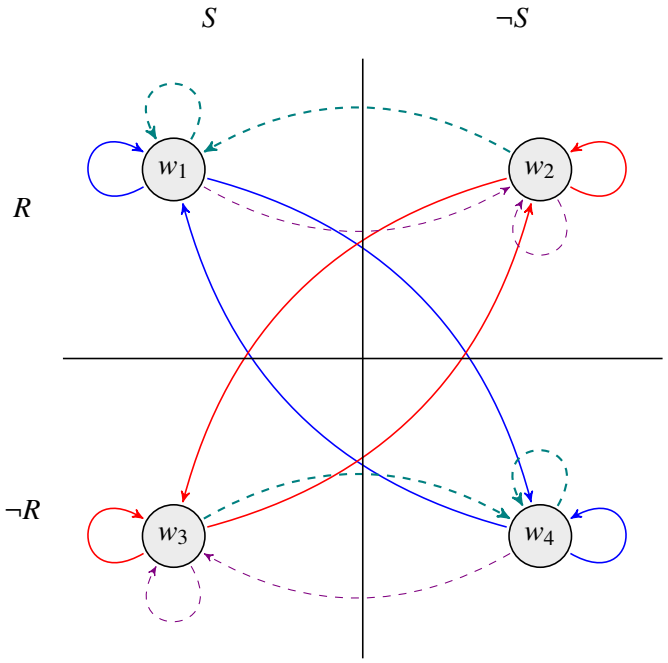

Figure 2.11: A Stalnaker model for Private Benjamin's belief state in the Judy Benjamin Problem. The blue arrows illustrate the specification of a similarity order $\leq'$ such that the received information $[R \Rightarrow S]$ is minimally informative. Note that each world having two outgoing blue arrows (one for $R > S$ and one for $\neg R > \neg S$) satisfies $R \Rightarrow S$. The red arrows illustrate the specification of another similarity order $\leq \neq \leq'$ such that the received information $[R \Rightarrow \neg S]$ is minimally informative. Each world having two outgoing red arrows (one for $R > \neg S$ and one for $\neg R > S$) satisfies $R \Rightarrow \neg S$). In sum, the similarity orders are specified such that one makes $[R \Rightarrow S] = \{w_1, w_4\}$ a minimally informative proposition and the other makes the complement proposition $[R \Rightarrow \neg S] = \{w_2, w_3\}$ a minimally informative proposition. By the causal difference assumption, we obtain $min_{\leq'_{w_2}}[R \Rightarrow S] = w_1$ and $min_{\leq'_{w_3}}[R \Rightarrow S] = w_4$. Furthermore, we obtain $min_{\leq_{w_1}}[R \Rightarrow \neg S] = w_2$ and $min_{\leq_{w_4}}[R \Rightarrow \neg S] = w_3$. The teal arrows represent the transfer of $k \cdot P(w)$, while the violet arrows represent the transfer of $1-k \cdot P(w)$. The application of Jeffrey imaging on $[R \Rightarrow S]$ with $k = \frac{3}{4}$ leads to the following calculation for the probability distribution: $P^{R \Rightarrow S}_{\frac{3}{4}}(w_1) = \frac{3}{4} \cdot P(w_1) + \frac{3}{4} \cdot P(w_2)$, and $P^{R \Rightarrow S}_{\frac{3}{4}}(w_2) = \frac{1}{4} \cdot P(w_1) + \frac{1}{4} \cdot P(w_2)$, and $P^{R \Rightarrow S}_{\frac{3}{4}}(w_3) = \frac{1}{4} \cdot P(w_3) + \frac{1}{4} \cdot P(w_4)$, and $P^{R \Rightarrow S}_{\frac{3}{4}}(w_4) = \frac{3}{4} \cdot P(w_3) + \frac{3}{4} \cdot P(w_4)$.

causal dependence relation. In Section 2.2.6, we treated the received uncertain conditional as merely carrying uncertain conditional information. Applying the method of learning uncertain conditional information allowed us to offer a solution to the Judy Benjamin Problem that agrees with Douven's desired distribution of (2.14).

The Judy Benjamin Problem illustrates quite vividly the main difference between learning conditional and causal information. A merely conditional understanding of the conditional in the Judy Benjamin Problem does not affect the (row of) $\neg\alpha$-worlds, whereas the difference-making or causal dependence interpretation of the conditional affects the (row of) $\neg\alpha$-worlds.

2.10 Stalnaker Inferences to the Explanatory Status of the Antecedent

The method of learning causal information provides a formally precise implementation for when and how Douven's explanatory status of the antecedent should change. Recall his idea from Section 2.6 that the explanatory power of the antecedent with respect to the consequent determines the probability of the antecedent after learning the conditional. The idea is related to abduction, nowadays more commonly referred to as 'inference to the best explanation', or at least to a good explanation. The schema of such an inference runs as follows: α explains γ (well), and γ obtains. Therefore, α is true, or at least more likely.

We may interpret a Stalnaker agent's learning of $\alpha \Rightarrow \gamma$ as inference to a good explanation. Suppose an agent believes the fact γ and receives the information $\alpha \Rightarrow \gamma$. Then the agent infers that α explains γ (well).[34] For, $\alpha \Rightarrow \gamma$ implies that $\neg\gamma$ would be the case, if α were not the case. But γ is the case and thus indicates that α is the case as well. The Ski Trip Example is an instance of this type of reasoning. Harry learns $E \Rightarrow S$, $S \Rightarrow B$ and the fact B. He infers by our method of learning causal information that S explains B and, in turn, that E explains S. Consequently, $P^{(E\Rightarrow S)\wedge(S\Rightarrow B)\wedge B}(E) \geq P(E)$. In general, $P_k^{(\alpha\Rightarrow\gamma)\wedge\gamma}(\alpha) \geq P(\alpha)$, if $k > \frac{1}{2}$. In such a case, we call α the antecedent in a 'Stalnaker inference to a good explanatory status of the antecedent', or simply the antecedent in a 'Stalnaker inference to a good explanans'.

In the Driving Test Example, Kevin's passing the driving test (D) is at odds with the parent's spading their garden (S). D does not explain S (well). There is rather a tension between the occurrence of D and S. We can again formally implement the reasoning. Suppose S and $S \Rightarrow \neg G$, where G stands for "Kevin's parents will throw a garden party". Betty receives the information that $D \Rightarrow G$. S and $S \Rightarrow \neg G$ implies that G is not the case. By $D \Rightarrow G$, we may therefrom infer that D is not the case either. For, if D were the case, G would be the case. Consequently, $P^{(D\Rightarrow G)\wedge(S\Rightarrow\neg G)\wedge S}(D) \leq P(D)$. In general, $P_k^{(\alpha\Rightarrow\gamma)\wedge\neg\gamma}(\alpha) \leq P(\alpha)$, if $k > \frac{1}{2}$. In such a case, we call α the antecedent in a 'Stalnaker inference to a bad explanans'. Notice that our framework allows for a probabilification of the Stalnaker inferences, if uncertain causal information is learned.

2.11 Conclusion

We have seen that Douven's dismissal of the Stalnaker conditional as a tool to model the learning of conditional and causal information is unjustified. Rather, this type of learning may be modeled by Jeffrey imaging on the meaning of Stalnaker conditionals under the following condition: the similarity order of the Stalnaker model is changed in a way such that the meaning of the conditional is minimally

[34] We may also say that α is a ground for γ that makes the difference as to whether or not γ obtains.

informative. Both methods of learning information align with the intuititively correct results in Douven's benchmark examples. However, Douven's intuitions about the Judy Benjamin Problem are only met, if we understand the conditional Judy receives as conveying merely conditional information.

We have shown that the method of learning (uncertain) conditional information may be adpated to a learning method of (uncertain) causal information. The adaptation is based on the Stalnaker conditional, for which Lewis's idea of causal dependence is implemented. The two methods come with two different assumptions, viz. the default assumption and the causal difference assumption, respectively. The combination of the two methods provides a unified framework that manages to clearly discern between a merely conditional and a causal reading of the conditional "If α, then γ". Hence, the general method cannot be attacked for not being applicable to conditionals that (are supposed to) express causal dependences. In detail, if no further contextual information is available, conjunctive information is strictly more informative than causal information, which is in turn strictly more informative than conditional information. For, the minimally informative conjunctive, causal and conditional propositions stand in the following strict subset relation: $[\alpha \wedge \gamma] \subset [\alpha \Rightarrow \gamma] \subset [\alpha > \gamma]$.

The causal dependence reading can be used to formalise Douven's explanatory status of the antecedent. We thereby convey the explanatory status a precise formal meaning that may be used to operationalize Douven's idea that explanatory considerations play a core role in learning conditionals. Furthermore, the results suggest that we should distinguish between a merely conditional or suppositional interpretation and a causal dependence interpretation of a conditional. A supposition should not affect those cases, in which the antecedent is not satisfied, whereas a difference-making conditional should. Based on this distinction, we hope that the proposed framework can help psychologists of reasoning to provide an empirically adequate account of actual reasoning behaviour with respect to the learning of conditional and causal information.

The advantages of our unified framework of learning uncertain information, as compared to alternative accounts, will be assessed elsewhere. We plan to compare our account in detail to Douven's account of learning conditional information and Bayesian accounts of learning conditionals. In particular, we will show that the Bayesian account of Hartmann and Rad (2017) – that minimizes the Kullback-Leibler divergence on a fixed Bayesian network – has severe problems to capture the merely conditional interpretation of conditionals. As a consequence the Judy Benjamin Problem remains troublesome for their account.

Chapter 3

On the Ramsey Test Analysis of 'Because'

In the previous chapter, we were concerned with the learning of conditional and causal information. We proposed a method of learning conditional and causal information based on Stalnaker's (1968) semantics of conditionals. Now, we move on to the natural language semantics of conditionals and because. The assertion of a conditional 'if α, then γ' commonly implies that the antecedent α is in some way relevant for the consequent γ. The semantics of variably strict conditionals by Lewis (1973a) and Stalnaker (1968) spells out the relation of relevance between antecedent and consequent via a similarity order between possible worlds.[1] However, this semantics fails to account for the relevance between antecedent and consequent, at least in the indicative case. That is, if α and γ are true in the actual world, then the conditional $\alpha > \gamma$ is true in that world, independently of whether there is any connection between α and γ. For example, 'If Munich is a town in Germany, then Lund is a town in Sweden' is true, provided that Munich is a town in Germany and Lund a town in Sweden. This seems to be an absurd consequence.

The Ramsey Test approach to conditionals by Gärdenfors (1978) and Gärdenfors (1988) faces an analogous problem: if α and γ are believed to be true, then $\alpha > \gamma$ must be accepted in that approach. Hence, the formal semantics of conditionals by Stalnaker (1968), Lewis (1973a), Gärdenfors (1978), and Gärdenfors (1988) fail to distinguish between trivially true and non-trivially true indicative conditionals. This problem has been addressed by Rott (1986) in terms of a strengthened Ramsey Test. In this chapter, we refine Rott's variant of a strengthened Ramsey Test and the corresponding analysis of because sentences. Given that because is used to express explanatory relations, we show that our final analysis captures the presumed asymmetry between explanans and explanandum much better than Rott's original analysis.

Sources. This chapter builds on joint work with Holger Andreas. Substantial content of Andreas and Günther (2018) is reprinted by permission from Springer Nature: Springer Netherlands, *Erkenntnis*, On the Ramsey Test Analysis of 'Because', Holger, A. and Günther, M., License Number 4364830123295 (2018), advance online publication, 7 June 2018 (https://doi.org/10.1007/s10670-018-0006-8, Erkenn).

[1] Grove (1988) has shown that Lewis's and Stalnaker's possible worlds semantics can equivalently be spelled out in systems of spheres of possible worlds.

3.1 Introduction

We attempt to analyse the presumed relevance between a conditional's antecedent and its consequent by means of a strengthened Ramsey Test. More specifically, we suggest that a conditional be accepted iff it passes the following test:

> First, suspend judgement about the antecedent and the consequent. Second, add the antecedent (hypothetically) to your stock of explicit beliefs. Finally, consider whether or not the consequent is entailed by your explicit beliefs.

We believe that this variant of a strengthened Ramsey Test has interesting applications in different areas of philosophical logic:

1. The analysis of indicative, subjunctive, and counterfactual conditionals in natural language.
2. The analysis of the conjunction 'because' in natural language.
3. The logical analysis of explanations.
4. The conditional analysis of causation.

The chapters 4 and 5 of this dissertation are devoted to analyse causation in terms of our strengthened Ramsey Test. In this chapter, however, we focus on the semantic analysis of indicative conditionals and the word 'because' in natural language. In passing, we shall also make references to explanations, which seem to be related to an analysis of 'because' in natural language.

Let us briefly explain why the problem of relevance between antecedent and consequent is particularly pressing for an analysis of 'because'. If a speaker asserts 'γ because of α', then he or she already believes, or knows, that α and γ. Hence, the standard Ramsey Test conditional $\alpha > \gamma$ is far too weak for a conditional analysis of 'because'. For, this conditional does not require the antecedent to be relevant for the consequent in case α and γ are believed to be true.

In addition to the relevance between explanans and explanandum, our analysis aims to account for the presumed asymmetry of explanatory relations. As is widely agreed upon, the presence of a tower may well explain the occurrence of a shadow, but not vice versa. That is, we endorse 'there is a shadow because of the tower', but not 'there is a tower because of the shadow'. The structure of this simple example captures a large class of asymmetric explanatory relations. We take it as a starting point to eventually work out a general account of scientific explanations.

The following analysis of 'because' will be shown to yield the intended results for the tower-shadow scenario. Let $\gg$ designate our strengthened Ramsey Test conditional. That is, $\alpha \gg \gamma$ iff, after suspending judgment about α and γ, an agent can infer γ from the supposition of α (in the context of further beliefs in the background). The schema of our definition is then:

$$\textit{Because } \alpha, \gamma \text{ relative to } K(S) \quad \text{iff} \quad \alpha \gg \gamma \in K(S) \quad \text{and} \quad \alpha, \gamma \in K(S)$$

where $K(S)$ designates the belief set of the epistemic state S. We represent epistemic states by belief bases rather than belief sets. This proves crucial to account for the asymmetry between explanans and explanandum.

The next step is to generalise the tower-shadow scenario. We shall specify the inferential relations on the basis of which our analysis verifies statements of the form 'γ because of α', provided the underlying representation of epistemic states satisfies certain conventions. While this analysis captures the presumed asymmetry for the tower-shadow scenario as well as further classes of explanatory

relations, there remain cases for which the analysis yields symmetric explanations. We therefore conclude with a proposal for a strictly asymmetric Ramsey Test conditional, which in turn yields a strictly asymmetric analysis of 'because'.

The present investigation is very much inspired by the work of Rott (1986) on strengthening the Ramsey Test and his corresponding analysis of 'because'. We will show, however, that Rott's analysis fails to account for the asymmetry of explanatory relations in the case of the tower-shadow scenario and a related class of explanatory relations. This is why we propose an alternative strengthening of the Ramsey Test.

Methodologically, we are working upward from the applications to the formal theory. This strategy seems preferable, for example, when it comes to choosing between different variants of a strengthened Ramsey Test. Moreover, it is worth noting that we take certain intuitions about the propriety of explanatory relations for granted, for instance, the intuition of asymmetry in the tower-shadow example. Our analysis thus aims to capture this and related intuitions about the propriety of explanatory directions.

3.2 Belief Revision Theory

3.2.1 Belief Revision: Basic Ideas

Belief revision theory provides us with a precise semantics of belief changes for the Ramsey Test. Let us therefore very briefly review the basic ideas of this theory. Let K be a set of formulas that represent the beliefs of an agent and α a formula that represents a single belief. In the AGM framework, developed by Alchourrón et al. (1985)), one distinguishes three types of belief change of a belief set K by a formula α:

1. Expansions $K + \alpha$
2. Revisions $K * \alpha$
3. Contractions $K - \alpha$.

An expansion of K by α consists in the addition of a new belief α to the belief set K. This operation is not constrained by any considerations as to whether the new epistemic input α is consistent with the set K of present beliefs. Hence, none of the present beliefs is retracted by an expansion.

A revision of K by α, by contrast, can be described as the *consistent integration* of a new epistemic input α into a belief system K. If α is consistent with K, it holds that $K + A = K * \alpha$. If, however, α is not consistent with K, some of the present beliefs are to be retracted in order to believe α.

A contraction of K by α, finally, consists in retracting a certain formula α from the presently accepted system of beliefs. This operation will be used to define the *suspension of judgement about* α in our strengthened version of the Ramsey Test.

Belief changes can be defined in various ways. In what follows, we explain two approaches to the determination of belief revisions and contractions. First, entrenchment based revisions, which are part of the classical AGM theory. Second, partial meet base revisions, which use the AGM framework but have been developed at a later stage. Before we do so, we briefly introduce the AGM postulates for belief revisions and contractions.[2]

[2]For an accessible exposition of the classical AGM theory, including a detailed discussion of the postulates, the reader is referred to Gärdenfors (1988).

3.2.2 AGM Postulates

AGM Belief Revision Theory aims to specify how a rational agent revises her belief state when accommodating a new belief. In part, the theory was devised to capture the effect on a rational agent's belief state when she learns new evidence. However, according to Gärdenfors (1988, p. 48) a "revision may also be made when you, for the sake of the argument, hypothetically accept a proposition", for example in a Ramsey Test situation.

In AGM Belief Revision Theory, belief states are modeled as belief sets. The intended interpretation of such a set is that it contains those and only those sentences that are accepted or believed to be true by the corresponding agent. A belief set K is a consistent set of sentences that is closed under logical consequence. Formally, let $\vdash$ denote some logical consequence and $Cn(X)$ the corresponding set of logical consequences of a set X. Then, for any belief set K, we have $K \nvdash \bot$ and $K = Cn(K)$. The consequence relation $\vdash$ represents the logic underlying the respective belief revision theory. As such $\vdash$ governs the object language $\mathcal{L}$ of the theory.

We write $\vdash \alpha$ iff the formula α is valid. Let $\rightarrow$ denote material implication, $\bot$ an arbitrary contradiction, and $\top$ an arbitrary tautology. The consequence relation $\vdash$ needs to satisfy certain minimal conditions:

1. If α is a truth-functional tautology, then $\vdash \alpha$.
2. $\dfrac{\vdash \alpha \rightarrow \beta \qquad \vdash \alpha}{\vdash \beta}$ (Modus Ponens)
3. $\vdash$ is consistent, i. e. $\nvdash \bot$.

Note that the logic governing $\mathcal{L}$ is underspecified. However, the logic requires to include at least Boolean propositional logic as given by the minimal conditions. Furthermore, $Cn(K)$ denotes the set $\{\alpha \mid K \vdash \alpha \text{ for } \alpha \in \mathcal{L}\}$. Therefore, the set of logically valid formulas $Cn(\top) = Cn(\emptyset)$ is included in every belief set, and the set of contradictory formulas $Cn(\bot) = K_\bot = \mathcal{L}$ is the unique largest belief set. We follow Gärdenfors in calling the set of all formulas $K_\bot$ the absurd belief set.

Here are the definitions of expansion, contraction, and revision as given by the basic AGM postulates.

Definition 8. Expansion
We say that $K + \alpha$ be the expansion of K by α, and is given by $K + \alpha = Cn(K \cup \{\alpha\})$. An expansion of K by α satisfies

1. $K \subseteq K + \alpha$
2. $\alpha \in K + \alpha$.

Definition 9. Contraction
We say that $K - \alpha$ be a contraction of K with respect to α only if

(K^-1) $K - \alpha$ is a belief set.

(K^-2) If $\vdash \alpha \leftrightarrow \beta$ then $K - \alpha = K - \beta$.

(K^-3) $K - \alpha \subseteq K$.

(K^-4) If $\nvdash \alpha$ then $\alpha \notin K - \alpha$.

(K^-5) If $\alpha \in K$ then $K = (K - \alpha) + \alpha$.

(K^-6) If $\alpha \notin K$ then $K - \alpha = K$.

(K^-5) is called the *recovery* postulate. Later, we will see that it is problematic when we want to model an *asymmetric* because.

Definition 10. Revision
We say that $K * \alpha$ be a revision of K with respect to α only if

(K^*1) $K * \alpha$ is a belief set.

(K^*2) If $\vdash \alpha \leftrightarrow \beta$ then $K * \alpha = K * \beta$.

(K^*3) If $\nvdash \neg\alpha$ then $K * \alpha \neq K_\perp$.

(K^*4) $\alpha \in K * \alpha$.

(K^*5) If $\neg\alpha \notin K$ then $K * \alpha = K + \alpha$.

K^-6 is an implementation of Gärdenfors's 'criterion of informational economy' which demands for contractions that as few beliefs as possible be removed from K. This means that the operation $-$ of belief contraction is in some sense supposed to be a minimal change of K. On logical grounds alone the minimality requirement is underdetermined, i. e. there are several minimal ways to determine which beliefs are to be kept and which are to be given up. The operation $*$ of belief revision has the same problem of underdetermination. For Gärdenfors has proven that (K^-1)–(K^-6) plus the Levi identity is equivalent to (K^*1)–(K^*5) plus the Harper identity.[3] Hence, it is now sensible why we used the indefinite article in the first lines of the definitions, since there is no unique $K - \alpha$ or $K * \alpha$ singled out by (K^-1) – (K^-6) or (K^*1) – (K^*5), respectively. Gärdenfors tries to solve this problem of underdetermination by invoking a relation of epistemic entrenchment.

3.2.3 Entrenchment Based Revisions

If we revise a belief set K by a new belief α that is not consistent with K, some beliefs of K need to be retracted so as to consistently integrate α into the belief system. In other words, we need to retract $\neg\alpha$ first in order to be able to accept α. This idea has been expressed by the Levi identity:

$$K * \alpha = (K - \neg\alpha) + \alpha. \qquad \text{(Levi identity)}$$

The challenge arising here is to find a sensible way of retracting $\neg\alpha$ from K. From a logical point of view, there is no unique solution to this problem, set aside trivial belief revision problems. For, there are several subsets K' of K such that $\neg\alpha \notin K'$. How shall we choose among those subsets?

Belief revision theory tells us that an operation of contraction (as well as that of a revision) should be guided by two principles. First, the 'criterion of informational economy' or *conservativity principle*: when forced to change our beliefs, we should retain as many as possible of the present beliefs. Second, certain beliefs are more firmly established than others. When revising our beliefs, we should maintain the former and be prepared to give up the latter, at least if this is logically possible. The two principles have been been formalised by the theory of entrenchment based revisions.

[3] See Gärdenfors (1988, Appendix A, p. 211 ff.). The Harper identity states that $K - \alpha = K \cap K * \neg\alpha$. We will not use it in this chapter.

Let us begin with the formal characterisation of the epistemic entrenchment relation. $\alpha \leq \beta$ means that α is at most as entrenched as β. The following postulates formally characterise this relation (Gärdenfors, 1988, pp. 89–91):

$$\text{If } \alpha \leq \beta \text{ and } \beta \leq \chi, \text{ then } \alpha \leq \chi \tag{EE1}$$

$$\text{If } \alpha \vdash \beta, \text{ then } \alpha \leq \beta \tag{EE2}$$

$$\alpha \leq \alpha \wedge \beta \text{ or } \beta \leq \alpha \wedge \beta \tag{EE3}$$

$$\text{When } K \neq K_{\perp}, \alpha \notin K \text{ iff } \alpha \leq \beta \text{ for all } \beta \in K \tag{EE4}$$

$$\text{If } \beta \leq \alpha \text{ for all } \beta \in \mathcal{L}, \text{ then } \alpha \in Cn(\emptyset). \tag{EE5}$$

where $\mathcal{L}$ is the set of all formulas of the formal language used to analyse belief changes, and $K_{\perp}$ the absurd belief set containing all elements of $\mathcal{L}$.

Epistemic entrenchment orderings and contractions are interdefinable by (G-) which has been introduced by Gärdenfors and Makinson (1988). In what follows, we shall use only the direction from epistemic entrenchment orderings to contractions:

$$\begin{aligned} &\beta \in K - \alpha \text{ iff } \beta \in K \\ &\qquad \text{and either } \alpha < (\alpha \vee \beta) \text{ or } \alpha \in Cn(\emptyset). \end{aligned} \tag{G-}$$

So, a belief β of K will remain in the belief set after a contraction with α iff either α is strictly less epistemically entrenched than $\alpha \vee \beta$ or α is a logical truth. As one would expect, the strict and equivalence relations of epistemic entrenchment are defined as $\alpha < \beta$ iff $\alpha \leq \beta$ but not $\beta \leq \alpha$, and $\alpha \sim \beta$ iff $\alpha \leq \beta$ and $\beta \leq \alpha$. Once contractions are defined, revisions can be determined using the Levi identity.

The classical AGM theory assumes that belief sets are logically closed. That is, $K = Cn(K)$, where Cn is a consequence operation that satisfies certain standard properties, such as monotonicity, compactness, and the deduction theorem (Hansson, 1999, Ch. 1^{+}). In this chapter, we assume Cn to be given by classical logic. Henceforth, Cn is always used to designate the consequence operation of classical propositional logic.

3.2.4 Belief Bases

The study of belief bases and revisions thereof is intended to achieve a more realistic representation of epistemic states and their dynamics. It can be seen as a cognitively more adequate refinement of classical belief revision theory, which only investigates changes of logically closed belief sets. Why are belief sets felt to be a deficient representation of epistemic states from a cognitive point of view? The problem is that even for languages of propositional logic, any belief set is infinite. This contrasts with the finiteness of human minds and computers. As human minds have only a finite capacity to memorise sentences that are accepted, so have computers only a finite storage.[4]

Unlike belief sets, belief bases are allowed to be finite and are usually assumed to be so. The idea is to have a set H of explicit beliefs that represents all further implicit beliefs in the sense that the latter beliefs are consequences of H. In formal terms:

$$K(H) = Inf(H).$$

[4]The study of belief base changes has been originated by Sven Ove Hansson. Much of what we are going to say about such revisions draws on his *Textbook of Belief Dynamics* (Hansson, 1999).

K contains all beliefs of the agent with a belief base H, i. e. the explicit beliefs and those beliefs that the agent is committed to accept because they are inferable from the explicit beliefs. Inf is an inferential closure operation. We assume that this operation is given by classical logic. Thus, $K(H) = Cn(H)$.

3.2.5 Partial Meet Base Revision

A contraction of a belief base H by α can be defined using the notion of a *remainder set* $H\perp\alpha$:

Definition 11. $H\perp\alpha$ (Hansson (1999, p. 12))
Let H be a set of formulas and α a formula. $H' \in H\perp\alpha$ iff

1. $H' \subseteq H$
2. $\alpha \notin Cn(H')$
3. there is no H'' such that $H' \subset H'' \subseteq H$ and $\alpha \notin Cn(H'')$.

A simple means to define the contraction of H by α is to take the intersection of the members of the remainder set $H\perp\alpha$:

$$H - \alpha = \bigcap H\perp\alpha. \tag{FMBC}$$

This way of defining a contraction is also referred to as *full meet base contraction*.

We can refine this way of determining contractions by invoking the idea of an epistemic ordering among the members of $H - \alpha$. Suppose $\leq$ is a binary transitive relation. $A \leq A'$ means that A' is epistemically not inferior to A. To put it more simply, $A \leq A'$ means that A' is epistemically at least as good as A. Using such an epistemic ordering, we can define a selection function for the remainder set as follows:

$$\sigma(H\perp\alpha) = \{H' \in H\perp\alpha \mid H'' \leq H' \text{ for all } H'' \in H\perp\alpha\}. \tag{Def σ}$$

Then, we take the selected members of the remainder set to define the contraction of H by α:

$$H - \alpha = \bigcap \sigma(H\perp\alpha). \tag{PMBC}$$

It remains to explain the expansion of a belief base H by α, which is straightforward:

$$H + \alpha = H \cup \{\alpha\}. \tag{H + α}$$

Now we are in a position to put everything together, thus defining partial meet base revisions:

$$H * \alpha = \bigcap \sigma(H\perp\neg\alpha) + \alpha. \tag{PMBR}$$

3.2.6 Prioritised Belief Bases

While the idea of an epistemic ordering of beliefs is quite plausible, it is far from clear how to order the subsets of a set of beliefs. This does not matter for studying the formal properties of belief changes, but it does so for studying concrete examples. Hence, we finally show how an epistemic ordering among the members of a belief base can be translated into an ordering among the subsets of such a base.

Drawing on the work by Brewka (1991), we assume the epistemic ordering among the items of H to be a strict weak ordering. Such an ordering can be represented by a sequence of subsets of H:

$$\mathbf{H} = \langle H_1, \ldots, H_n \rangle.$$

where $H_1, \ldots, H_n$ is a partition of H. $\mathbf{H}$ is called a *prioritised belief base*. $H_1, \ldots, H_n$ are sets of formulas that represent explicit beliefs, and the indices represent an epistemic ranking of the beliefs. H_1 is the set of the most firmly established beliefs, the beliefs in H_2 have secondary priority, etc.

This prioritisation of beliefs can be used to define an epistemic ordering among the subsets of H:[5]

Definition 12. $H'' \leq H'$
Let H be a set of formulas, and H'' and H' be subsets of H. $H'' \leq H'$ iff there is no i $(1 \leq i \leq n)$ such that

1. $H' \cap H_i \subset H'' \cap H_i$
2. for all $j < i$ $(j \geq 1)$, $H'' \cap H_j = H' \cap H_j$.

In the following investigation, we assume that our belief base has exactly two levels of epistemic priority: the upper level, containing the generalisations, and the lower level, which contains our beliefs about atomic facts. These levels of epistemic priority affect the determination of belief changes: when we retract a belief α, we retract first beliefs about atomic facts before we retract generalisations. If necessary, we also retract generalisations, but only if the retraction of α cannot be achieved by retractions of beliefs about atomic facts. For the considerations to follow, it may be helpful to have a graphical representation of such a prioritised belief base in mind:

G
L

G stands for the set of generalisations, while L contains the beliefs about the atomic facts. L is a set of literals. A literal is an atomic formula or its negation.

If we need to distinguish between strict and *ceteris paribus* laws, we can do so by distinguishing between two corresponding levels of generalisations. Strict laws have priority over *ceteris paribus* laws. The notion of a generalisation subsumes strict and non-strict laws.

The present convention about generalisations and literals allow us to define the revision of epistemic states (in addition to the revision of belief sets of epistemic states). That is,

$$(H, <) * \alpha$$

has a well defined meaning insofar as Definition 12, (Def σ), and (PMBR) together define the revised belief base H', while the revised epistemic ordering $<'$ is determined by the simple convention that generalisations have priority over literals. So there is an epistemic state $(H', <')$ such that $(H', <') = (H, <) * \alpha$. Iterated belief base revisions are thus well understood.

Such are the basic ideas and definitions about belief changes that will be used in the present analysis of 'because'. We study the properties of belief set revisions using an epistemic entrenchment ordering of beliefs. Belief base revisions are studied in terms of partial meet base revisions with an underlying selection function that is defined by a prioritised belief base.

Why do we not study belief set revisions in terms of partial meet belief set revisions? The simple reason for this choice is that the idea of an epistemic entrenchment ordering is easier applicable than the idea of an epistemic ordering of subsets of a logically closed (and so infinite) belief set. Despite the differences between partial meet belief set revisions and entrechment based belief set revisions, it has been shown that any entrenchment based belief set revision can be represented by a partial meet belief set revision, and vice versa (Gärdenfors, 1988, Ch. 4). Hence, results about the former can be translated into results about the latter.

[5]This definition is inspired by Brewka (1991), but the resulting belief revision operation is not equivalent with the one defined there.

3.2.7 Why Belief Bases?

In the final analysis of 'because', we shall use belief base revisions rather than belief set revisions. For, the combination of belief bases with our novel variant of a strengthened Ramsey Test allows us to capture the asymmetry of explanatory relation for a large class of scenarios, including the famous tower-shadow scenario. Belief set revisions of the classical AGM theory, by contrast, turn out not to be suited for this purpose (cf. sections 3.5 and 3.6).

Admittedly, belief base revision theory is less well established than belief set revision theory. This is surprising in light of distinctive merits of belief bases if compared with belief sets. First, there is the above indicated finite-memory argument in favour of belief bases. A belief set is an infinite entity and so cannot be fully comprehended by a human mind, at least on a literal understanding of comprehension. Likewise, a computer cannot store a belief set for obvious reasons.

Second, relatedly, while the study of formal properties is not much impeded by the infinite character of belief sets, the study of concrete examples certainly is. Note that even the notation $K = Cn(\alpha, \beta, \gamma)$ is misleading if K is supposed to be a belief set. For, it suggests that the belief set K is generated by the belief base $\{\alpha, \beta, \gamma\}$. This is misleading because belief set revisions differ from belief base generated revisions as regards their formal properties (cf. Hansson (1999, Ch. 4)). Moreover, it is quite a bit more work to completely specify the epistemic priority ordering for a belief set than it is for a belief base. As a consequence of this, iterated belief changes are easier to define and to describe using belief bases (cf. Section 3.2.6). It is therefore not surprising that, in the belief set revision literature, examples of concrete belief changes are hardly formalised. It is next to impossible to find a fully formalised application of belief set revision theory, even to toy examples. Using belief bases, by contrast, makes applying belief revision theory to particular examples much easier.

In sum, belief bases are cognitively more plausible and much easier to use when it comes to formalising belief systems that concern specific examples. However, there is also an influential objection to belief bases. This objection appeals to the *principle of the irrelevance of syntax* (Dalal, 1988), which is sometimes violated in the belief base approach. Contrary to this principle, Brewka (1991) has pointed out that choosing the formulation $\{p \wedge q\}$ over $\{p, q\}$ may well be intended to make a difference. This choice is justified, for example, if p and q are only to be given up together.

We are not convinced that the principle of the irrelevance of syntax is justified from a cognitive perspective. Contrary to this principle, one can point out that the study of belief bases carries on what Van Benthem (2008) has termed the *cognitive turn* in philosophical logic, i. e. the development of logical systems that aim to represent and to theoretically explain human reasoning. For the above indicated reasons, a finitely bounded human mind has no alternative to working with belief bases. Note, finally, that there is a very simple way to respect the principle of the irrelevance of syntax within the belief base approach. It suffices to require that the members of a belief base conform to a specific logical form. For example, we can require to represent generalisations by disjunctions of literals, while beliefs about atomic facts be represented by literals. The latter requirement has already been made explicit.

3.3 The Ramsey Test

3.3.1 The Ramsey Test by Ramsey

Ramsey (1950, footnote 1) proposes the following evaluation procedure for conditionals that is known as the Ramsey Test (RT):

> If two people are arguing 'If p will q?' and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q; so that in a sense 'If p, q' and 'If p, $\overline{q}$' are contradictories. [...] If either party believes *not* p for certain, the question ceases to mean anything to him except as a question about what follows from certain laws or hypotheses.

The RT is an epistemic evaluation recipe for conditionals in the sense that the evaluation depends on the beliefs of the agent(s) involved in the hypothetical discussion. This evaluation recipe for an epistemic agent has been pointedly expressed by Stalnaker (1968, p. 102):

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true.

3.3.2 The Ramsey Tests by Gärdenfors and Levi

The AGM theory allows Gärdenfors (1988, Ch. 7) to concisely formalise the epistemic recipe of the Ramsey Test:

$$\alpha > \gamma \in K \text{ iff } \gamma \in K * \alpha. \tag{RT$_G$}$$

Thus, a conditional $\alpha > \gamma$ is accepted in K iff γ is believed in the course of revising K by α.

Writing $K * \alpha$ and speaking of belief set revisions is misleading insofar as this suggests that it is the belief set itself that is revised. This is not quite correct because there is no sensible way of uniquely determining the revision of a belief set by a new epistemic input. It is rather the belief set of a particular epistemic state that is revised, according to the AGM theory. As indicated in the previous section, epistemic states can be represented in various ways. Syntactic representation schemes commonly have the form of a pair $(A, <)$, where A is a set of formulas and $<$ an epistemic ordering among formulas or sets of formulas. A is logically closed for belief set revision schemes, while it does not have to be so for belief base revision schemes. Most possible world approaches to belief revision work with epistemic states of the form $(W, <)$, where W is a set of possible worlds and $<$ a plausibility ordering among these worlds (cf. (Grove, 1988)).

It is thus more appropriate to write $K(S) * \alpha$ or $K(A, <) * \alpha$ and to speak of the revision of the belief set of an epistemic state. In this notation, S stands for an epistemic state. As regards the Ramsey Test, it seems consequently more appropriate to write:

$$\alpha > \gamma \in K_>(S) \text{iff } \gamma \in K(S) * \alpha. \tag{RT$_L$}$$

$K(S)$ stands for the beliefs in non-modal propositions, i. e. beliefs that can be expressed by formulas of classical logic without any conditional or modal operator. $K_>$, by contrast, stands for the conditionals accepted, or believed, by the agent on the basis of the Ramsey Test.

In spirit, the distinction between $K(S)$ and $K_>(S)$ goes back to Levi (1988). It was also Levi (1988) who emphasised another distinction, viz. between believing and merely accepting conditionals, in light of a famous triviality theorem proved by Gärdenfors (1986). If we merely accept conditionals without viewing them as truth-apt, we avoid the fatal consequences of the triviality theorem for the Ramsey Test.[6]

[6]The triviality theorem continues to provoke lively research in belief revision theory (see, e.g., Rott (2011); Leitgeb (2010); Bradley (2007)).

We shall discuss triviality briefly in Section 3.6.4. There, it will be shown that an important premise of the proof by Gärdenfors (1986) is violated for our variant of a strengthened Ramsey Test. This allows us to remain neutral as to whether conditionals are properly believed or merely accepted. As (RT_L) gives us a clearer instruction of how to carry out a particular Ramsey Test, we prefer (RT_L) over (RT_G) as formulation of the Ramsey Test. A similar strategy has been recommended by Hansson (1992) who shows that triviality can be avoided by taking belief bases to represent the epistemic states underlying the Ramsey Test. Following Hansson (1992), we favour (RT_L) without making a commitment as regards the belief-acceptance distinction for conditionals.

3.3.3 Absurdity: Relevance Issues of the Ramsey Test

(RT_G) leads to the absurdity that any two accepted formulas α, β bear a conditional relation between each other, as has been shown by Rott (1986). Suppose $\alpha, \beta \in K$. By the AGM belief revision postulates, we know that, if $\alpha \in K$, then $K * \alpha = K$. Therefore, $\beta \in K * \alpha$. We conclude by (RT_G) that $\alpha > \beta \in K$. Hence,

$$\text{If } \alpha, \beta \in K, \text{ then } \alpha > \beta \in K. \qquad \text{(Absurdity)}$$

(Absurdity) expresses that a (RT_G) agent accepts a conditional connection between any two formulas she accepts. If, for example, 'Munich is a town in Germany' and 'Lund is a town in Sweden' is accepted by an agent, then (RT_G) prescribes that 'If Munich is a town in Germany, then Lund is a town in Sweden' should also be accepted.

The just observed problem carries over to a Ramsey Test analysis of 'because' as proposed by Ramsey (1950, p. 156, our emphasis) himself; there he relates conditional sentences 'If α, then γ' and sentences 'Because α, γ' by stating:

> *because* is merely a variant on *if*, when [the antecedent] p is known to be true.

It is a consequence of this view that, if $\alpha \in K$, then 'if' and 'because' coincide. Thus, given Ramsey's view, (RT_G) prescribes that our agent accepts the sentences 'Because Munich is a town in Germany, Lund is a town in Sweden', and the converse 'Because Lund is a town in Sweden, Munich is a town in Germany'. In more general terms, (Absurdity) entails that a (RT_G) agent accepts any because sentence composed of any two accepted formulas, once Ramsey's analysis of 'because' is adopted. This is in any case not less troubling than the (Absurdity) of the merely conditional reading. Moreover, the absurdity violates the asymmetry of one usage of because in natural language, viz. the one according to which the acceptance of 'because α, β' precludes the acceptance of 'because β, α', at least for some α, β. We might, for instance, accept that we sometimes get injured because we often play football, but then we would not accept that we often play football because we get sometimes injured.[7]

In sum, (RT_G) fails to capture the semantics of indicative conditionals. This failure leads to verifying absurd explanatory relations if we accept Ramsey's analysis of 'because' in terms of the Ramsey Test. The underlying problem is that the conditional connective $>$ does not express a proper relation of relevance between the antecedent and the consequent. Rott (1986) proposes to invalidate (Absurdity) by modifying (RT_G) such that the mere acceptance of α, γ does not result in the acceptance of 'If α, γ', as we shall see in the next section.

[7] Although the latter because sentence seems totally fine in a context, in which the agent performs a so-called *inference to the best explanation*: repeatedly playing football may be the best explanation for occasional injuries. This reasoning *towards* (as opposed to *from*) the putative explanatory 'causes' seems to justify the usage of 'because' in the other direction. So peculiar as natural language is, we do not want to ban this usage of 'because' from natural language. For now, we just want to focus on the one usage showing the asymmetry without being entirely sure that this usage is strictly asymmetric. For example, 'because p and q are true, $p \wedge q$ is true' does not seem to preclude 'because $p \wedge q$ is true, p and q are true'.

3.4 Rott's Ramsey Test Analysis of 'Because'

Rott (1986) embeds a parallel analysis of 'if' and 'because' in a systematic theory of universal conditionals. Universal conditionals, Rott claims, are not instantiated in natural language. However, he proposes a semantics of the natural language conjunctions 'if', 'if ... might', 'because', 'though' and 'even if' by specifying constraints on the acceptance of 'antecedents' and 'consequents' of the respective universal conditionals.[8]

Rott aims to analyse the type of 'because' that points to a reason or an explanation. The basic idea of Rott's analysis is that 'Because α, γ' be synonymous to 'α is a reason or an explanation for γ'. In an explanatory sentence, for example, 'because' may be seen as a connective that relates explanans and explanandum. As such a pointer 'because α, γ' expresses a relation of positive relevance between (explanans) α and (explanandum) γ. However, we have seen in the last section that (RT_G) does not capture a proper conditional connection of (positive) relevance between antecedent and consequent.

Rott's analysis of 'because' is driven by considerations of how to establish a relation of positive relevance, and this means for a start to find ways to invalidate (Absurdity). One such way consists in (i) the modification of (RT_G) to his Strong Ramsey Test; another way in (ii) the contraction of the belief set by the consequent, before the set is revised by the respective antecedent. The implementation of (i) and (ii) in Rott's analysis gives rise to the scheme of universal pro-conditionals. This scheme allows him to derive a semantics of the indicative and subjunctive ifs and of a certain 'because' of natural language.[9]

3.4.1 The Strong Ramsey Test and the Contraction of the Belief Set by the Consequent

In Section 3.3.3, we have seen that (RT_G) together with Ramsey's analysis of 'because' fails since this analysis validates (Absurdity). This failure requires a modification of (RT_G) that invalidates (Absurdity). Rott's idea is to strengthen (RT_G), which results in his 'Strong Ramsey Test':

$$\alpha \ggg \gamma \in K \text{ iff } \gamma \in K * \alpha \text{ and } \gamma \notin K * \neg\alpha. \qquad (\mathrm{SRT}_R)$$

We obtain in the situation where a (SRT_R) agent already accepts α, γ:

$$\text{If } \alpha, \gamma \in K, \text{ then } [\alpha \ggg \gamma \in K \text{ iff } \gamma \notin K * \neg\alpha]. \qquad (1)$$

Implication (1) shows how (SRT_R) invalidates the (Absurdity) that the mere acceptance of α and γ is sufficient for the acceptance of 'Because α, γ'. The reason is that the second conjunct of the right-hand-side of (SRT_R) still needs to be satisfied. This modification makes (SRT_R) "more adequate for natural language conditionals than" Gärdenfors's Ramsey Test, so Rott argues, since "it explicitly requires the antecedent to be positively relevant for the consequent" (Rott, 1986, p. 352).[10]

The situation represented by implication (1) requires the (SRT_R) agent to perform a contrary-to-fact supposition. The counterfactual supposition of $\neg\alpha$ needs to retract the accepted γ from the belief

[8] 'Antecedent' is here a generalisation of the antecedent of a conditional sentence. It stands for 'subordinate clause' of the respective sentence. This mirrors Rott's view that all of the mentioned conjunctions are derived from a framework of universal conditionals. In detail, the indicative and subjunctive ifs and 'because' fall into the category of universal pro-conditionals, 'though' into the category of universal contra-conditionals, and 'even if' into the category of universal un-conditionals. See Rott (1986, pp. 355-363).

[9] Considerations of how to systematically categorise conditionals result in the schemes of universal contra- and un-conditionals as well.

[10] This idea has recently been exploited in an analysis of evidential support by Chandler (2013).

set. In this sense the contrary-to-fact supposition 'makes a difference' as to whether γ is accepted.[11]

An alternative method to invalidate (Absurdity) consists in the contraction of the belief set by the consequent before (RT_G) is applied.

$$\alpha > \gamma \in K \text{ iff } \gamma \in (K - \gamma) * \alpha. \tag{RT_G^-}$$

In the situation where a (RT_G^-) agent already accepts α, γ, the consequent γ may not be in the belief set K after a contraction by γ and a subsequent revision by α, i. e. $\gamma \notin (K - \gamma) * \alpha$. The consequent γ is only accepted if it is a consequence of the contracted belief set $(K - \gamma)$ revised by the antecedent α. We may say that (the belief expressed by) α is an inferential epistemic reason for γ such that the supposition of α epistemically brings about the acceptance of γ.

3.4.2 Universal Pro-Conditionals and 'Because'

If we amend the (SRT_R) and its dual by (RT_G^-), we obtain the scheme of universal pro-conditionals that is according to Rott (1986, p. 355) "perfect for the analysis of what you can call a 'conditional connection'". We write $\Rightarrow$ for the universal pro-conditional. The scheme is then given by

$$\begin{aligned} \alpha \Rightarrow \gamma \in K \text{ iff } & [\alpha \ggg \gamma \in (K - \gamma)] \text{ or } [\neg\alpha \ggg \neg\gamma \in (K - \gamma)] \\ \text{iff } & [\gamma \in (K - \gamma) * \alpha \text{ and } \gamma \notin (K - \gamma) * \neg\alpha] \\ \text{or } & [\neg\gamma \notin (K - \gamma) * \alpha \text{ and } \neg\gamma \in (K - \gamma) * \neg\alpha]. \end{aligned} \tag{UPC}$$

(UPC) says that there is a conditional connection (of positive relevance) between antecedent α and consequent γ iff (i) α lets us infer γ in the context of $K - \gamma$, and the supposition of $\neg\alpha$ makes a difference as to whether γ is accepted, *or* (ii) $\neg\alpha$ lets us infer $\neg\gamma$ in the context of $K - \gamma$, and the supposition of α makes a difference as to whether $\neg\gamma$ is accepted.

Rott derives the natural language ifs and 'because' from the scheme (UPC) by specifying acceptance constraints on antecedent and consequent of the respective connective. For universal pro-conditionals the acceptance constraint is that the acceptance status of the antecedent and the consequent is the same. In accordance with Ramsey's view, 'because α, γ' is only accepted if the antecedent α is accepted. Let ${}^{a}\Rightarrow$ be the connective of Rott's 'because', where the superscript ^{a} indicates that the antecedent is accepted. Then Rott's analysis of the natural language 'because' is given by

$$\begin{aligned} \alpha \, {}^{a}\Rightarrow \gamma \in K \text{ iff } & \alpha \Rightarrow \gamma \in K \text{ and } \alpha, \gamma \in K \\ \text{iff } & \gamma \in (K - \gamma) * \alpha \text{ or } \neg\gamma \in (K - \gamma) * \neg\alpha \\ & \text{and } \alpha, \gamma \in K. \end{aligned} \tag{Because$_R$}$$

(Because$_R$) can be derived from (UPC) using a proposition in Rott (1986, p. 350):

Proposition 1. If $\gamma \in K * \alpha$ and $\gamma \in K * \neg\alpha$ then $\gamma \in K$.

Proposition 2. Let K be a non-absurd belief set and γ a non-tautology. Then (UPC) and $\alpha, \gamma \in K$ implies (Because$_R$).

Proposition 2 and all subsequent propositions are proven in Appendix B.

[11] (SRT_R) structurally resembles Lewis (1973c)'s notion of causal dependence in terms of counterfactual conditionals. Using $\Rightarrow$ for causal dependence, we can transcribe Lewis's idea into the notation of belief revision: $\alpha \Rightarrow \gamma \in K$ iff $\gamma \in K * \alpha$ and $\neg\gamma \in K * \neg\alpha$. Note that Lewis's causal dependence requires a stronger version of difference making than (SRT_R), viz. the adoption of $\neg\gamma$ in $K * \neg\alpha$ in contrast to the mere retraction of γ. Moreover, Lewis might say that 'γ because α' means γ is causally dependent on α, when α and γ are (believed to be) true. Given Lewis (1973a)'s semantics for counterfactuals, we obtain the following implication paralleling (1): If $\alpha, \gamma \in K$, then $[\alpha \Rightarrow \gamma \in K \text{ iff } \neg\gamma \in K * \neg\alpha]$.

3.5 Symmetry Problems of Rott's 'Because'

3.5.1 A General Symmetry Problem

We can show now that (Because$_R$) is symmetric for a large class of potentially explanatory relations. For this to be achieved, we distinguish between trivial and non-trivial implications in K. As the members of K are non-modal propositional formulas, laws and generalisations are to be represented by material implications. However, not all implications in a belief set K represent instances of generalisations. As is well known, if $\neg\alpha \in K$, then, for any γ, $\alpha \to \gamma \in K$. Likewise, if $\gamma \in K$, then, for any α, $\alpha \to \gamma \in K$. How can we distinguish, then, between trivial implications and non-trivial implications in K, on the understanding that only the latter represent instances of genuine generalisations? Arguably, a material implication is non-trivial in a belief set K iff it 'survives' a contraction by the negation of the antecedent and a contraction by the consequent:

Definition 13. Non-trivial implication in K
K contains an implication $\alpha \to \gamma$ non-trivially iff

1. $\alpha \to \gamma \in K$, and
2. $\alpha \to \gamma \in K - \neg\alpha$, and
3. $\alpha \to \gamma \in K - \gamma$.

Now, if non-trivial implications represent instances of generalisations, it is reasonable to assume that non-trivial implications are more entrenched than literals and conjunctions of literals.[12] For, this assumption guarantees that generalisations are available for counterfactual considerations. To see this consider the following: suppose $\neg\alpha$, $\alpha \to \gamma \in K$, where $\alpha \to \gamma$ is a non-trivial implication, such as 'if it snows on the street, the street gets white'. Further, suppose that $\neg\alpha < \alpha \to \gamma$, where $\alpha < \beta$ means that β is strictly more entrenched than α. By (G-) and the entrenchment postulate (EE2), it holds then that $\alpha \to \gamma \in K - \neg\alpha$. By the Levi identity, this implies that $\gamma \in K * \alpha$, as it should be. The street would get white if it were to snow on the street.

If, by contrast, $\alpha \to \gamma \leq \neg\alpha$, we would have (i) $\neg\alpha \vee \gamma \leq \neg\alpha$. By (EE2), however, we know that (ii) $\neg\alpha \leq \neg\alpha \vee \gamma$. Using $\vdash (\alpha \to \gamma) \leftrightarrow (\neg\alpha \vee \gamma)$, we can infer from (i), (ii), and (G-) that $\alpha \to \gamma \notin K - \neg\alpha$. Hence, our generalisation represented by $\alpha \to \gamma$ would not be available for counterfactual considerations on the hypothetical assumption of α. We could not infer that the street would get white if it were to snow on the street. It goes without saying that this result is highly counterintuitive.

We must wonder, finally, whether or not a non-trivial implication $\alpha \to \gamma$ can be retracted in the course of a revision by a proposition β that is neither related to the antecedent α nor to the consequent γ. Such a retraction does not seem reasonable at all, even though there may be an entrenchment ordering that requires it. For example, we should not retract 'if it snows on the street, the street gets white' if we get to know that Munich is a town in Germany, Anna goes to the party, etc. The case study in sections 3.5.2 and 3.6.3 will further support the claim that implications representing generalisations must be more entrenched than literals and conjunctions of literals.

Hence, it is reasonable to make the following assumption:

Assumption 1. Let $\alpha \to \gamma$ be a non-trivial implication in K. Let δ, β be literals or conjunctions of literals. $<_K$ denotes the entrenchment ordering associated with the beliefs of K and $<_{K-\delta}$ the entrenchment ordering of the beliefs of $K - \delta$. Then, $\beta <_K \alpha \to \gamma$ and $\beta <_{K-\delta} \alpha \to \gamma$.

[12] See Hansson (1999, p. 96) for a brief justification of why law-like statements should – in most cases – be epistemically more entrenched than factual statements.

The second conjunct of this assumption says that the implication $\alpha \rightarrow \gamma$ remains more entrenched than literals and conjunctions thereof, after a contraction of K by a literal or a conjunction of literals.

We can show that ${}^{a}\Rightarrow$ is symmetric, if $\alpha \rightarrow \gamma$ is an accepted non-trivial implication.

Proposition 3. Let α and γ be literals or conjunctions of literals. Further, $\alpha, \gamma \in K$. Suppose that $\alpha \rightarrow \gamma$ is a non-trivial implication in K and Assumption 1 holds for $\alpha \rightarrow \gamma$. Then $\alpha \, {}^{a}\Rightarrow \gamma \in K$ and $\gamma \, {}^{a}\Rightarrow \alpha \in K$.

Proposition 3 says that, under Assumption 1, if $\alpha \rightarrow \gamma$ is a non-trivial implication in K, then ${}^{a}\Rightarrow$ is symmetric since both $\alpha \, {}^{a}\Rightarrow \gamma \in K$ and $\gamma \, {}^{a}\Rightarrow \alpha \in K$. In particular, we have the following problem for Rott's ${}^{a}\Rightarrow$: if $\gamma \rightarrow \alpha$ is a non-trivial implication in K, then $\alpha \, {}^{a}\Rightarrow \gamma \in K$. This means, for instance, if an agent accepts the sensible generalisation 'if there is lightning, then there is thunder', the agent is also committed to accept 'because there is thunder, there is lightning'. Of course, from an information-theoretic point of view, it is sensible to say that we *believe* there is lightning, because we *believe* that there is thunder. However, we are interested in the asymmetric usage of 'because' that goes beyond a purely information-theoretic relation. To be more precise, we aim to define an epistemic relation of bringing about according to which we *accept*, for instance, that lightning brings about thunder, but we should not *accept* the converse.

3.5.2 Further Symmetry Problems

We have seen that the notion of because implemented in ${}^{a}\Rightarrow$ is symmetric on some reasonable assumptions. However, as we shall see shortly, the scope of Proposition 3 is limited. By means of a simple example scenario, we lift the limitation by showing that ${}^{a}\Rightarrow$ does not capture intuitively asymmetric relations of relevance. The scenario illustrates the symmetry of ${}^{a}\Rightarrow$ over and above Proposition 3. It may be understood as revealing further symmetry problems for (Because$_R$) by characterising another class of problematic applications. The characterisation of a class means here that once the reader understands the underlying structure of the example scenario, she may easily come up with her own examples of the problematic class.

Suppose there is a tower (t), the sun is shining (s), so that the sun casts a shadow (sh).[13] (t, s, and sh are propositional constants to be used for the below formalisation of the example.) Intuitively, the presence of the tower and the sunlight *explain* that there is a shadow, but not vice versa, i.e. there being a shadow does not explain that there is a tower. After all, there might be, for instance, another opaque object exposed to sunlight. However, it seems that the following common-sense generalisation is entailed by our background knowledge:

$$t \wedge s \rightarrow sh. \tag{2}$$

We assume in our scenario that (2) is epistemically more entrenched in the agent's background knowledge than the facts t, s, and sh. Let us assume, moreover, that $K = Cn(\{t, s, sh, t \wedge s \rightarrow sh\})$. Note that (2) plays the role of a non-trivial implication that remains more entrenched than any literals or conjunctions of literals after the contraction of the epistemic state by some literals or conjunctions thereof. Assumption 1 is thereby satisfied, and thus (2) constitutes a special case of Proposition 3 according to which a non-trivial implication may be of the form $\alpha_1 \wedge \alpha_2 \wedge ... \wedge \alpha_n \rightarrow \gamma$ for a finite $n \in \mathbb{N}$,

[13] The example is similar to the famous tower-shadow scenario, for which there is wide agreement that the height of the tower together with the altitude of the sun explain the length of the shadow, but not vice versa. However, see Van Fraassen (1980b, pp. 132-34) for an interesting challenge of this agreement involving the notion of relevance. Note that we simplified the original tower-shadow scenario such that a wider class of examples succumbs to the asymmetry problem of Rott's (Because$_R$).

where the α_i are literals. If such a non-trivial implication with a conjunction of antecedent conditions is accepted, we obtain $\alpha_1 \wedge \alpha_2 \wedge ... \wedge \alpha_n \stackrel{a}{\Rightarrow} \gamma$ as well as $\gamma \stackrel{a}{\Rightarrow} \alpha_1 \wedge \alpha_2 \wedge ... \wedge \alpha_n$. In particular, suppose the non-trivial implication (2) with $n = 2$. Then a (Because$_R$) agent is committed to both the plausible direction that 'There is a shadow, because there is a tower and the sun is shining' and the less plausible direction that 'There is a tower and the sun is shining, because there is a shadow'.

Following our cognitive habits, we do not list all of the antecedent conditions when using 'because', especially if there are many. Rather we only state the pertinent ones given a particular contextual knowledge. If this is true, the scope of Proposition 3 is limited. For, this proposition does not tell us whether $\alpha_1 \wedge \alpha_2 \wedge ... \wedge \alpha_n \rightarrow \gamma$ being a non-trivial implication entails that $\alpha_1 \wedge \alpha_2 \wedge ... \wedge \alpha_{n-k} \stackrel{a}{\Rightarrow} \gamma$ and $\gamma \stackrel{a}{\Rightarrow} \alpha_1 \wedge \alpha_2 \wedge ... \wedge \alpha_{n-l}$, where $1 \leq k, l < n$. In what follows we show that further symmetry problems arise for non-trivial implications of the logical form $\alpha_1 \wedge \alpha_2 \rightarrow \gamma$, and because sentences of the form $\alpha_1 \stackrel{a}{\Rightarrow} \gamma$ and $\gamma \stackrel{a}{\Rightarrow} \alpha_1$, where α_i and γ are literals, respectively. We illustrate this further class of symmetry problems by proving the following proposition about the tower-shadow scenario, in which an antecedent condition of a generalisation remains implicit in the background knowledge, viz. 'the sun is shining'. Thereby we show that $\stackrel{a}{\Rightarrow}$ may express a symmetric relation even if Proposition 3 is not applicable.

Proposition 4. Assume a (Because$_R$) agent accepts all facts and the generalisation of the tower-shadow scenario, i. e. $t, s, sh, t \wedge s \rightarrow sh \in K$, where the order of epistemic entrenchment is $t, s, sh < t \wedge s \rightarrow sh$. Then, $t \stackrel{a}{\Rightarrow} sh \in K$ if $t \leq sh$ and $sh \stackrel{a}{\Rightarrow} t \in K$ if $sh \leq t$.

The proposition shows that Rott's analysis only verifies the desired direction 'because of the tower there is a shadow', if 'there is a tower' is at most as entrenched as 'there is a shadow', and thus both beliefs are given up when the belief set is contracted by 'there is a shadow'. Moreover, a (Because$_R$) agent is committed to believe the undesired direction 'because of the shadow there is a tower', if 'there is a shadow' is at most as entrenched as 'there is a tower'. The question is, of course, why should some of those atomic beliefs be more or less entrenched? It seems to be an ad hoc strategy to assume the entrenchment ordering that verifies the intended results. Why should the belief that there is a tower be strictly less entrenched than the belief that there is a shadow?

Even if we grant that $t < sh$ in order to obtain the desired asymmetry, a change in the epistemic entrenchment ordering could change the set of accepted because statements. There is a general problem surfacing here: a given epistemic entrenchment ordering alone cannot do both (i) represent firmness of belief and, somewhat independently, (ii) encode explanatory relations of because sentences. In other words, either an ordering of epistemic entrenchment is used for the representation of epistemic firmness, or it is used to obtain the desired asymmetry – but you do not get both from one ordering. To be clear, it seems too much of a demand for one ordering of epistemic entrenchment that it can satisfy the role of representing epistemic firmness and ensure the relatively stable asymmetry of because. This trade-off between the representation of firmness and asymmetry seems to apply to any account that directly and *only* relies on an ordering of epistemic entrenchment. It is by no means obvious that there should not be a certain independence between the firmness of belief and the acceptance of explanatory because statements.

Our analysis, in contrast, has no such sensitivity to the firmness of particular beliefs. The firmness of particular beliefs can change without changing the direction of accepted because sentences, at least when the relevant generalisations are in place. Of course, our analysis hinges on the assumption which generalisations the agent accepts. But the generalisations also allow us to vary the firmness of particular beliefs, while preserving the asymmetry of because.

The underlying structure of the tower-shadow scenario illustrates the tendency of (Because$_R$) to express a symmetric relation. In addition to the symmetry shown by Proposition 3, the scenario

illustrates why $^{a}\Rightarrow$ is prone to symmetry problems, even if certain antecedent conditions are not explicitly stated. The tower-shadow scenario thus lets us recognize symmetry problems for (Because$_R$) even beyond the assumptions of Proposition 3.

Rott seems to be aware of the symmetry problems. He writes in Rott (1986, p. 347):

> [If] someone insists that *because* is positively about an asymmetric causal relation in the world, I have to confess that I cannot give a satisfactory interpretation of this 'causal' *because*. I shall concentrate on the 'informative' *because* specifying just reasons. Yet I conjecture that this 'informative' *because* is the more common and the more general one, and that the 'causal' *because* can eventually be characterised as a special case by a few non-epistemic conditions.

The quote is interesting in at least three respects. (i) Rott distinguishes between several interpretations of the word 'because' that correspond to different usages of the word. He calls the usage that points to reasons 'informative', and the usage that expresses an 'asymmetric causal relation' 'causal'. (ii) Rott assumes that a concept of causation needs to satisfy 'a few non-epistemic conditions'. Hence, he writes that a causal relation be 'in the world', although he outlines a purely epistemic account. (iii) Rott conjectures that the ontological 'causal' usage is derivative as a special case from the epistemic 'informative' usage.

We agree that the word 'because', like many natural language connectives, has several interpretations, and thus is used in a variety of ways. As for the non-epistemic conditions that a concept of causation may have to satisfy, the requirement that a cause precedes its effect seems still promising. As is well known, this requirement is central to both Hume's and Ramsey's account of causation.[14] Moreover, the requirement of temporal precedence has been adopted by Spohn (2006) in his ranking-theoretic elaboration of the basic Humean idea about causation. We shall not further pursue this line here, but confine ourselves to finding an epistemic interpretation of 'because' that is asymmetric. We leave it open, however, whether or not this interpretation deserves to be called 'causal'. In doing so, we bracket the topic of causation for now and return to it in the next chapters.

Recall from Section 3.4 Rott's basic idea that the 'informative' usage of 'because' expresses a reason or an explanation. But then Rott's analysis runs into the following difficulty: (Because$_R$) fails to capture the asymmetry associated with some explanations, as we have seen in the tower-shadow scenario. In other words, if an explanation should intuitively be asymmetric, then (Because$_R$) is too permissive as it allows for intuitively incorrect converse explanations. We consider this difficulty to call for a complementation of Rott's merely 'informative' because with an epistemically asymmetric because.

3.5.3 Using Belief Bases

So far the symmetry of Rott's 'because' has been characterised within the original AGM belief revision theory, assuming validity of all *AGM postulates* as established in Gärdenfors (1988, Ch. 3). This accords with the framework assumed in Rott's analysis. Now we switch from the original AGM theory, which employs belief sets, to belief revision using belief bases (as outlined in sections 3.2.6-3.2.4) and Levi's formulation of the Ramsey Test.

[14] "A cause is an object precedent and contiguous to another, and so united with it, that the idea of the one determines the mind to form the idea of the other, and the impression of the one to form a more lively idea of the other." (Hume (1739/1978, p. 170)) One might wonder whether the temporal order of cause and effect is a properly non-epistemic condition in the context of Hume's work, but this is a question that need not concern us here. See also Ramsey (1929/1990).

As regards revisions and contractions, belief bases behave somewhat differently if compared to belief sets.[15] Notably, recovery (K^-5), which is needed in the proof of Proposition 4, is not valid for belief base revisions.[16] Thus the question arises: could we resolve the symmetry problems of (Because$_R$) by resorting to belief base revisions instead of belief set revisions? The answer to this question is no – as the below proposition shows.

Proposition 5. Assume a (Because$_R$) agent accepts all the formulas in $K(H, <) = K(S)$ for $H = \{t, s, sh, t \wedge s \rightarrow sh\}$, where the order of epistemic priority is $t \sim s \sim sh < t \wedge s \rightarrow sh$. Then $t \,{}^{a}\!\Rightarrow sh \notin K_>(S)$ and $sh \,{}^{a}\!\Rightarrow t \in K_>(S)$.

As compared to Proposition 4, Proposition 5 makes things even worse for Rott's analysis of because. Using belief bases, (Because$_R$) does not verify the desired direction 'because of the tower there is a shadow'. Moreover, a (Because$_R$) agent is still committed to believe the undesired direction 'because of the shadow there is a tower'. This shows that it is by no means a trivial task to find a Ramsey Test operator capturing the asymmetry of the tower-shadow example, even if we employ belief bases.

In sum, switching from belief sets to belief bases does not resolve the problem that Rott's analysis of 'because' verifies the undesired direction that there is a tower because of the shadow. In the next section, we propose an alternative strengthening of (RT$_G$) that avoids this troublesome result and fares better in capturing asymmetric relations of relevance.

3.6 Our Ramsey Test Analysis of 'Because'

3.6.1 Further Strengthening the Ramsey Test Semantics

Ramsey (1950, p. 247) expresses his idea about the semantics of conditionals as follows:

> In general we can say [...] that 'If ϕ then ψ' means that ψ is inferable from ϕ, that is, of course, from ϕ together with certain facts and laws not stated but in some way indicated by the context.

In Ramsey's approach to conditionals, so-called 'variable hypotheticals' guide an agent's inferences. More specifically, a set of these law-like variable hypotheticals, together with a set of factual beliefs, including the antecedent ϕ, let the agent infer the consequent ψ. The inferability from law-like generalisations and facts is thus tantamount to the acceptability of conditionals, as has also been pointed out in Levi's (2007) interpretation of Ramsey's approach. Ramsey's test question is whether the consequent can be inferred from generalisations judged to be reliable and some facts that specify the boundary conditions or contextual knowledge so that the generalisations are applicable. A conditional 'If ϕ then ψ' is thus acceptable just in case the consequent is inferable from the antecedent, the atomic facts judged to be true, and the judged to be reliable generalisations. Notice that Ramsey's idea requires the retention of reliable generalisations. Otherwise, the conditional cannot be inferred. This is very much in line with the discussion of non-trivial implications of Section 3.5.1.

Inspired by Ramsey, our test question is: after the suspension of judgment on everything that entails antecedent and/or consequent, is an agent disposed to infer the consequent from the antecedent and the remaining background beliefs including the generalisations? This basic idea of our semantics may be expressed by the following evaluation recipe:

[15] See Hansson (1999) for a very comprehensive study of belief base revisions and contractions, including a detailed comparison to belief set revisions and contractions.

[16] Hansson (1999, Ch. 2) discusses the recovery postulate in detail. In Van Benthem and Smets (2015, p. 302) recovery is justifiedly called "the most conroversial" AGM postulate.

> First, suspend judgement about the antecedent and the consequent. Second, add the antecedent (hypothetically) to your stock of explicit beliefs. Finally, consider whether or not the consequent is entailed by your explicit beliefs.

Our basic idea is thus split into two steps. The first step consists in an 'agnostic move', i. e. our agent suspends acceptance and/or rejection of antecedent α and consequent γ with respect to her epistemic state.[17] The second step then consists in supposing or hypothesising the antecedent α, and checking whether the consequent γ is thereby inferred.

In order to render our idea precise, we introduce a belief function that helps us formally implement the agnostic move.

Definition 14. Belief Function
Let $\perp$ be some arbitrary classical contradiction, and ϕ a formula.

$$B(\phi) = \begin{cases} \phi & \text{if } \phi \in K \\ \neg\phi & \text{if } \neg\phi \in K \\ \perp & \text{otherwise.} \end{cases}$$

Now, we are in a position to present the core of our strengthened Ramsey Test semantics. Let $\gg$ be the conditional connective of the Strengthened Ramsey Test. Then our evaluation recipe can be formally expressed as follows:

$$\alpha \gg \gamma \in K_> \text{ iff } \alpha > \gamma \in K_> - (B(\alpha) \vee B(\gamma)) \text{ iff } \gamma \in K - (B(\alpha) \vee B(\gamma)) * \alpha. \qquad (SRT_P)$$

The evaluation of $\alpha \gg \gamma$ consists of two steps. (i) The agnostic move is implemented by a contraction of the belief set K by $B(\alpha) \vee B(\gamma)$. The result is a new belief set K' such that $\neg\alpha, \alpha, \neg\gamma, \gamma \notin K'$ is guaranteed. Moreover, K' does neither contain $B(\alpha) \vee B(\gamma)$ nor $\neg B(\alpha) \vee \neg B(\gamma)$. The contraction by $B(\alpha) \vee B(\gamma)$ amounts to the agent's operation of suspending acceptance and/or rejection with respect to α and γ. We call the result of this contraction *the agnostic belief set* K'. (ii) $\alpha \gg \gamma \in K_>$ iff $\alpha > \gamma \in K'_>$ iff $\gamma \in K' * \alpha$. The second step requires for $\alpha \gg \gamma$ to be accepted that (RT_L) is satisfied for $\alpha > \gamma$ with respect to the agnostic belief set K' of step (i).

We noted that $\neg\alpha \notin K - B(\alpha) \vee B(\gamma)$. By the Levi identity, we obtain:

$$\gamma \in K - (B(\alpha) \vee B(\gamma)) * \alpha \text{ iff } \gamma \in K - (B(\alpha) \vee B(\gamma)) + \alpha. \qquad (3)$$

Moreover, we have:

$$\gamma \in K - (B(\alpha) \vee B(\gamma)) + \alpha \text{ iff } K - (B(\alpha) \vee B(\gamma)), \alpha \vdash \gamma \qquad (4)$$

where $\vdash$ is the provability relation of classical logic. We arrive thus at an alternative formulation of (SRT_P):

$$\alpha \gg \gamma \in K_> \text{ iff } K - (B(\alpha) \vee B(\gamma)), \alpha \vdash \gamma. \qquad (5)$$

This formulation emphasises the inferential character of our strengthened Ramsey Test: $\alpha \gg \gamma$ means that γ is inferable from α together with the beliefs in $K - (B(\alpha) \vee B(\gamma))$.

[17] The first step is reminiscent of the Pyrrhonian epoché by Edmund Husserl (1913, §§31-33). This phenomenological epoché denotes the method of suspending or bracketing (German: *Einklammerung*) the acceptance status of one's beliefs about the world. We apply the Pyrrhonian idea with a – by far – smaller scope: we demand an agent to suspend her respective belief status of the particular antecedent and consequent under consideration. We call the bracketing or suspension of antecedent and consequent 'agnostic move' and credit Pyrrho by labelling our Strengthened Ramsey Test (SRT_P).

Interestingly, (SRT_P) validates (Absurdity) if we employ belief sets. Suppose $\alpha, \gamma \in K$. Applying step (i) yields $\alpha, \gamma \notin K'$. However, by the recovery postulate, $(\alpha \vee \gamma) \rightarrow \gamma \in K'$ and thus, by closure, $\alpha \rightarrow \gamma \in K'$. Hence, $\gamma \in K - (\alpha \vee \gamma) * \alpha$.

In contrast, (SRT_P) invalidates (Absurdity) if we employ belief bases. For, then, recovery is not satisfied any more. Let $S = (H, <)$ be an epistemic state. Then, it is an open question whether or not $\gamma \in K(S') * \alpha$, and thus whether or not $\alpha \gg \gamma \in K_>(S)$.

Using belief bases provides Ramsey's semantics of conditionals a transparent meaning: $\alpha \gg \gamma$ means that γ is inferrable from α together with the 'facts and laws not stated' in the conditional, but explicitly stored in the agnostic epistemic state S'. In other words, our agent will accept $\alpha \gg \gamma \in K_>(S)$ only if she is disposed to (classically) infer γ from α together with the literals and generalisations stored in the agnostic epistemic state S'.

According to Ramsey's quote in Section 3.3.1 'two people' can disagree when arguing 'If p, will q?', even if both believe $\neg p$ for certain. Our (SRT_P) clarifies the sense in which $p \gg q \in K(S)$ and $p \gg \neg q \in K(S)$ are contradictories: the 'laws or hypotheses' of S must be different. It could be that two agents accept the same facts while they accept different generalisations. Hence, they would not have a dispute about facts but about how to revise the beliefs. If for 'either party' $\neg p \in K(S)$, then $K(S') * p$ entails either q or $\neg q$ or none of $q, \neg q$. A consistent epistemic agent cannot accept both $p \gg q$ and $p \gg \neg q$. Based on different sets of generalisations, however, two agents may well have different inferential dispositions. This is Ramsey's wisdom wherefore he speaks about 'two people'.

Let us compare Rott's (SRT_R) and ($\mathrm{Because}_R$) with our (SRT_P). Our semantics is closer to Gärdenfors's (RT_G) than Rott's is. The only difference, apart from using the formulation (RT_L), to (RT_G) consists in bracketing the epistemic status of antecedent and consequent. After this suspension of judgement, (RT_G) is applied in the standard way. The bracketing of the epistemic status in the agnostic epistemic state may be seen as a further strengthening of (RT_G^-) in the sense that not only the consequent is contracted from the belief set, but also the antecedent. Almost ironically, we solve problems of excessive symmetry by a 'more' symmetric contraction as compared to (RT_G^-). The additional epistemic suspension of the antecedent is the reason in virtue of which our semantics does not require a contrary-to-fact-supposition, but nevertheless expresses a relation of positive relevance. In contrast, ($\mathrm{Because}_R$) requires a counterfactual supposition in view of cases in which the antecedent α remains in $K - \gamma$. For, then $K - \gamma = (K - \gamma) * \alpha$, and so $\gamma \notin (K - \gamma) * \alpha$. Without a contrary-to-fact-supposition, our semantics does not rely on a notion that structurally resembles a counterfactual notion of causal dependence.

3.6.2 Another Analysis of 'Because'

By Ramsey's view on the relation between 'if' and 'because' and Rott's constraint on universal pro-conditionals, we can read our (SRT_P) 'if' as 'because' in the case when $\alpha, \gamma \in K$. Thus, we obtain the following analysis:

$$\alpha \ {}^{P}\!\Rightarrow \gamma \in K_> \text{ iff } \alpha \gg \gamma \in K_> \text{ and } \alpha, \gamma \in K. \qquad (\mathrm{Because}_P)$$

To avoid well-known paradoxes involving tautologies, we may furthermore require that the consequent of an explanatory relation is contingent. That is, we may require that $\gamma \notin Cn(\emptyset)$. But we shall not further explore paradoxes with logical truths surrounding conditionals and explanatory relations as this is a different topic.

Arguably, the semantics of ($\mathrm{Because}_P$) is simpler than that of ($\mathrm{Because}_R$) insofar as it does not rest on counterfactual suppositions. In comparison to (RT_G^-), our semantics puts more emphasis on the inference relation between antecedent (plus context knowledge) and consequent by bracketing the

epistemic status of the antecedent *and* the consequent – which is in the spirit of Ramsey's ideas about conditionals. Moreover, our semantics solves the class of symmetry problems associated with the tower-shadow scenario of Section 3.5.2, as we shall see in the next section.

3.6.3 Symmetry Problems Resolved

We reconsider now the tower-shadow scenario with respect to our analysis of 'because'. In the original AGM framework, we will see that our analysis does not provide the desired asymmetry, because of the recovery postulate governing the belief set. If we use belief bases, however, we can show (1) that our semantics for 'because' validates the intuitively correct (explanatory) because statement, and (2) that our semantics invalidates the intuitively incorrect, converse because statement.

Proposition 6. Assume a (Because$_P$) agent accepts all facts and the single, more entrenched generalisation of the tower-shadow scenario, i. e. $t, s, sh, t \wedge s \rightarrow sh \in K$. Then $t \stackrel{P}{\Rightarrow} sh \in K_>$ and $sh \stackrel{P}{\Rightarrow} t \in K_>$.

The proposition shows that (Because$_P$) validates (with respect to K) that 'because there is a tower, there is a shadow', as desired. However, in the original AGM framework, our semantics validates (with respect to K) also the undesired direction 'because there is a shadow, there is a tower'.

Let us move on to belief bases. The following proposition shows the asymmetry we were looking for.

Proposition 7. Assume a (Because$_P$) agent accepts all the formulas in $K(H, <) = K(S)$ for $H = \{t, s, sh, t \wedge s \rightarrow sh\}$, where the agent may assume whatever epistemic ordering $<$. Then $t \stackrel{P}{\Rightarrow} sh \in K_>(S)$, but $sh \stackrel{P}{\Rightarrow} t \notin K_>(S)$.

The proposition shows that (Because$_P$) validates (with respect to $K(S)$) that 'because there is a tower, there is a shadow', as desired. Using belief bases, our semantics invalidates (with respect to $K(S)$) the undesired direction 'because there is a shadow, there is a tower'. Note that this result holds independently of any assumptions about the epistemic ordering among the members of the belief base H.

The result is reasonable since a shadow may be cast by various things. It does not have to be a tower. To this claim, one may object that the shadow cast by *this* tower has a particular shape that is normally only produced by the very tower. This objection presupposes that an agent can uniquely infer the antecedent from the consequent. But there are frequently occurring examples where the agent is not able to do this. Here is such an example: person A sees person B taking poisonous arsenic, which leads to the death of B. Once A has suspended judgement about B's poisoning himself and his death, the assumption of B's taking arsenic lets A infer B's death, but the assumption of B's death does not allow A to infer B's intake of arsenic.

We note that the generalisation $t \wedge s \rightarrow sh$ figures as 'directed inference ticket' when using belief bases in virtue of the absence of recovery. In general, it is easy to show that an implication $\alpha \rightarrow \gamma$ is 'non-trivially' in a belief base H iff $\alpha \gg \gamma \in K_>(H, <)$, where α, γ are literals or conjunctions thereof. In contrast to $\stackrel{a}{\Rightarrow}$, the 'non-triviality' of an implication $\alpha \rightarrow \gamma \in H$ is not sufficient for the acceptance of $\gamma \gg \alpha \in K_>(H, <)$.

In the original AGM framework, (Because$_P$) succumbs to the same class of symmetry problems as Rott's analysis. In contrast to (Because$_R$), however, our semantics provides the desired asymmetry, but only if we use belief bases. We conclude that, using belief bases, our strengthened Ramsey Test semantics solves the class of symmetry problems characterised by the tower-shadow scenario. The idea behind (Because$_P$) is thus able to capture these asymmetric relations of relevance.

It is worth noting, finally, that the asymmetry of 'because' may also be captured by (Because$_P$) together with a belief set revision scheme that does not validate the recovery postulate. Severe withdrawals by Rott and Pagnucco (1999) and the related scheme of mild contractions by Levi (2004) seem to be the most obvious choices for such a scheme. These findings may cast doubt on the validity of the recovery postulate from a different angle.[18]

3.6.4 Note on Non-triviality

A note on *triviality* is in order here. As is well known, Gärdenfors (1986, 1988, Ch. 7) has shown that his version of the Ramsey Test implies, in the context of the full set of AGM postulates, that there are only *trivial* belief revision systems. (The precise meaning of triviality need not concern us here.) Therefrom, he concluded that either the Ramsey Test or a rationality postulate called *preservation*

$$\text{if } \neg\alpha \notin K \text{ and } \beta \in K, \text{then } \beta \in K * \alpha \qquad (K^*P)$$

has to be given up (where β may well be a conditional). Does our strengthened Ramsey Test fall prey to the triviality theorem? It does not. For here is a counterexample to (K^*P). Suppose $K = Cn(q \rightarrow \neg r)$. Hence, $q \gg \neg r \in K_>$. Now, let us revise K (consistently) with $q \rightarrow r$ such that (i) $\neg q < \neg q \vee r$ for the beliefs of $K' = K * (q \rightarrow r) = Cn(\{q \rightarrow r, q \rightarrow \neg r\})$. Using $(G-)$, we can infer from (i) that $(q \rightarrow r) \in K' - (\neg q \vee \top)$, where $\top$ stands for a tautology. Hence, $(q \rightarrow \neg r) \notin K' - (\neg q \vee \top)$. By the definition of $\gg$, this implies that $q \gg \neg r \notin K'_>$. Hence, (K^*P) is violated. (For belief bases, an analogous result can easily be obtained with the same formulas.) Our semantics of $\gg$ is therefore non-trivial in the sense that a crucial premise of Gärdenfors' triviality theorem is violated.

Gärdenfors's triviality result forces us to chose between preservation (for conditionals) and the Ramsey Test. Our semantics for $\gg$ does not validate preservation in the first place. Hence, we can side with the Ramsey Test without falling prey to the triviality result.

3.6.5 Note on Package Contraction

We have expressed the suspension of judgement about the antecedent α and the consequent γ using a contraction by the disjunction $\alpha \vee \gamma$. We should acknowledge, however, that the suspension of judgement can also be expressed by an operation called *package contraction*. This operation contracts a belief set K by another belief set A. Such contractions can be determined using ideas about partial meet revision, which are based on the notion of a *package remainder*:

Definition 15. $K \perp A$ (Fuhrmann and Hansson (1994, Sect. 8))
Let K and A be two sets of formulas. $B \vdash A$ means that B entails at least one member of A. $K' \in K \perp A$ iff

1. $K' \subseteq K$
2. $K' \nvdash A$
3. there is no K'' such that $K' \subset K'' \subseteq K$ and $K'' \vdash A$.

A selection function for $K \perp A$ can then be invoked, as explained in Section 3.2.5. (The notion of a selection function applies to remainder sets of belief sets and belief bases.)

[18] For the usual criticisms of this postulate, see, for instance, Hansson (1991) and Makinson (1987).

Why did we not chose package contractions to define the suspension of judgement about the antecedent and the consequent? This question is easy to answer if we work with belief bases and adopt the conventions of Section 3.2.6 as well as Definition 12, (Def σ), and (PMBC) for belief base contractions. Let us assume these definitions (i. e. Definition (12), (Def σ) and (PMBC)) also for belief base package contractions. On these conditions, it is easy to show that $(H, <) - \alpha \vee \gamma = (H, <) - \{\alpha, \gamma\}$. So it does not make a difference which operation is used. For simplicity, we chose the contraction by $\alpha \vee \gamma$ as opposed to the package contraction by $\{\alpha, \gamma\}$.

If we do not work with belief bases or deviate from the conventions in Section 3.2.6, using a package contraction by $\{\alpha, \gamma\}$ rather than a contraction by $\alpha \vee \gamma$ may well have unintended consequences. Suppose the strengthened Ramsey Test conditional $\gg_p$ is defined using a package contraction. Further, assume $\alpha, \gamma \in K(S)$ and $\alpha \to \gamma$ has high epistemic priority, and is epistemically superior to both α and γ. (Epistemic priority may be spelled out in terms of an ordering among the members of a belief base, an entrenchment ordering, or an ordering of subsets of K that defines a selection function σ.) Then, it is reasonable to expect that $\alpha \gg_p \gamma$. For, there is an inferential connection between α and γ that is based on a generalisation with high epistemic priority. Now, suppose that $\alpha \vee \gamma$ has even higher epistemic priority than $\alpha \to \gamma$. Suppose, for contradiction, $\alpha \gg_p \gamma$. By the deduction theorem, this implies that (i) $\alpha \to \gamma \in K(S')$, where $S' = S - \{\alpha, \gamma\}$. Because of the high priority of $\alpha \vee \gamma$, we have $\alpha \vee \gamma \in K(S')$ and so (ii) $\neg\alpha \to \gamma \in K(S')$. Since $K(S')$ is closed under classical logic, (i) and (ii) imply that $\gamma \in K(S')$. This, however, contradicts $S' = S - \{\alpha, \gamma\}$. Hence, $\alpha \gg_p \gamma \notin K_>(S)$.

The underlying problem is that, if $\alpha \vee \gamma$ has epistemic priority over $\alpha \to \gamma$ (while we do believe $\alpha \to \gamma$ quite firmly), the package contraction by $\{\alpha, \gamma\}$ forces us to give up $\alpha \to \gamma$. We avoid this problem if we define the suspension of judgement via a contraction by $\alpha \vee \gamma$.

3.6.6 Note on the Logic of the Strenghtened Ramsey Test Conditional

Unfortunately, working out the logical theory of our strengthened Ramsey Test Conditional $\gg$ is a comprehensive enterprise, in particular if we consider belief bases part of its semantics. On the surface, it seems that $\gg$ behaves similar to the conditional operators due to Stalnaker (1968), Lewis (1973a) and Gärdenfors (1988), or the systems of non-monotonic reasoning à la Kraus et al. (1990). However, this appearance is misleading. The logical theory of $\gg$ is quite different from the conditional logics of the mentioned operators. One of the reasons is that $\gg$ does not validate a rule called RCM by Chellas (1980), which corresponds to the rule Right Weakening (RW) in the context of non-monotonic reasoning systems. These rules are needed in many proofs of the standard properties of conditional logics, even very weak ones.

Here are the rules, where $\to$ stands for the material implication and $>$ for a conditional operator:

$$\text{RW}: \quad \frac{\vdash \alpha \to \beta \quad \gamma > \alpha}{\gamma > \beta} \tag{6}$$

$$\text{RCM}: \quad \frac{\vdash \alpha \to \beta}{(\gamma > \alpha) \to (\gamma > \beta)} \tag{7}$$

Here is a counterexample to RW for $\gg$: Suppose $H = \{p, q, r, q \wedge r \to p\}$, where $p, q, r, < q \wedge r \to p$. Let $S = (H, <)$.

(i) $S - (p \vee q) = S' = \{r, q \wedge r \to p\}$. Therefore, $S' * q \vdash p$, and so $q \gg p \in K_>(S)$.

(ii) $S - (p \vee q \vee r) = S' = \{q \wedge r \to p\}$. Therefore, $S' * q \nvdash p \vee r$ as $K(S') = Cn(\{q, q \wedge r \to p\})$. Consequently, $q \gg (p \vee r) \notin K_>(S)$.

(i) and (ii) show that the following instance of RW or RCM is *not* valid:

$$\text{RW}: \quad \frac{\vdash p \rightarrow p \vee r \quad q \gg p \in K_{>}(S)}{q \gg (p \vee r) \in K_{>}(S)} \tag{8}$$

$$\text{RCM}: \quad \frac{\vdash p \rightarrow p \vee r}{(q \gg p \in K_{>}(S)) \rightarrow (q \gg (p \vee r) \in K_{>}(S))} \tag{9}$$

We see that the logical weakening of the consequent of $\gg$ is not valid. This property might be interesting in the context of relevance logics. For now, let us resume the main thread of the chapter.

3.7 Generalising the Tower-Shadow Scenario

3.7.1 Conjunctive and Disjunctive Scenarios

We have spent quite a bit of time investigating the explanatory directions of the famous tower-shadow asymmetry. It proved anything but trivial to capture these directions in a Ramsey Test framework. Our solution to this problem is of course intended to work not only for a single example, but for a wider class of explanatory relations. Let us therefore specify further classes of explanatory relations that are well captured by (Because$_P$).

Suppose our prioritised belief base consists of two levels: an upper level G of generalisations and a lower level L of literals, as explained in Section 3.2.6. Further, let us distinguish between different types of generalisation:

$$\alpha_1 \wedge \ldots \wedge \alpha_n \rightarrow \gamma \tag{C}$$

$$\alpha_1 \vee \ldots \vee \alpha_n \rightarrow \gamma \tag{D}$$

where $\alpha_1, \ldots, \alpha_n$ are literals. We say that a generalisation of type (C) represents a conjunctive explanatory scenario, whereas a generalisation of type (D) represents a disjunctive explanatory scenario.[19] It seems as if these generalisations give rise to corresponding explanatory relations, in the sense of the present Ramsey Test analysis of 'because'. Suppose δ is a generalisation of type (C). Then, 'γ because of α_i' is verified by an epistemic state $(H, <)$ if (i) $\delta \in G$, and $\alpha_1, \ldots, \alpha_n \in L$. Suppose δ is a generalisation of type (D). Then, 'γ because of α_i' is verified by $(H, <)$ if (i) $\delta \in G$, and $\alpha_i \in L$. Recall that $H = G \cup L$.

These explanatory relations do in fact hold for a large class of conjunctive and disjunctive explanatory scenarios, but there are exceptions. Suppose a lit match that is dropped and lightning individually suffice to bring about a forest fire, on condition that oxygen is present. Further, assume that both a lit match has actually been dropped and lightning has actually occurred. So, there is a forest fire. Intuitively, we would endorse 'there is a forest fire because of lightning' and 'there is a forest fire because of the lit match'. The example can be formalised by the following prioritised belief base:

$m \wedge o \rightarrow f, \quad l \wedge o \rightarrow f$
m, l, o, f

where the propositional constants have the following natural language interpretations. m: a lit match has been dropped in the forest. l: there is lightning with electrical discharges to the ground of the forest. f: there is a forest fire. o: oxygen is present.

[19] The distinction between conjunctive and disjunctive explanatory scenarios is taken from an analogous distinction in the literature on actual causation (cf. Halpern and Pearl (2005, Sec. 3)). Disjunctive scenarios amount to cases of overdetermination if more than one of the antecedent conditions is satisfied.

Let us test for $l \gg f$. $(H,<)-(l\vee f) = (\{m\wedge o \to f, l\wedge o \to f\},<')$. Hence, $f \notin (K(H,<)-(l\vee f))+l$. Therefore, $l \gg f \notin K_>(H,<)$. So our Ramsey Test analysis of 'because'does not verify 'there is a forest fire because of lightning', which is counterintuitive.

The present example is a variant of a scenario of overdetermining causes in the literature on actual causation (Halpern and Pearl, 2005, Sec. 3). So we can describe this example as one of overdetermining causal explanations. It is easy to see that the problem in question arises not just because of the overdetermination structure but because of the combination of this structure with a background condition that is needed for the two explanations of the forest fire. We try to solve this problem by drawing on ideas about causal graphs, as introduced in the literature on actual causation (cf. Halpern and Pearl (2005)).

Let us view the members of a belief base H with two levels G and L in terms of an undirected graph in the following way. The propositional constants of the literals in L are represented by nodes, while any two nodes whose propositional constants occur together in some generalisation in G are connected by an edge. We can derive the literal γ from the literal α, using certain generalisations in G, only if these literals are connected by a path. Let us call such a path *explanatory* iff there is a derivation of γ from α that uses generalisations of G and, possibly, also literals of L. In the case of a scenario of overdetermination, we have two different explanatory paths with two different literals α and α'. Arguably, for α to explain γ, it suffices if there is a subgraph that contains one explanatory path from α to γ. This view is analogous to the widely shared intuition that overdetermining causes are proper causes.

Drawing on this picture of explanatory paths, we can account for overdetermining explanations by weakening our strengthened Ramsey Test:

$$\begin{aligned}&\alpha \gg_s \gamma \in K_>(H,<) \text{ iff there are } (H',<') \text{ and } L^-\in H \text{ s.t.}\\ &(H',<') = (H,<) - \bigvee L^-,\ \ \alpha \gg \gamma \in K_>(H',<'), \text{ and} \qquad (SRT_{P'})\\ &L^- \text{ is a possibly empty set of literals.}\end{aligned}$$

$\bigvee A$ designates an arbitrary disjunction of the members of the set A of formulas. This translates directly to our analysis of 'because':

$$\begin{aligned}&\text{Because } \alpha,\gamma \text{ (relative to } K(H,<)) \text{ iff}\\ &\quad \alpha,\gamma \in K(H,<) \text{ and } \alpha \gg_s \gamma \in K_>(H,<).\end{aligned} \qquad (\text{Because}_{P'})$$

The motivation for this refinement may be summarised as follows: to capture explanatory relations in terms of inferential connections between literals, it is sometimes necessary to ignore explanatory paths that are parallel to the one under consideration. Fortunately, there is no need to specify the meaning of 'sometimes' in this justification. In the literature on actual causation it is a common strategy to identify *active causal paths* in terms of a subset of the nodes of a given causal graph (Hitchcock, 2007; Halpern and Pearl, 2005). The index 's' in $\gg_s$ stands for 'subset', thus indicating that a subset of H suffices as background theory for α to be inferentially relevant for γ.

This refined analysis of 'because' solves our problem. For this to be seen, observe that $(H,<)$ contracted by $\bigvee\{m\}$ yields $(H',<') =$

$l \wedge o \to f, \quad m \wedge o \to f$
l, o, f

and $l \gg f \in K_>(H',<')$. In the next section, we shall see that $(\text{Because}_{P'})$ works for any combination of conjunctive and disjunctive scenarios. That is, we can combine generalisations of type (C) and (D), in an arbitrary way, to form explanatory paths.

3.7.2 Inferential Ramsey Test Explanations

Our Ramsey Test analysis of 'because', combined with belief bases, qualifies as an inferential approach to explanation. For, it is essential to this analysis that the explanandum can be inferred from the explanans, in the context of certain generalisations and possibly further background conditions. In this section, we shall specify which explanatory inferential relations are captured by our analysis, given the members of the belief base satisfy the conventions of Section 3.2.6. Thereby, we characterise a large class of explanatory relations for which our analysis of 'because' works correctly and completely.

We confine ourselves to explanatory relations between presumed facts that are expressed by literals. An inferential explanation of this type may be characterised as follows:

Definition 16. Inferential explanation of γ by α
We say that α inferentially explains γ – in the eyes of an agent a – iff there are G and L such that

1. G is a set of generalisations
2. L is a set of literals
3. α and γ are literals and believed to be true by a
4. all members of $G \cup L$ are believed to be true by a
5. $G \cup L, \alpha \vdash \gamma$
6. $G \cup L \nvdash \gamma$.

Figuratively speaking, we can say that α explains γ iff there is an inferential path from α to γ such that α is an essential premise of this path, and all premises are believed to be true. Our analysis of 'because' by (Because$_{P'}$) captures precisely this inferential understanding of an explanation:

Proposition 8. Let α and γ be literals. Epistemic states are represented by prioritised belief bases with two levels: an upper level G of generalisations and a lower level L of literals, as explained in Section 3.2.6. A (Because$_{P'}$) agent accepts 'γ because of α' with respect to $(H, <)$ iff α inferentially explains γ – in the sense of Definition 16 – in the eyes of the agent accepting all members of H.

One must wonder, however, whether (Because$_{P'}$) is strictly asymmetric in the sense that 'γ because of α' implies that 'α because of γ' does not hold. This is not so. Symmetric explanations can be constructed if we have $G \cup L, \alpha \vdash \gamma$ and $G \cup L, \gamma \vdash \alpha$.

Should we therefore further strengthen our semantics so as to yield a strictly asymmetric conditional? Is our common sense and scientific notion of 'because' asymmetric in the sense that 'γ because of α' always precludes 'α because of γ'? While it is difficult to provide a clear-cut example of a properly symmetric explanation, we hesitate to answer this question in the affirmative. We could enforce strict asymmetry, of course, by simply defining $\alpha \gg_a \gamma$ iff $\alpha \gg_s \gamma$ and $\gamma \not\gg_s \alpha$. This would rule out a number of further cases, such as inferential relations that are based on definitions or mathematical laws.

The question of whether explanatory relations are strictly asymmetric is strongly related to research in philosophy of science on explanation and causation. It is an open question whether or not any explanation must be causal, as suggested by Woodward (2003). If so, then there is good reason to suppose that explanatory relations must be asymmetric, given that causation is an asymmetric relation. Our analysis of 'because' thus becomes intertwined with an analysis of causation. Future research must show how our strengthened Ramsey Test semantics can be exploited for a fully fledged account of scientific explanation.

3.8 Conclusion

We strengthened Gärdenfors's Ramsey Test semantics for conditionals in a way which is well-motivated by Ramsey's original remarks. Like Rott's Strong Ramsey Test, but unlike Gärdenfors Ramsey Test semantics, our semantics avoids the absurdity that any two accepted formulas constitute an accepted because sentence. However, Rott's analysis of 'because' is susceptible to symmetry problems, as was shown by Proposition 3 and the tower-shadow scenario. Using belief bases, we could show that our strengthened Ramsey Test semantics captures the asymmetry of the tower-shadow scenario in an intuitively correct manner: the presence of the tower explains the presence of the shadow, but not vice versa.

We moved on to generalising the tower-shadow scenario by characterising the beliefs on the basis of which 'γ because of α' is verified by an epistemic state, according to our analysis. This generalisation reveals that our analysis of 'because' is not strictly asymmetric. At least for causal explanatory relations in scientific language, a strictly asymmetric analysis of 'because' seems to be called for. We envision to achieve such an analysis by an epistemic analysis of causation that is likewise based on the present variant of a strengthened Ramsey Test. The idea is that we can use this epistemic analysis of causation to identify actual causes and distinguish them from their effects. If we succeed, we can test whether 'because α, γ' expresses a causal relation. We turn to this epistemic analysis of actual causation in the next chapter.

Chapter 4

Causation in Terms of Production

In this chapter, we aim to analyse actual causation in terms of *production*. The latter concept is made precise by the strengthened Ramsey Test semantics of conditionals that we developed in the previous chapter. In brief, we recall that our strengthened Ramsey Test conditional $A \gg C$ is satisfied just in case C is believed in the course of assuming A, after judgement has been suspended about A and C. The idea is that this test allows us to (epistemically) verify or falsify that an event brings about another event. We will see that complementing the concept of production by a weak condition of difference-making gives rise to a full-blown analysis of causation.

Inspired by Hall (2004) and Hall (2007), we regard production rather than counterfactual dependence as the more central notion of causation. As we just mentioned, we aim to analyse the concept of production by a strengthened Ramsey Test conditional. Based on the Ramsey Test conditional, we define a concept of causation that is reductive and solves the problems of overdetermination, preemption, and double prevention. Moreover, we aim to solve the problem of spurious causation by a best system account of generalisations.

Sources. This chapter builds on joint work with Holger Andreas. Substantial content of Andreas and Günther (2019) is reprinted by permission from Springer Nature: Springer Netherlands, *Philosophical Studies*, Causation in Terms of Production, Holger, A. and Günther, M., License Number 4553020860480 (2019), advance online publication, 19 March 2019 (https://doi.org/10.1007/s11098-019-01275-3, Philos Stud).

4.1 Introduction

We aim to propose an analysis of causation that builds upon a strengthened Ramsey Test semantics of conditionals. The evaluation recipe for such conditionals can be expressed as follows:

> First, suspend judgement about the antecedent and the consequent. Second, add the antecedent (hypothetically) to your stock of explicit beliefs. Finally, consider whether or not the consequent is entailed by your explicit beliefs.

In brief, $A \gg C$ iff, after suspending judgement about A and C, C is believed in the course of assuming A. We suggest that such a test allows us to (epistemically) verify or falsify that an event brings about a certain other event, and thus qualifies as a candidate cause of the latter event. Hence, as a preliminary starting point, we define:

$$C \text{ is a cause of } E \text{ iff } C \text{ and } E \text{ occur, and } C \gg E. \qquad \text{(Def C)}$$

The logical foundations of the belief changes that define the conditional $\gg$ are explicated using AGM-style belief revision theory, as founded by Alchourrón et al. (1985) and fleshed out by Gärdenfors (1988).

We aim to set forth the first analysis of causation in terms of production that is formally as rigorous as the counterfactual accounts of causation by Lewis (1973c) and Halpern and Pearl (2005). As pointed out by Paul and Hall (2013), these counterfactual accounts face persistent problems, especially the combination of overdetermination, preemption, and double prevention. For our analysis the problem of overdetermination does not even arise. Early and late preemption turn out to be the same problem, which is solved by imposing constraints on the inferential relations underlying the conditional $\gg$. Halpern and Pearl (2005) solve the problems of overdetermination and preemption in a formally rigorous way as well, but at the price of rather complex set-theoretic constructions that are difficult to supplement with an intuitive motivation. We do not discuss causal models in this chapter. In the next chapter, however, we will translate the present analysis into the framework of causal models, and then compare it to the definition of actual cause due to Halpern and Pearl (2005) and Halpern (2015).

The plan of this chapter is straightforward: we work upward from belief changes via the strengthened Ramsey Test to an analysis of causation. In Section 4.2, we explain the basic ideas of AGM-style belief revision theory and the Ramsey Test, then moving on to our strengthened Ramsey Test conditional $\gg$. On the basis of the conditional $\gg$, Section 4.3 develops an epistemic approach to causation. In a stepwise fashion, we deal with the problems of joint effects, overdetermination and conjunctive scenarios, early and late preemption, switches and double prevention. Section 4.4 aims to answer the challenge of spurious causation by amending our analysis with a best systems account of generalisations. Section 4.5 concludes the chapter.

4.2 Belief Changes and the Ramsey Test

4.2.1 Belief Changes: Basic Ideas

AGM-style belief revision theory provides a precise semantics of belief changes for the Ramsey Test. Let us therefore make ourselves familiar with the basic ideas of belief revision. Suppose K is a set of formulas that represent the beliefs of an agent, while A is a formula that represents a single belief. In the AGM framework, one distinguishes three types of belief change of a belief set K by a formula A:

1. Expansions $K + A$
2. Revisions $K * A$
3. Contractions $K \div A$.

An expansion of K by A consists in the addition of a new belief A to the belief set K. This operation is not constrained by any considerations as to whether the new epistemic input A is consistent with the set K of present beliefs. Hence, none of the present beliefs is retracted by an expansion. $K + A$ designates the expanded belief set.

A revision of K by A, by contrast, can be described as the *consistent integration* of a new epistemic input A into a belief system K. If A is consistent with K, it holds that $K + A = K * A$, i. e. the revision by A is equivalent to the expansion by A. If, however, A is not consistent with K, some of the present beliefs are to be retracted, as a consequence of adopting the new epistemic input. $K * A$ designates the revised system of beliefs.

A contraction of K by A, finally, consists in retracting a certain formula A from the presently accepted system of beliefs. This operation will be used to define the *suspension of judgement about A* in our strengthened version of the Ramsey Test. $K \div A$ designates the belief set after the retraction of A.

In some contexts, it is helpful to distinguish between the belief system K and the epistemic state S that underlies it. Henceforth, we shall make this distinction, and write $K(S)$ for the belief system K of the epistemic state S.

Belief changes can be defined in various ways. A large number of different belief revision schemes have been developed in the spirit of the original AGM theory. We shall assume that epistemic states are represented by *belief bases*. In symbols, $S = H$. A belief base H is a set of formulas that represent the explicit beliefs of an agent. Belief base revision schemes are guided by the idea that the inferential closure of a belief base H gives us the belief set K of H:

$$K(H) =_{df} Inf(H).$$

K contains all beliefs of the epistemic state H, i. e. the explicit beliefs and those beliefs that the agent is committed to accept because they are inferable from the explicit beliefs. Inf is an inferential closure operation that may or may not be given by the consequence operation of classical logic.[1]

The definition of an expansion is straightforward for belief bases:

$$K(H) + A =_{df} K(H + A)$$

where $H + A$ stands for adding the new epistemic input to the belief base H.

Note that we can define revisions in terms of contractions and expansions:

$$K(S) * A = (K(S) \div \neg A) + A. \qquad \text{(Levi identity)}$$

Once we have retracted $\neg A$, we obtain a belief set $K(S')$ that is consistent with A. Hence, we have $K(S') * A = K(S') + A$.

In the following analysis of causation, we assume that our belief base has exactly two levels of epistemic priority: the upper level, containing the generalisations, and the lower level, which contains our beliefs about atomic facts. These levels of epistemic priority affect the determination of belief changes: when we retract a belief A, we first retract atomic beliefs that imply A (in the context of the

[1] Nonmonotonic and paraconsistent inference operations have proved useful as well, for example Brewka (1991).

generalisations). If necessary, we also retract generalisations, but only if the retraction of A cannot be achieved by retractions of beliefs with lower epistemic priority. For the considerations to follow, it may be helpful to have a graphical representation of such a prioritised belief base in mind:

G stands for the set of generalisations, while F contains the beliefs about the atomic facts. For some causal scenarios, it is necessary to distinguish between strict and *ceteris paribus* generalisations. Then, we need to have at least two levels of epistemic priority for the generalisations. The canonical scenarios of preemption, overdetermination, and double prevention can be represented using just one level of generalisations. We shall use boldfaced symbols, such as $\mathbf{H}$ and $\mathbf{H}'$, to refer to prioritised belief bases.

Such are the basic ideas about belief changes that will be used in the present analysis of causation. In the Appendix, we expound precise definitions of belief changes for prioritised belief bases. These definitions will enable the reader to formally verify, or to falsify, certain claims to be made about applications of our analysis to causal scenarios. For an intuitive understanding of what follows, the Appendix is not needed.

4.2.2 The Ramsey Test

Ramsey (1929/1990) devised an epistemic evaluation recipe for conditionals that is nowadays known as the Ramsey Test. Its core idea has been pointedly expressed by Stalnaker (1968, p. 102):

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent), finally, consider whether or not the consequent is then true.

It was then Gärdenfors (1978) who translated this test into the language of belief changes and who insisted more forcefully than Stalnaker on an epistemic understanding of conditionals. Using the AGM framework, he was able to define a semantics of conditionals explicitly in terms of belief changes:

$$A > C \in K(S) \text{ iff } C \in K(S) * A \tag{RT}$$

where $>$ designates the conditional connective. Recall that $K(S) * A$ designates the revision of the beliefs of an epistemic state S with the formula A. So the Ramsey Test defines that a conditional $A > C$ is to be accepted in a belief system $K(S)$ iff the consequent C is in $K(S)$ when revised by the antecedent A. Unlike Gärdenfors (1978), we require that A and C be non-conditional formulas.[2]

4.2.3 Strengthening the Ramsey Test

As we have seen in the previous chapter, in order to analyse the logic of 'because', Rott (1986) introduces a strengthened version of the Ramsey Test:

$$A \ggg C \text{ iff } C \in K(S) * A \text{ and } C \notin K(S) * \neg A.$$

[2]Gärdenfors (1986) has proven a triviality theorem concerning the Ramsey Test, after a conditional logic was developed on the basis of this test in Gärdenfors (1978). Recently, however, there have been various, apparently successful attempts at defending the Ramsey Test in light of this result (see, e.g., Bradley (2007)). We have shown that our strengthened Ramsey Test does not imply triviality in Chapter 3, Section 3.6.4, and Andreas and Günther (2018).

In the wake of this work, we define a conditional $\gg$ with the following intuitive meaning: $A \gg C$ iff, after suspending any beliefs in $K(S)$ as to whether A and C are true or false, it holds that $C \in K(S) * A$. The evaluation of $A \gg C$, thus, consists of two steps: (i) contracting $K(S)$ in such a manner that we become agnostic about A and C; (ii) testing whether or not C is in $K(S) * A$. In more formal terms:

Definition 17. Belief function $B(A)$
Let T be some arbitrary classical tautology and A a formula.

$$B(A) = \begin{cases} A & \text{if } A \in K(S) \\ \neg A & \text{if } \neg A \in K(S) \\ \neg T & \text{otherwise.} \end{cases}$$

$$A \gg C \in K(S) \text{ iff } C \in (K(S) \div B(A) \vee B(C)) * A. \tag{SRT}$$

Contracting $K(S)$ by $B(A) \vee B(C)$, where A and C are contingent, results in a belief system $K(S')$ that contains none of A, $\neg A$, C, $\neg C$. Moreover, $K(S')$ does neither contain $B(A) \vee B(C)$ nor $\neg B(A) \vee \neg B(C)$. Subsequently, the ordinary Ramsey Test is carried out on the belief set $K(S')$. That is, $A \gg C \in K(S)$ iff $C \in K(S') * A$.

In other words, the first step of (SRT_{SE}) consists in an *agnostic move* that lets us suspend judgement about the antecedent and the consequent. Then, we check whether or not we can infer the consequent C from the antecedent A, in the context of the remaining beliefs of the epistemic state. If so, $A \gg C \in K(S)$. Otherwise, $A \gg C \notin K(S)$.

Note that $\neg A \notin K(S) \div B(A) \vee B(C)$. By the Levi identity, this implies that

$$C \in (K(S) \div B(A) \vee B(C)) * A \text{ iff } C \in (K(S) \div B(A) \vee B(C)) + A.$$

Moreover, we have:

$$C \in (K(S) \div B(A) \vee B(C)) + A \text{ iff } (K(S) \div B(A) \vee B(C)), A \vdash C$$

where $\vdash$ designates the relation of provability in classical logic. Hence,

$$A \gg C \in K(S) \text{ iff } (K(S) \div B(A) \vee B(C)), A \vdash C.$$

This is another formulation of (SRT_{SE}). $A \gg C$ therefore means that the consequent C is inferable from the antecedent A, together with certain beliefs in $K(S)$, after judgement has been suspended about A and C. The above definition of $\gg$ is defined for epistemic states S in general, which may be represented in various ways. For the analysis to follow, however, we assume that epistemic states are represented by prioritised belief bases. In symbols, we assume that there is a prioritised belief base $\mathbf{H}$ such that $S = \mathbf{H}$.

4.3 Causation

4.3.1 Actual Causation

We are aiming at an analysis of actual causation between events. Actual causation concerns the question of whether or not a particular occurrent event causes another occurrent event. It is related to, but different from causation at the type-level, which concerns causal relations between repeatable events. Unless otherwise stated, upper case Latin letters stand for propositions saying that a certain

event occurs. By abuse of notation, we shall also speak of the event C, and thereby refer to the event whose occurrence is claimed by the proposition C.

We express the requirement that the causal relata be occurring events as follows:

$$C, E \in K(S). \tag{C1}$$

$K(S)$ designates the belief set of the epistemic state S. Our analysis is thus relative to an epistemic state, in a similar way as the analysis of Spohn (2006) and Ramsey (1929/1990) is, and perhaps the one of Hume (1739/1978).[3]

4.3.2 Production

Recall that our analysis of causation is driven by the idea that a cause brings about, or produces, its effect. Consequently, we advance the strengthened Ramsey Test as an inferential means to verify or falsify that a certain event brings about a certain other event. Hence, for C to be a cause of E it must hold that

$$C \gg E \in K(S).$$

This condition allows us to capture a large range of causal relations. It is somewhat too liberal, however. Sometimes, we are not only able to infer the effect from the cause, but also the other way around. For example, thunder caused by lightning seems to be unique in the sense that it is different from thunder caused by blasts or supersonic aircraft. (If it is not, then the better for our analysis.) Hence, lightning strongly conditionally implies – in the sense of (SRT_{SE}) – thunder. But it seems counterintuitive and wrong to view thunder as a cause of lightning. For thunder does not produce lightning.

The idea of production seems to imply a temporal asymmetry between the producing event and the effect: the cause must precede its effect. Hence,

$$t(C) < t(E) \tag{C2}$$

where $t(C)$ is a function that yields the time at which the event C occurs. (C2) expresses an old Humean dictum on causation, which is also central to Spohn's ranking-theoretic analysis of causation. We thus take the temporal order of events as not relying on causal relations. Once (C2) is in place, our analysis rules out that the thunder is a cause of the lightning, as intended. If A, B, or A and B are temporally extended events, we take $t(A) < t(B)$ to mean that A comes into being before B, while A and B may well overlap.

4.3.3 Joint Effects

Joint effects of a common cause can pose a problem for an inferential approach to causation. Take the following neuron diagram from Paul and Hall (2013, p. 71):

[3] In the *Treatise*, Hume defines: "A cause is an object precedent and contiguous to another, and so united with it, that *the idea of the one determines the mind to form the idea of the other*, and the impression of the one to form a more lively idea of the other." (Hume (1739/1978, p. 170), our emphasis). Hume can be read to slightly favour this essentially epistemic account of causation over a non-epistemic variant.

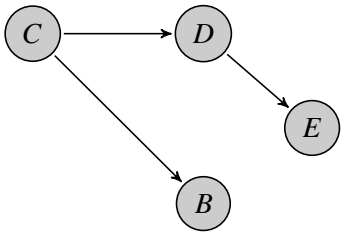

Figure 4.1: A neuron diagram for joint effects of a common cause.

C fires and, thereby, sends signals to D and B so that D and B are excited. E is excited in the course of receiving a signal from D. Intuitively, the excitation of B is not a cause of the excitation of E. However, the excitation of B strongly conditionally implies – in the sense of (SRT) – the excitation of E. For, once we have suspended judgement about the excitation of both B and E, the excitation of B lets us infer the excitation of C because there is no other way to excite B. Therefrom, we can infer that D fires, which in turn implies the excitation of E.

As observed by Paul and Hall (2013, 71-72), most counterfactual approaches solve the problem by excluding *backtracking* counterfactuals. We can adopt a similar strategy. The counterintuitive result about joint effects is avoided if all inferences that lead from the presumed cause C to the putative effect E are non-backtracking. We can make this idea precise in proof-theoretic terms. Recall that any inferential step in a natural deduction proof consists of a set P of premises and a conclusion C, where P may contain subproofs as premises. With this in mind, we can define the notion of a *forward-directed proof*. Such a proof conforms to the temporal order of events in the following sense:

Definition 18. $H \vdash_F C$
Let H be a set of formulas and C be a formula. Only literals and conjunctions of literals are taken to assert the occurrence of an event. We say there is a *forward-directed natural deduction proof* of C from H – in symbols $H \vdash_F C$ – iff there is a natural deduction proof of C from H such that (i) for all inferential steps P/I (of the main proof and any subproof), if I asserts the occurrence of an event, then this event does not precede any event that is asserted by a premise in P or by a premise in a subproof that is a member of P, and (ii) the assumption of any subproof is consistent with H.

The notion of an event is understood, in this definition, in the broad sense that includes negative events. A negative event is simply the failure of a corresponding positive event to occur. If A, B, or A and B are temporally extended events, we say that B does not precede A iff B does not come into being before A.

Why do we require that the assumption of any subproof is consistent with H? In our inferential approach to causation, subproofs are intended to represent possible ways how an event may bring about another event. Any assumption made in a subproof must therefore be consistent with H, the beliefs that form our premises. This condition will be relevant for the resolution of late preemption in Section 4.3.6.

Using this notion of a forward-directed proof, we can impose a temporal constraint on our Ramsey Test:

$$A \gg_F C \in K(S) \text{ iff } (K(S) \div B(A) \vee B(C)), A \vdash_F C. \qquad (SRT_F)$$

That is, C is a forward-directed strong conditional implication of A iff there is a forward-directed proof of C from A and the explicit beliefs of the epistemic state $\mathbf{H}$, after suspending judgement on the antecedent A and the consequent C. Recall that the epistemic state S is represented by a prioritised belief base $\mathbf{H}$.

Now, we require there to be a forward-directed proof of the putative effect from the presumed cause:

$$C \gg_F E \in K(S). \tag{C3}$$

This condition solves the problem of joint effects. For, the inference from the firing of B to the firing of C is backward-directed. Hence, there is no forward-directed proof of E from A.

The requirement of using only forward-directed inferences is well motivated by our intuitions about causation: a cause must bring about, or produce, its effect. The inferential test of this production must be forward-directed in a manner that represents the temporal order of actual productive processes in the real world.

Note that a proof is forward directed iff it is not backward directed, i. e. does not involve inferences to the occurrence of an event that precedes any event asserted in the premises. Hence, a forward-directed proof may still involve inferences where the events asserted in the premises and the conclusion are simultaneous. The exclusion of backward-directed inferences will also prove crucial to solving the problem of preemption in sections 4.3.5 and 4.3.6.

4.3.4 Overdetermination and Conjunctive Scenarios

The problem of overdetermination is severe for a counterfactual approach to causation in the tradition of Lewis (1973c). This approach is centred on the idea of counterfactual dependence between cause and effect: if the cause had not occurred, the effect would not have occurred either – according to the simple counterfactual analysis. As is well known, this approach fails in cases described as *overdetermination*. Suppose a prisoner is shot by two soldiers at the same time, and each of the bullets is fatal without any temporal precedence. Then, intuitively, both shots would qualify as causes of the death of the prisoner. However, if one of the soldiers had not shot at the prisoner, the prisoner would still have died. The death of the prisoner does therefore not counterfactually depend on the shooting by a single soldier. Hence, on the counterfactual approach, neither of the soldiers is causally responsible for the death of the prisoner.

Or, consider the following neuron diagram:

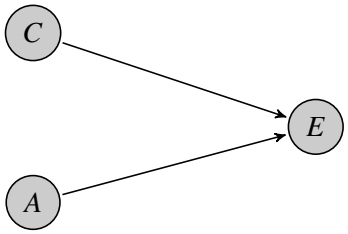

Figure 4.2: A neuron diagram for overdetermination and conjunctive scenarios.

Neurons C and A fire so that neuron E is excited. The firing of each of C and A suffices to excite E. Hence, if C had not fired, E would still have been excited. This implies that, on the counterfactual approach, neither C nor A causes E to be excited. Such results are counterintuitive and not acceptable.

Our condition that $C \gg_F E$ allows us to recognise relations of actual causation even in cases of overdetermination. It is easy to verify that the firing of C strongly conditionally implies the firing of E, in a forward directed manner. In formal notation, $C \gg_F E$, where C and E, respectively, stand for the firing of the neurons C and E. To verify $C \gg_F E$, we need to suspend judgement about the firing

of the neurons C and E. Suspending judgement about E forces us also to suspend judgement about A because our beliefs are inferentially closed. More precisely, if we were to retain the belief that A fires, we would also have to retain the belief that neuron E is excited because we believe that the firing of A triggers the firing of E. The latter belief has priority over beliefs about atomic events because it is intended to represent law-like connections between types of events. In sum, since A implies E and since we must retract our belief in E, we must also retract the belief in A.

In this way the suspension of judgement about C and E leads to an epistemic state in which we continue to believe the relations of firing between the neurons, but have no beliefs as regards the firing of C, A, and E. If we then assume that C is firing, we can infer that E is excited, in a forward-directed manner. Hence, $C \gg_F E$. Thereby, we have epistemically verified that the firing of C produces the firing of E. These inferential considerations can easily be generalised so as to capture other examples, including those with fatal bullets.

We can therefore conclude that – thanks to (SRT_{SE}) – the problem of overdetermination does not arise in the first place for our analysis of causation. This is an important advantage over counterfactual approaches to causation. As we do not have to introduce further conditions to take overdeterministic causation into account, our analysis remains relatively simple and less likely to fall prey to further counterexamples that resemble scenarios of overdetermination. In particular, we have more leeway to deal with overdetermination's close relatives known as preemption, switches, and double prevention.

Halpern and Pearl (2005) call the above neuron diagram a *conjunctive scenario* if the firing of both C and A is necessary to excite E. To give an example, it seems plausible that lightning together with a preceding drought is an – if not 'the' – actual cause of a forest fire. Indeed, it is again easy to verify that the firing of C strongly conditionally implies the firing of E, in a forward directed manner. This time the suspension of judgement regarding C and E does not force us to suspend judgement on A. The reason is that the firing of A alone is not sufficient to excite E. Hence, the suspension of judgement on C and E results in an epistemic state in which we continue to believe that A is firing. If we then assume that C is firing, we can infer that E is excited, in a forward directed manner. The same reasoning applies to the verification of $A \gg_F E$ due to the symmetry of the conjunctive scenario.

Unlike the account of Halpern and Pearl (2005), our analysis also verifies $C \wedge A \gg_F E$. As proven by Eiter and Lukasiewicz (2002), Halpern and Pearl's definition of actual causation precludes any conjunction of two (or more) events to be an actual cause (at least when the causal model contains only finitely many variables). In a conjunctive scenario, where two events are necessary for an effect to occur, the conjunction of both events is thus never an actual cause, while both events individually are recognised as actual causes. It seems surprising – to say the least – that the conjunction of two actual causes is not also an actual cause. They are forced to say that the lightning and the preceding drought together are no actual cause of the forest fire, while the lightning and the preceding drought individually are. Our analysis escapes such artefacts that stem from Halpern and Pearl's formal apparatus.

4.3.5 Early Preemption

Preemption is about backup processes: there is an event C that, intuitively, causes E. But even if C had not occurred, there is a backup event A that would have brought about E. Paul and Hall (2013, p. 75) take the following neuron diagram as canonical example of early preemption:

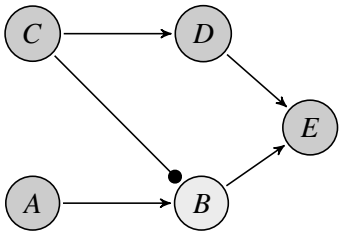

Figure 4.3: Paul and Hall's (2013) canonical neuron diagram for early preemption.

C's firing excites neuron D, which in turn leads to an excitation of neuron E. At the same time, C's firing inhibits the excitation of B. Had C not fired, however, A would have excited B, which in turn would have led to an excitation of E. Such scenarios are described as *early preemption* because the backup process does not get started in the first place, or is cut off prior to the occurrence of the effect.

How does our approach fare with respect to such scenarios? The good news is that we can easily, and without further modification, verify that C brings about E. That is, $C \gg_F E$, where C and E, respectively, stand for the excitation of the neurons C and E. Once we have suspended judgement about C and E, the assumption C lets us infer, in the context of our background theory, that D and E. These inferences are forward directed. The simple counterfactual approach, by contrast, fails here because it does not hold that the putative effect does not occur if the presumed cause is absent. In fact, E fires, even if C does not, due to the firing of A. This problem forced counterfactual approaches to be refined in various ways, see for example Lewis (1986a), and Halpern and Pearl (2005).

$A \gg E$ holds as well, but the inferential path from A to E is not strictly forward directed. Hence, $A \gg_F E$ does not hold, which is more good news. To see this, let us suspend judgement as regards A and E. Clearly, the retraction of E requires the retraction of D and C. Otherwise, we could infer E using our beliefs about the relations of firing. These beliefs have priority over our beliefs about the firing of a neuron insofar as the relations of firing represent law-like connections between events. Trivially, retraction of A requires us to retract A. Notably, however, we do not retract $\neg B$ (the belief that B is not firing), as a consequence of retracting $A \vee E$. For, $\neg B$ neither implies A nor C, nor E. B may not fire because A does not fire or because C does fire, or both. After the agnostic move, we no longer know the reason why B does not fire. $\neg B$ merely implies $\neg A \vee C$, which does not suffice to establish C inferentially, as we shall see more clearly below.

So suspension of judgement results in an epistemic state in which we continue to believe $\neg B$, but have no beliefs about A, C, D, and E. If we now assume A, we can infer E, but only via the inferential path that goes through C and D. We cannot infer E in a forward-directed manner because we believe that B does not fire, and so we believe that B cannot excite E. The inferential path via C is backward directed because it involves an inference from $\neg B$, A, and $(A \wedge \neg C) \leftrightarrow B$ to C. Clearly, this inference violates the condition of forward directedness as explained in Definition 21. For, from A firing and B not firing at a certain time t_B we can infer that C is firing at a certain time t_C, with the constraint that $t_C < t_B$. Hence, there is no forward-directed proof that would allow us to derive C from $\{\neg B, A, (A \wedge \neg C) \leftrightarrow B\}$. For simplicity, we leave the temporal order implicit, but it is obvious from the neuron diagram that a full explication of the temporal order would enable us formally to verify that the inference from A to E is not forward directed.

It is worth formalising the example. This is the epistemic state **H** that represents the scenario:

$C \leftrightarrow D$, $D \rightarrow E$, $(A \wedge \neg C) \leftrightarrow B$, $B \rightarrow E$
A, $\neg B$, C, E, D

Note that we do not only believe that $C \rightarrow D$ but also $D \rightarrow C$ because – in the confines of our causal scenario – D can only be excited by C. E, by contrast, can be excited by two different neurons. Hence, we do not believe that $E \rightarrow D$. Suspension of judgement as regards A and E results in an epistemic state $\mathbf{H}'$:

$C \leftrightarrow D$, $D \rightarrow E$, $(A \wedge \neg C) \leftrightarrow B$, $B \rightarrow E$
$\neg B$

As explained above, it is obvious that we have to retract A, C, E, and D, as a consequence of retracting $A \vee E$. Why can we retain $\neg B$? Here is the diagram of a model that verifies all members of $\bigcup \mathbf{H}'$ but fails to verify $A \vee E$:[4]

$$\{\neg A, \neg B, \neg C, \neg D, \neg E\}. \tag{1}$$

Hence, (i) $A \vee E$ is not a logical consequence of $\bigcup \mathbf{H}'$. Note furthermore that (ii) a contraction cannot produce new beliefs. Because of (i) and (ii), we can retain $\neg B$ without having any belief as to whether A or E. Retaining $\neg B$ is even mandatory by the principle of conservativity, which is central to AGM-style belief revision theory (see Gärdenfors (1988) and Hansson (1999)).

To carry out (SRT_{SE}), let us expand $K(\mathbf{H}')$ by A. Using $(A \wedge \neg C) \leftrightarrow B$ and $\neg B$, we can infer C from A, but not in a forward directed manner, for the above explained reasons. In brief, C starts to fire before B is actually inhibited; the inhibition of B by C is not immediate. Hence, $A \not\gg_F E$, and so Condition (C3) is violated.

Note, finally, that we can generalise the argument that shows us why we retain certain beliefs about an intermediate event between the preempted cause A and the effect E, after the agnostic move as regards A and E has been carried out. Retaining such beliefs is feasible and mandatory on the following grounds: (i) The intermediate event in question does not occur because the causal process from A to E is cut off by the genuine cause C. (ii) Suspension of judgement concerning A and E requires us to suspend judgement about A, C, D and E, where D is an intermediate event between the genuine cause and the effect. (iii) The failure of the intermediate event B to occur may be due to A's failure to occur, the occurrence of C, or both. (iv) $\neg B$ does not allow us to infer the effect E via the preempted pathway. (ii) and (iii) imply that (v) we can neither infer C (which would imply D and E) nor $\neg A$ from $\neg B$. (v) and (iv) imply that there is no way to infer E from $\neg B$, once A, C, D, and E have been retracted. Hence, $\neg B$ is consistent with being agnostic about A and E. Since belief changes are to be as conservative as possible, it is mandatory to retain $\neg B$. (For further details, please consult Appendix C.)

Most decisive and by no means trivial in this line of reasoning is (iii): the failure of the intermediate event to occur may be due to the non-occurrence of the preempted cause or the occurrence of the genuine cause, or both. As we are agnostic as to whether any of these causes occurs, we cannot infer from the non-occurrence of the intermediate event any claim as regards the occurrence of the genuine and the preempted cause. (If, by contrast, we were to believe that the preempted cause is present, we could infer the presence of the genuine cause from the non-occurrence of the intermediate event.) Hence, we can remain agnostic as regards the presence of the preempted and the genuine cause, while believing that the intermediate event does not occur. We shall explain why it is intuitive to require that the effect be inferable from the cause in a forward directed manner in the next section.

[4] A diagram of a model $\mathcal{A}$ is the set of all closed literals (in a given language) that are true in $\mathcal{A}$. The notion of a first order diagram has a clear analogue for propositional languages. In the case of propositional logic, a diagram contains for any propositional constant A, either A or $\neg A$. Such a diagram represents a valuation of a language of propositional logic.

4.3.6 Late Preemption

The canonical example of late preemption is presumably the most frequently cited causal scenario in the recent literature on actual causation. Suppose Billy and Suzy are throwing rocks at a bottle. Suzy's rock hits the bottle first, and thus is the genuine cause of the bottle's shattering. Billy, however, is also very skilful at throwing rocks. If Suzy had not thrown her rock, Billy's rock would have hit the bottle, and thus the bottle would have shattered a little bit later. Billy's throw is a preempted cause of the shattering of the bottle. Paul and Hall (2013, p. 99) describe this causal scenario as *late preemption*. In such a scenario, there is a backup process present that fails to go to completion merely because another process is effective prior to the backup one.

As is easy to verify, Suzy's throw strongly conditionally implies – in the sense of (SRT_{SE}) – that the bottle shatters. Likewise for Billy's throw. But only the first conditional implication can be established by a forward-directed proof. Hence, Billy's throw does not qualify as a genuine cause of the shattering of the bottle, as it should be.

Why is there no forward-directed proof of the shattering of the bottle from the assumption that Billy throws a rock at the bottle, once judgement has been suspended as regards the occurrence of these two events? The reasons for this are perfectly analogous to the reasons why there is no such proof in the case of early preempted causes. After judgement has been suspended, we continue to believe that Billy's rock did *not* hit the bottle. Therefore, there is no forward-directed inferential path from the assumption that Billy throws his rock to the shattering of the bottle. We can infer from Billy's throw and his rock not hitting the bottle that Suzy's rock has hit the bottle first. This inference, however, is not forward-directed because, by assumption, Suzy's rock arrives at the bottle before Billy's rock had a chance to hit.

Again, it is helpful to have a formal representation of the example:

ST: Suzy throws a rock at the bottle.

BT: Billy's throws a rock at the bottle.

SH: Suzy's rock hits the bottle.

BH: Billy's rock hits the bottle.

BS: The bottle shatters.

We have adopted the symbols from Halpern and Pearl (2005), with the qualification that they stand for sentences instead of variables in a causal model. Here is the epistemic state $\mathbf{H}$ that represents the causal scenario:

$ST \leftrightarrow SH, \quad SH \rightarrow BS, \quad (BT \wedge \neg SH) \leftrightarrow BH, \quad BH \rightarrow BS$
$ST, \quad BT, \quad SH, \quad \neg BH, \quad BS$

After the agnostic move as regards BT and BS, we have the epistemic state $\mathbf{H}'$:

$ST \leftrightarrow SH, \quad SH \rightarrow BS, \quad (BT \wedge \neg SH) \leftrightarrow BH, \quad BH \rightarrow BS$
$\neg BH$

Why do we continue to believe $\neg BH$ after judgement has been suspended as regards BT and BS? In other words, why is $BT \vee BS$ not a logical implication of $\bigcup \mathbf{H}'$? Here is the diagram of a countermodel to this inference:

$$\{\neg ST, \neg BT, \neg SH, \neg BH, \neg BS\}.$$

It is easy to show that the model of this diagram verifies all members of $\bigcup \mathbf{H}'$, while it does not verify $BT \vee BS$. Hence $BT \vee BS$ is not a logical implication of $\bigcup \mathbf{H}'$.

Using $BT \wedge \neg SH \leftrightarrow BH$, we can infer from BT and $\neg BH$ that SH. This inference however is not forward-directed because the event asserted by SH precedes, at least slightly, the event asserted by $\neg BH$. (Recall that our notion of a forward-directed inference applies also to premises that assert the occurrence of a negative event, understood as the absence of a certain positive event.) Hence, $BT \not\gg_F BS$. Billy's throw of the rock, though perfectly accurate, does therefore not qualify as the genuine cause of the shattering of the bottle. As is obvious, our solution exploits the consideration of *intermediate events* in similar ways as the solution by Halpern and Pearl (2005) does in the framework of structural equations.

Note once more that our insistence on forward-directed inferences is well motivated: we want to analyse causation as production by an epistemic verification that the presumed cause brings about the putative effect. Consequently, this epistemic verification must proceed by forward directed inferences. Otherwise, our inference would not model, or reconstruct, a process of production.

We should finally discuss a potential objection to the present solution to the problem of preemption. Suppose, we were to adopt the following implication laws in our belief base, in addition to those already specified:

$$BT \rightarrow (BH \vee SH)$$

$$(BH \vee SH) \rightarrow BS.$$

Then, there would be a forward-directed proof from BT to BS, in the context of the remaining beliefs, after judgement has been suspended as regards BT and BS. Both implication laws hold true of the causal scenario under consideration. Hence, $BT \gg_F BS$, after all.

There are two ways to answer this objection. First, arguably, these laws are derived and do therefore not count as *explicit beliefs*. Hence, they should not be members of the belief base. Second, we can prevent the adoption of the above implication laws by the following requirement: any consequent of any implication law of $\mathbf{H}$ asserts the occurrence of n ($n \geq 1$) distinct, non-disjunctive events. (In the case of a biimplication, we have two consequents.) This is a reasonable requirement because it forces our inferences to reconstruct the precise way in which the presumed cause has brought about the effect. In deterministic scenarios, the condition is easy to meet. Hence, we shall adopt it.

4.3.7 Switches

Switching scenarios are problematic for counterfactual accounts of causation. Lewis (1973c), for example, defines actual causation as the transitive closure of causal dependence. If the distinct events C and E occur, then E causally depends on C just in case if C had not occurred, E would not have occurred. As a consequence, counterfactual dependence of $\neg E$ on $\neg C$ is sufficient for actual causation of the occurring event E by the occurring event C. This sufficiency for causation is widely shared among the counterfactual accounts in the tradition of Lewis, e.g. by Hitchcock (2001), Woodward (2003), Hall (2004), Hall (2007), and Halpern and Pearl (2005).

The following neuron diagram represents a simplified version of a switching scenario:

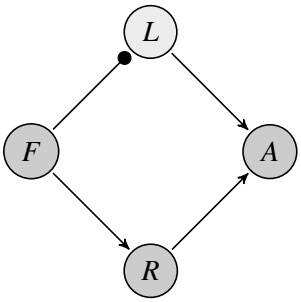

Figure 4.4: A neuron diagram for a switching scenario.

Neuron L is special: it is excited unless inhibited. The firing of neuron F excites R's firing, which in turn excites neuron A. At the same time, F's firing inhibits the excitation of L, which would have been excited in case F had not fired. In brief, F determines which one of L and R is firing, and thus acts like a switch. Speaking of events, F, R and A occur, and both $\neg R$ counterfactually depends on $\neg F$, and $\neg A$ counterfactually depends on $\neg R$. By the transitive closure imposed on the one-step causal dependences, Lewis (1973c) is forced to say that F is a cause of A.

The counterfactual accounts based on structural equations due to Hitchcock (2001) and Halpern and Pearl (2005) reject the transitivity of causation. Still, Hitchcock (2001) counts F to be a cause of A. The reason is that there is an active path from F over R to A and keeping the off-path variable L fixed at its actual value induces a counterfactual dependence of $\neg A$ on $\neg F$. Similarly, Halpern and Pearl (2005) count F to be a cause of A, since A counterfactually depends on F under the contingency that $\neg L$.

Hall (2007, p. 28) puts forth a switching scenario in which F should intuitively not count as a cause of A: Flipper is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch (F), so that the train travels down the right-hand track (R), instead of the left (L). Since the tracks reconverge up ahead, the train arrives at its destination all the same (A).

Flipping the switch does not seem to be a cause of the train's arrival. After all, there is no net effect. By assumption, 'the train arrives at its destination all the same' independent of the flipping. In other words, flipping the switch makes no difference to the train's arrival. However, the above mentioned accounts, including the ones based on structural equations, say that F is a cause of A. Furthermore, the flipping of the switch is a cause of the train traveling on the right-hand track, which in turn causes the arrival of the train. Hence, switching scenarios pose indeed a severe challenge for the transitivity of causation.

Our analysis verifies that F brings about A. Once we have suspended judgement on F and A, the assumption of F lets us infer – in a forward-directed manner – that R and thus A. Here, we see that production is necessary but not sufficient for causation. Although F is a producer of A, F is not a cause of A. Our diagnosis of the problem is: while it is possible that F brings about A and the absence of F also brings about A, in the sense of (SRT_F), it sounds paradoxical to say that F causes A and the absence of F causes A as well. The concept of causation seems to preclude that both the presence and absence of an event can genuinely cause the same effect. In the same vein, Sartorio (2005, p. 90) states "if causes are difference-makers, it is in virtue of the fact that events and their absences would not have caused the same effects."

Therefore, we complement our analysis by a weak condition of difference-making. For C to be a cause of E, the absence of C does not strongly conditionally imply E, in a forward-directed manner:

$$\neg C \gg_F E \notin K(S). \tag{C4}$$

Condition (C4) says that an event is only an actual cause if its absence does not also bring about the effect. Importantly, this condition is weaker than the difference-making in counterfactual approaches as it does not require the putative effect to be absent if the presumed cause fails to occur. The condition merely demands that the putative effect cannot be inferred from the absence of the presumed cause, once the agnostic move has been made.[5]

Let us now return to the switching scenario. We know already that F brings about A. That is, $F \gg_F A \in K(S)$. However, the absence of F also brings about A, in a forward directed manner. Hence, $\neg F \gg_F A \in K(S)$. Flipping the switch and not flipping the switch individually bring about the arrival of the train. Once Condition (C4) is in place, flipping the switch is no cause of the train's arrival. However, our analysis verifies that F causes R and R causes A. Here, the transitivity of causation fails because we can infer 'too much': the effect A will trivially obtain independent of whether or not F occurs.

To ease the verification of the details, we explicate the epistemic state $\mathbf{H}$ that represents the causal scenario:

$F \leftrightarrow R, \quad \neg F \leftrightarrow L, \quad R \rightarrow A \quad L \rightarrow A$
$F, \quad R, \quad A$

After the agnostic move as regards F and A, we have the epistemic state $\mathbf{H}'$:

$F \leftrightarrow R, \quad \neg F \leftrightarrow L, \quad R \rightarrow A \quad L \rightarrow A$

If we now assume F, we obtain R and thus A, both in a forward-directed manner. The assumption of $\neg F$ results in L and thus A, again both inferences are forward-directed.

Unlike the standard counterfactual condition of difference-making, (C4) does not imply counter-intuitive results about overdetermination. Let C be a putative cause of an effect E in a scenario of overdetermination. If we test $\neg C \gg_F E \notin K(S)$, the agnostic move forces us to suspend judgment about the occurrence of all overdetermining causes. The reason is that any overdetermining cause would imply the occurrence of the effect, which would violate the suspension of judgement as regards cause and effect. Hence, if we assume – after the agnostic move – that one of the overdetermining causes is absent, we are unable to infer the effect. $\neg C \gg_F E \notin K(S)$ is therefore satisfied for causal relations in scenarios of overdetermination.

Similar considerations apply to conjunctive scenarios and cases of preemption. In conjunctive scenarios, where two (or more) events are necessary for an effect to occur, the agnostic move forces us to suspend judgement only on the presumed cause C and the effect E. The reason is that the set of other events alone is by assumption not sufficient to bring about the effect E. In cases of preemption, observe that to test $\neg C \gg_F E \notin K(S)$ results in the same belief set $K(S \div B(C) \vee B(E))$ than the test of $C \gg_F E \in K(S)$, given that $C, E \in K(S)$. (Of course, if the latter condition is not satisfied, C is no candidate for being a cause of E.) Hence, the suspension of judgement required by Condition (C4) leads to an epistemic state in which the absence of C does not strongly conditionally imply, in a forward-directed manner, that E. In Figure 4.3, for example, the belief that $\neg B$ is retained after the agnostic move, and thus the absence of C does not bring about E. Hence, $\neg C \gg_F E \notin K(S)$ is satisfied for cases of preemption.

To sum up, our forward-directed conditional $\gg_F$ verifies some cases where the antecedent does not cause the consequent in any reasonable sense. Hence, we have complemented Condition (C3) by the weak difference-making Condition (C4) to yield a more appropriate account of causation in terms of production. As an upshot, our analysis of causation combines a notion of production and a weak

[5] Condition (C4) is inspired by Rott (1986).

notion of counterfactual dependence. Here our analysis stands in stark contrast to the postulate of Hall (2004) that there be two stand-alone concepts of causation. We rather agree with Hall (2007) that production and weak counterfactual dependence are metaphysically entangled in *one* concept of causation.

4.3.8 Double Prevention

Problems concerning double prevention have been observed at a time later than those of preemption. Suppose a boulder is dislodged and, hence, is rolling toward a hiker. The hiker sees the boulder coming and ducks so that he does not get hurt. If the hiker had not ducked, the boulder would have hurt him, in which case the hiker would not have continued the hike. Since, however, he was clever enough to duck, the hiker continues the hike. This scenario due to Hitchcock (2001, p. 276) is described as a case of *double prevention*: the boulder threatens to prevent the continuation of the hike, but provokes an action that prevents this prevention from being effective. Double prevention is thus about threat canceling.

Paul and Hall (2013, p. 216) take the following neuron diagram to represent cases of double prevention similar to Hitchcock's boulder example:

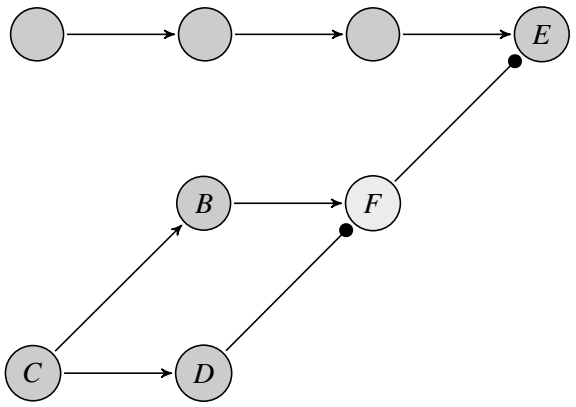

Figure 4.5: A neuron diagram for one type of double prevention.

In Figure 4.5, an event C starts a process via B that threatens to prevent E, but at the same time C initiates another process via D that prevents the threat. In particular, the firing of neuron D inhibits F's firing, which would have inhibited E's firing.

For clarity, it is helpful to have an explicit correspondence between the neuron diagram and the boulder example:

C: A boulder is dislodged.

B: The boulder rolls toward the hiker.

D: The hiker sees the boulder and ducks.

F: The hiker gets hit by the boulder.

E: The hiker continues the hike.

Like switching scenarios, the scenario seems to show that there are cases where the transitivity of causal relations is violated: the dislodged boulder C causes the ducking of the hiker D, which in turn enables the hiker to continue the hike E. But it is counterintuitive to say that the dislodging of the boulder C causes the continuation of the hike E. After all, the boulder has no net effect on the continuation of the hike.

Now, the dislodged boulder strongly conditionally implies – in the sense of (SRT_F) – that the hiker continues the hike. Once we have suspended judgement about the events C and E, we can infer from the boulder that the hiker ducks, which in turn lets us infer that the hiker continues the hike. All inferences are forward-directed. Hence, we have $C \gg_F E \in K(S)$. Conditions (C1) and (C2) are satisfied as well. However, if the boulder had not been dislodged, the hiker would have continued the hike all the same. That is, we can infer, in a forward-directed manner, E from $\neg C$, once the agnostic move has been carried out. But this means that $\neg C \gg_F E \in K(S)$ – in violation of Condition (C4). Therefore, our analysis does not accept the dislodged boulder as a cause of the continuation of the hike.

Again, it is worthwhile formalising the example considered. The following epistemic state represents our beliefs about the causal scenario:

$C \leftrightarrow B, \quad C \leftrightarrow D, \quad B \wedge \neg D \leftrightarrow F, \quad F \leftrightarrow \neg E$
$C, \quad B, \quad D, \quad \neg F, \quad E$

With respect to this epistemic state, we have $C \gg_F E \in K(S)$, but also $\neg C \gg_F E \in K(S)$. After the agnostic move, the implications are still in place, whereas all the literals are given up. Hence, the assumption of C or $\neg C$, respectively, lets us infer E, in a forward-directed manner. Due to Condition (C4) the dislodged boulder is no cause of the continuation of the hike.[6]

The structural equation accounts of Hitchcock (2001) and Halpern and Pearl (2005) classify C as a cause of E. Even worse, Hall (2007, p. 36) shows that they count C a cause of E "for *exactly the same reason*" they count, in Figure 4.3, C a cause of E, or Suzy's throw a cause of the bottle shattering. Hence, our analysis has a major advantage over these accounts by being able to capture both, scenarios of preemption and double prevention.

4.4 Spurious Causation

So far, we have relied on intuitions which generalisations should be believed or accepted in the different scenarios. For an epistemological account of causal reasoning, this relativity to what an epistemic agent judges plausible seems warranted. However, if we are interested in analysing a notion of causation, we need to take care that we can deal with the problem of spurious causation. In our framework, a spurious cause occurs if there is a common cause of at least two effects and a non-zero temporal distance between the occurrences of the two effects. This problem is nicely illustrated by the neuron diagram of Figure 4.1 in Section 4.3.3 that represents joint effects of a common cause. As there is a stable conjunction between the firing of B and E, it seems on first sight reasonable to have the implication $B \rightarrow E$ as a generalisation in the belief base. (In Section 4.3.3, by contrast, it was assumed that $B \rightarrow E$ is not a member of the belief base.) Hence, $B \gg E \in K(S)$. If the firing of B precedes that of E, the inference is forward directed, and so $B \gg_F E \in K(S)$ holds as well. The other conditions of

[6] Halpern and Pearl (2005), Halpern (2015), Halpern and Hitchcock (2015), and Spohn (2006) do not address this particular scenario of double prevention. Other scenarios of double prevention that Halpern and Pearl (2005) take into account are captured by the present analysis as well. See Example 4.4 in Halpern and Pearl (2005). Moreover, our account gives the solutions, widely acknowledged to be correct, to the so-called 'Bunzel diagrams' in Lewis (1986a, pp. 208-9).

our analysis are satisfied in an obvious way. So we would have to consider the firing of B as a cause of the firing of E. The problem is now that, by assumption, B is no actual cause of E, but only a spurious cause. (The arrows in a neuron diagram represent the genuine causal relations. In Figure 4.1, there is no arrow from B to E.) Notice that the problem of spurious causation reduces in our framework to having the 'correct' generalisations (in form of material implications) in the respective belief base. Here, 'correct' means something like 'corresponding to the arrows'. So, how do we figure out which generalisations should be in the belief base?

Inspired by Ramsey (1928/1978) and Lewis (1973a), we answer with a mild *best systems account* of generalisations: we should believe the generalisations which we would take as axioms when organising all of our beliefs as simply as possible in a deductive system. The idea is roughly that, once organised in a 'most simple' system, certain generalisations become redundant in the sense that the agent can still infer what she could infer before, but now without these generalisations. Hence, the redundant generalisations are not needed any more and can be forgotten for the sake of simplicity, whereas the non-redundant generalisations remain in the belief base to retain inferential power.

We exemplify our idea by elaborating the lightning example from Section 4.3.3.[7] A causal analysis of a thunderstorm should distinguish at least three events: (i) the electrical discharge between a cloud and the ground, (ii) the flash of the lightning, and (iii) the thunder. Arguably, the modern theory of physics contains non-redundant generalisations in the sense of a best system, namely *laws of nature*. Physics tells us that the electrical discharge between a cloud and the ground is – via the rapid production of heat within the region of the air where electricity is conducted – the common cause of the flash and the thunder. Physics would deny that the flash is a genuine cause of the thunder. So we have to conclude that the flash is a mere spurious cause of the thunder. But the event of the flash satisfies conditions (C1) – (C4) with respect to the thunder. So we must further refine our analysis.

Why is the electrical discharge rather than the flash a genuine cause of the thunder that temporally follows? Let us take a closer look at the various inferential paths in this causal scenario. The flash and the electrical discharge of the lightning are alike in that we can infer from their occurrence the occurrence of the thunder. The inferential paths, however, differ from one another. We can infer the thunder from the electrical discharge, in a forward-directed manner, using only generalisations that are laws of nature. In essence, it is the the laws of electrodynamics, atomic theory, and acoustics that are used in this path. By contrast, there is no forward-directed inferential path from the flash to the thunder such that all generalisations thereof are laws of nature. The inference from the flash of the lightning to the thunder is either not strictly forward-directed (when it goes via the electrical discharge) or not based on laws of nature (when the thunder is directly inferred from the flash, without going through the common cause of the flash and the thunder).

Why is the generalisation 'if there is a flash at the sky, then there will be a thunder a few seconds later' not a law of nature? What are laws of nature? We do not attempt to answer the latter question, but observe that the generalisation in question is not a non-redundant generalisation or law of nature in the sense of a best system account. In fact, a very weak understanding of what a best system is suffices to make the crucial distinction.

Explanation 1. Law of Nature
A generalisation is a law of nature iff it is a member of a deductive system that does a good job at balancing strength and simplicity and that is not clearly inferior to an alternative system.

This formulation is intentionally sloppy and thus acknowledges that the optimisation of strength

[7] One paradigm of a spurious cause is the drop of the barometer which does not count as a genuine cause of any storm. Since, however, the correlation between the drop of the barometer and the storm is commonly considered to be probabilistic, this example is not well suited for the present analysis of deterministic causation.

and simplicity cannot be accomplished in a straightforward manner.[8] We deem the criticism, thoroughly worked out in Van Fraassen (1989, pp. 40-64), correct that the laws of best systems are mind-dependent in virtue of the appeal to simplicity, strength, and their best balance. For obvious reasons, this criticism does not affect our account: the laws are always relative to an epistemic state. Nevertheless, Explanation 1 is still sufficient to rule out that the generalisation asserting a regular connection between a flash at the sky and a thunder is a law of nature. For, (i) it seems plausible that the laws of electrodynamics, atomic theory, acoustics, etc. (i. e. the laws that are used in the inferential path from the flash to the thunder via the electrical discharge) are members of any deductive system that does a good job at maximising strength and simplicity, and that is not clearly inferior to an alternative system. Further, (ii) our generalisation concerning flash and thunder can be derived from the laws described in (i). More precisely, it can be derived that the generalisation holds universally for a system that contains clouds above the ground and an atmosphere of a specific composition.[9] Otherwise, there would not be an inferential path from the flash to the thunder via the electrical discharge. (i) and (ii) imply that our generalisation concerning flash and thunder is not a law of nature in the sense of Explanation 1. For this to be seen more clearly, suppose we have a deductive system that contains (a) the laws of electrodynamics, atomic theory, etc. (i. e. the laws that are used in the inferential path from the flash to the thunder via the electrical discharge), (b) a general description of the type of system in which the generalisation about flash and thunder is supposed to hold, and (c) the generalisation about flash and thunder for systems described in (b). If we drop from this system component (c), we obtain a deductive system that is simpler, while being equally strong. The system with components (a), (b), and (c) is therefore clearly inferior to an alternative system.

In sum, the generalisation about flash and thunder is *redundant* in all systems that contain those generalisations that are indispensable for achieving a powerful deductive systematisation of certain optical, electrical, acoustic, and meteorological phenomena. In a similar vein, it can be shown that the generalisation about the drop of the barometer and the storm is redundant in the best deductive systems of the corresponding domain. Finally, the assertion of a regular connection between the firing of B and E (in the above neuron diagram) is redundant in the just explained sense. Readers whose intuitions about laws of nature diverge from best system accounts may well replace 'law of nature' by 'non-redundant generalisation' in what follows. So we can leave it open what laws of nature actually are.[10]

In light of the different types of inferential path in a scenario of spurious causation, an inferential account of spurious causation almost falls into place:

Definition 19. Spurious Cause
C is a spurious cause of E iff C satisfies conditions (C1) – (C4) with respect to E, and

1. there is an inferential path from C to E such that all generalisations thereof are laws of nature, but this inferential path is not forward directed, and

2. there is no forward-directed inferential path from C to E such that all generalisations thereof are laws of nature.

[8] For an accessible summary, sympathetic reconsideration, and refinement of the best system account, the reader is referred to Cohen and Callender (2009). Notably, their account is prepared to face a variety of best systems, as Explanation 1 is. It is worth noting, moreover, that Cohen and Callender (2009) are concerned with deductive systematisations of our *knowledge* of whatever domains.

[9] The description of this type of system is part of the description of a flash at the sky and the thunder. The existential statement that there are such systems may or may not be part of a best system.

[10] Strictly speaking, Explanation 1 is an account of what we judge to be a law of nature and so should explicitly be relativized to our beliefs about a certain domain. For simplicity, we continue to speak of laws of nature and leave the relativisation implicit.

Let us now cast this inferential account of spurious causation into our framework of strengthened Ramsey Test conditionals. For this to be done, a few more concepts need to be introduced:

Definition 20. $H \vdash_L C$
$H \vdash_L C$ iff there is a natural deduction proof of C from H such that all generalisations of this proof are laws of nature.

Definition 21. $H \vdash_{F-L} C$
$H \vdash_{F-L} C$ iff there is a forward-directed natural deduction proof of C from H such that all generalisations of this proof are laws of nature.

Now we can move on to defining the two conditionals needed in the account of spurious causation:

$$A \gg_L C \in K(S) \text{ iff } (K(S) \div B(A) \vee B(C)), A \vdash_L C \qquad (SRT_L)$$
$$A \gg_{F-L} C \in K(S) \text{ iff } (K(S) \div B(A) \vee B(C)), A \vdash_{F-L} C. \qquad (SRT_{F-L})$$

These two conditionals give us the formal means to exclude spurious causation:

$$\text{If } C \gg_L E \in K(S), \text{ then } C \gg_{F-L} E \in K(S). \qquad \text{(C5)}$$

This condition says that, if there is an inferential path from C to E such that all generalisations thereof are laws of nature, then there must be a forward-directed inferential path with this property. Otherwise, it is not a genuine cause. Note that we do not require that quotidian causal claims be based on laws of nature or non-redundant generalisations, and thus free of spurious causal judgements. Without a modern theory of physics, for example, a flash might well be categorised as the cause of a thunder. If we want to restrict ourselves to a scientific notion of causation, we can ban spurious causes by dropping (C5) and modifying (C3) and (C4) to:

(C3*) $C \gg_{F-L} E \in K(S)$, and

(C4*) $\neg C \gg_{F-L} E \notin K(S)$.

4.5 Conclusion

It is time for a concluding summary of our analysis. We represent causal scenarios by epistemic states, which in turn are given by prioritised belief bases. Such a base $\mathbf{H}$ has two levels: an upper level of generalisations and a lower level of beliefs about atomic facts. On grounds explained in Section 4.3.6, any consequent of any generalisation in $\bigcup \mathbf{H}$ asserts the occurrence of n ($n \geq 1$) distinct, non-disjunctive events. This in mind, we define:

Definition 22. C **is a cause of** E
The event (asserted by) C is a cause of the event (asserted by) E – relative to an epistemic state S – iff

(C1) $C, E \in K(S)$,

(C2) $t(C) < t(E)$,

(C3) $C \gg_F E \in K(S)$,

(C4) $\neg C \gg_F E \notin K(S)$, and

(C5) if $C \gg_L E \in K(S)$, then $C \gg_{F-L} E \in K(S)$.

Definition 22 puts forth an analysis of causation in terms of a strengthened Ramsey Test conditional, which is meant to express a relation of production. By condition (C3), production is necessary but not sufficient for actual causation. Similarly, (C4) amounts to a weak condition of difference-making that is necessary but not sufficient for actual causation. Causes are difference-makers in the weak sense that events and their absences do not bring about the same effects. From this follows the principle of difference making convincingly argued for by Sartorio (2005): if C is a cause of E, $\neg C$ is not a cause of E. The reason is that C is a cause of E requires both $C \gg_F E \in K(S)$ and $\neg C \gg_F E \notin K(S)$, while $\neg C$ is a cause of E requires $\neg C \gg_F E \in K(S)$ and $C \gg_F E \notin K(S)$. But this is impossible. Hence, our analysis satisfies Sartorio's constraint which she imposes on any theory of causation as a condition "that the true analysis of causation (if there is such a thing) would have to meet" (ibid., p. 75).

In the Appendix D, we explain how counterfactual dependence and our analysis of causation relate. There, we have shown that, within the original AGM belief revision theory, Lewis's (1973c) notion of counterfactual dependence is sufficient for causation as understood by our analysis when $t(C) < t(E)$. Specifically, if $C, E \in K$ for contingent propositions C, E, and counterfactual dependence is transcribed into the AGM framework as $\neg C > \neg E \in K$, it follows that C is a cause of E according to our analysis, where the epistemic state S is a classic belief set K, as opposed to a prioritised belief base $\mathbf{H}$. Back in our framework using prioritised belief bases, the two 'counterfactual' generalisations $C \rightarrow E$ and $\neg C \rightarrow \neg E$ being in $K(\mathbf{H})$ are sufficient for causation, at least when the temporal precedence and forward-directedness is assumed.

We have achieved a formally elaborated solution to the problems of overdetermination, conjunctive scenarios, (early and late) preemption, switches, and double prevention. To the best of our knowledge, there is no other formally well-defined account in the literature that solves the whole set of these problems. Finally, we have shown how to tackle the problem of spurious causation by a best system account of generalisations. Our analysis is based only on (non-redundant) generalisations and knowledge about temporal relations among events. At no point, causal relations are taken as primitive or antecedently known. In this sense, our analysis qualifies as reductive. Arguably, this reductive property is an important advantage over accounts based on causal models, such as those put forward by Halpern and Pearl (2005) and Woodward (2003). In the next chapter, we develop the idea behind the present analysis of causation within the framework of causal models. We will see that this framework simplifies our analysis considerably at the cost of losing its reductiveness. Moreover, embedding our analysis into causal models makes the comparison with Halpern and Pearl's (2005) and Halpern's (2015) definition of actual cause easier.

Chapter 5

A Ramsey-Test Analysis of Causation for Causal Models

In this chapter, we aim to provide a Ramsey-Test analysis of actual causation within the framework of causal models. Methodically, we define the strengthened Ramsey Test, as developed in the chapters 3 and 4, for causal models. The result is a variant to the definitions of actual causation put forth by Halpern and Pearl (2005) and Halpern (2015). On the way, we point out that the account of Halpern and Pearl (2005) and Halpern's modification thereof in Halpern (2015) both still struggle with the issues of overdetermination and conjunctive scenarios. We show that our variant, in contrast, deals satisfactorily with both overdetermination and conjunctive scenarios.

The goal of Halpern and Pearl (2005) is to provide a formal framework to answer the question: given all the information about a scenario (considered relevant by a modeler), is/was event C the actual cause of the event E? The answer, so Halpern and Pearl, depends on the causal model with respect to which the question is asked. They aim thus to extract the relations of actual causation (as we did in Chapter 4) from a given causal model (unlike our attempt in Chapter 4). A causal model consists of, roughly speaking, a set of structural equations and a set of (random) variables. The structural equations are supposed to represent primitive causal mechanisms or 'law-like' relationships that support a counterfactual interpretation. A structural equation assigns a value x to a random variable X. Halpern and Pearl take such an assignment to a (random) variable to represent an event.

Halpern and Pearl (2005) define actual causation with respect to a model of structural equations, or equivalently a causal model. I want to stress that claims of causation are always relative to a causal model. This means that each claim of causation is evaluated with respect to a particular causal model. Their approach allows only to claim that the value assignment $X = x$ (actually) causes the other value assignment $Y = y$ (in a particular context) in a particular model of structural equations. As a consequence the modeler must decide which events are explicitly included in the model (i. e. the endogenous variables) and which events to leave in the background (i. e. the exogenous variables). Subsequently, the modeler specifies the structural equations that model the causal influence of exogenous and endogenous variables on other endogenous variables.

Causal models can be represented by causal networks. A causal network is a directed (acyclic) graph, where the nodes correspond to the endogenous variables and there is a directed edge from a node labeled X to one labeled Y iff Y's structural equation depends on the value of the variable X. Causal networks can thus be used to depict the (recursive) dependencies of the structural equations. Notice the similarity of causal networks to Lewis's (1973c) *neuron diagrams*. Neuron diagrams depict the dependencies among a set of binary variables for one particular context. Causal networks lift the

restriction to binary variables. While the edges in a neuron diagram represent only excitatory or inhibitory functional relations, a causal network allows for arbitrary functional relations. Hence, the class of functions represented by neuron diagrams is more limited than the class represented by causal networks.

The basic idea behind Halpern and Pearl's (2005) definition is to extend Lewis's notion of causal dependence to a notion of contingent dependence. To recall, Lewis (1973c, p. 563) defines causal dependence between occurring events C, E in terms of counterfactual dependence: E causally depends on C iff C and E happened, and if C had not happened, E would not have happened. Furthermore, he identifies actual causation with the transitive closure of causal dependence. Hence, C is an actual cause of E iff there is a chain of causal dependencies from C to E. Halpern and Pearl extend this definition by (possibly non-actual) contingencies: C is an actual cause of E iff E causally depends on C *under certain contingencies*. Roughly, contingent dependence says that even if E does not counterfactually depend on C in the actual situation, E counterfactually depends on C under certain contingencies.[1]

Halpern and Pearl (2005, p. 843) give the following example to illustrate the notion of contingent dependency.[2]

Example 8. Consider two fires advancing toward a house. Fire A burned the house before fire B. Hence, we (and many juries nationwide) consider fire A 'the actual cause' for the damage, even supposing that the house would definitely have been burned down by fire B, if it were not for A.

The event of the house burning down does not counterfactually depend on fire A. However, the house's burning down counterfactually depends on fire A under the contingency that, for example, firefighters reach the house inbetween the actual arrival of fire A and (the then non-actual arrival of) fire B. The house's burning down also counterfactually depends on fire A under the contingency that fire B was not started. Symmetrically, the house's burning down also counterfactually depends on fire B under the contingency that fire A was not started. In the definition of actual causation Halpern and Pearl (2005, p. 844) offer, they aim to exclude the contingencies of the latter type that interfere with actual causal processes.

All in all, Halpern and Pearl (2005) provide a powerful framework to model actual causation. Their definition of actual cause in terms of contingent causal dependence copes well with many of the benchmark examples in the literature. As we will see in Section 5.2, they established a language to express causal relations based on the asymmetry of structural equations and the notion of an intervention (or, to be precise, a submodel).

Now, what are we up to? We will use the framework established by Halpern and Pearl (2005) to define an alternative notion of actual causation. We borrow their models of structural equations on which we define our strengthened Ramsey Test. The core idea transferred to the framework of causal models can be paraphrased as follows: for two variables A, C endogenous to the causal model under consideration, $C = c$ agnostically depends on $A = a$ if and only if, after suspending judgement on $C = c$ and $A = a$, $C = c$ will be derived by the structural equations once we set $A = a$. If we write $A = a \gg C = c$ for $C = c$ agnostically depends on $A = a$, our concept of causation can be roughly given as follows: $C = c$ is an actual cause of $E = e$ iff

(i) $C = c$ and $E = e$ occur, and

(ii) $C = c \gg E = e$.

[1] Note that Halpern and Pearl do not take the chain of the causal dependencies for their definition of actual causation. In contradiction to Lewis's stipulation, they think (Halpern and Pearl, 2005, p. 844) that causation is not always transitive.

[2] We modified the example slightly.

Remarkably, our analysis of actual causation boils down to only two conditions (due to the asymmetry of the structural equations and the definition of an intervention).

Sources. This chapter builds on joint work with Holger Andreas. Substantial content of Andreas and Günther (2018) is reprinted by permission from Oxford University Press, *The British Journal for the Philosophy of Science*, A Ramsey Test Analysis of Causation for Causal Models, Andreas, H. and Günther, M., License Number 4516510013975 (2018), advance online publication, 10 December 2018 (https://doi.org/10.1093/bjps/axy074, BJPS).

5.1 Introduction

We propose a Ramsey Test analysis of actual causation in the framework of causal models. The basic idea is to define a Ramsey Test conditional for causal models. The evaluation recipe of this new Ramsey Test conditional $\gg$ can be informally stated as follows:

> First, suspend judgment about the antecedent and the consequent. Second, add the antecedent (hypothetically) to your stock of beliefs. Finally, consider whether or not the consequent is entailed by your beliefs.

In brief, the conditional $A \gg C$ should be believed iff, after suspending judgment on A and C, C is believed as a result of assuming A.[3] The conditional $\gg$ gives rise to a template for an analysis of causation:

$$C \text{ is a cause of } E \text{ iff } C \text{ and } E \text{ occur, and } C \gg E.$$

In the chapters 3 and 4, we have shown that $\gg$ can be explicated using AGM-style belief revision theory, as founded by Alchourrón et al. (1985) and Gärdenfors (1988). Still using some concepts of belief revision, however, we aim to define in this chapter the conditional $\gg$ within the framework of causal models. Thereby, the above template becomes a surprisingly powerful analysis of causation.

Why analyse causation by a Ramsey Test conditional in the framework of causal models? After all, Halpern and Pearl (2005) already put forward a formally precise definition of actual causation that captures a wide range of causal scenarios, including troublesome cases of preemption. However, as they admit, their definition is "complicated" (ibid., p. 880). Moreover, even this highly elaborate definition runs into a problem. As proven by Eiter and Lukasiewicz (2002), this definition precludes any conjunction of two (or more) events to be an actual cause. In a conjunctive scenario, where two events are necessary for an effect to occur, the conjunction of both events is thus no actual cause, while both events individually are recognized as actual causes. It seems surprising – to say the least – that the conjunction of two actual causes is not also an actual cause. If lightning and a preceding drought are necessary factors for a forest fire to occur, the conjunction of these factors should qualify as an – if not 'the' – actual cause.

Halpern's (2015) modification of the Halpern and Pearl (2005) definition is more elegant and allows – in principle – to count conjunctions as actual causes. Henceforth, we refer to the Halpern and Pearl (2005) definition also as 'HP definition' and to Halpern's (2015) modification as 'modified HP definition'. As we shall see later, the modified HP definition inherits the problem of the conjunctive scenario. The modification comes with another cost. In scenarios of overdetermination, the modified definition does not recognize the overdetermining causes as actual causes. (Surprisingly, the modified HP definition classifies the conjunction of the overdetermining causes as an actual cause.) We observe that no Halpern-Pearl definition satisfactorily solves both conjunctive scenarios and cases of overdetermination. We show that this predicament can be remedied in the framework of causal models by our Ramsey Test analysis of causation.

The plan of this chapter is straightforward: we work upward from causal models via the strengthened Ramsey Test to an analysis of causation. Section 5.2 extends Halpern and Pearl's causal model semantics by the notion of an agnostic model. Based on this extension, Section 5.3 presents our strengthened Ramsey Test and our analysis of actual causation. Section 5.4 applies the analysis to overdetermination, conjunctive scenarios, and preemption. Section 5.5 compares our Ramsey Test variant to the Halpern-Pearl definitions of actual causation. Section 6.4 concludes the chapter.

[3] For a detailed presentation and embedding of our strengthened Ramsey Test into the literature, see Andreas and Günther (2018) or Chapter 3.

5.2 An Extension of Causal Model Semantics

We extend Halpern and Pearl's (2005) causal model semantics by the notion of an agnostic model. This model represents the suspension of judgment in our strengthened Ramsey Test. We briefly introduce causal models and the Arsonists Example, before we define the notion of an agnostic model and its corresponding operator of suspension.

5.2.1 Halpern and Pearl's Causal Model Semantics

The semantics of conditionals due to Halpern and Pearl (2005, pp. 851-852) is defined with respect to a causal model over a signature.

Definition 23. Signature $\mathcal{S}$
A signature $\mathcal{S}$ is a triple $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$, where $\mathcal{U}$ is a finite set of exogenous variables, $\mathcal{V}$ is a finite set of endogenous variables, and $\mathcal{R}$ maps any variable $Y \in \mathcal{U} \cup \mathcal{V}$ to a non-empty (but finite) set $\mathcal{R}(Y)$ of possible values for Y.

Definition 24. Causal Model M
A causal model over signature $\mathcal{S}$ is a tuple $M = \langle \mathcal{S}, \mathcal{F} \rangle$, where $\mathcal{F}$ maps each endogenous variable $X \in \mathcal{V}$ to a function $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} \setminus \{X\}} \mathcal{R}(Y)) \mapsto \mathcal{R}(X)$.

The mapping $\mathcal{F}$ defines a set of (modifiable) structural equations that model the causal influence of exogenous and endogenous variables on other endogenous variables. The function F_X determines the value of $X \in \mathcal{V}$ given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. Note that $\mathcal{F}$ defines no structural equation for any exogenous variable $U \in \mathcal{U}$.

A structural equation such as $x = F_X(\vec{u}, y)$ says (in a context where the exogenous variables take the values $\vec{u}$), if Y *were* set to y (by means of an intervention), then X *would* take on the value x. Notice the difference to a direct intervention on X: an intervention that assigns a value $x' \neq x$ to X (by means external to the model) overrules the value x assigned by $F_X(\vec{u}, y)$. The difference points to a 'causal' asymmetry. In any structural equation, the values of the variables on the right-hand side jointly determine the value of the variable on the left-hand side; if, however, the value of the variable on the left-hand side is changed by means of an external intervention, the values of the variables on the right-hand side remain unaffected.

Halpern and Pearl (2005, p. 849) confine themselves to models of recursive structural equations.[4] In such models, the causal dependences among the variables in $\mathcal{V}$ can be represented by a directed acyclic graph. Hence, there is a strict partial order such that, if $X \prec Y$, then the value of X may affect the value of Y, but the value of Y cannot have any effect on the value of X. A consequence of the restriction to models M of recursive equations is that, given a context $\vec{U} = \vec{u}$, there is always a unique solution to the equations in M; for we can always solve the equations in the order given by $\prec$.

In the spirit of Halpern and Hitchcock (2010, p. 397), we make two assumptions concerning the variables of a causal model: (i) no value of a variable X logically implies a value of another variable Y, and (ii) the different values of the same variable are mutually exclusive and jointly exhaustive.

Given a signature $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$, we say that a formula of the form $X = x$, where $X \in \mathcal{U} \cup \mathcal{V}$ and $x \in \mathcal{R}(X)$, is a primitive value assignment. Halpern and Pearl (2005) and Halpern (2015) interpret

[4] However, Halpern and Pearl (2005, pp. 883-884) provide a definition of actual causation for non-recursive models in their appendix.

primitive value assignments as 'primitive events'.[5] We write $\vec{X}$ for a (finite) set $\{X_1, .., X_n\}$ of variables, and $\vec{x}$ for a (finite) set $\{x_1, .., x_n\}$ of values of the variables in $\vec{X}$. We abbreviate the set of primitive value assignments $\{X_1 = x_1, .., X_n = x_n\}$ by $\vec{X} = \vec{x}$.[6]

The values of the variables in $M = \langle \mathcal{S}, \mathcal{F} \rangle$ are not determined. If the exogenous variables are not set to a value, the structural equations in $\mathcal{F}$ are not triggered and thus no endogenous variable is assigned a value. Only if we set all the exogenous variables $U \in \mathcal{U}$ to exactly one of their values, the structural equations propagate the causal influence from parent to child variables such that each endogenous variable $V \in \mathcal{V}$ is assigned exactly one value. In other words, if a context $\vec{u} = \{u_1, ..., u_n\}$ is set for all n variables $U_i \in \mathcal{U}$, the structural equations recursively determine the value of each endogenous variable. The result is that in a contextualized causal model $\langle M, \vec{u} \rangle$ every variable in $\mathcal{U} \cup \mathcal{V}$ is assigned exactly one value.

In Halpern and Pearl's (2005) semantics, a simple conditional $[Y = y]X = x$ is satisfied in a contextualized causal model $\langle M, \vec{u} \rangle$ if the intervention setting the variable Y to the value y results in X taking on the value x.[7] Such an intervention is defined by the notion of a submodel $M_{Y=y}$ of M.

Definition 25. Submodel $M_{\vec{X}=\vec{x}}$
Let $M = \langle \mathcal{S}, \mathcal{F} \rangle$ be a causal model, $\vec{X}$ a (possibly empty) set of variables in $\mathcal{V}$ and $\vec{x}, \vec{u}$ sets of values for the variables in $\vec{X}, \vec{U}$. We call the causal model $M_{\vec{X}=\vec{x}} = \langle \mathcal{S}_{\vec{X}}, \mathcal{F}^{\vec{X}=\vec{x}} \rangle$ over signature $\mathcal{S}_{\vec{X}} = \langle \mathcal{U}, \mathcal{V} \setminus \vec{X}, \mathcal{R} \restriction_{\mathcal{V} \setminus \vec{X}} \rangle$ a submodel of M. $\mathcal{F}^{\vec{X}=\vec{x}}$ maps each variable in $\mathcal{V} \setminus \vec{X}$ to a function $F_Y^{\vec{X}=\vec{x}}$ that corresponds to F_Y for the variables in $\mathcal{V} \setminus \vec{X}$, and sets the variables in $\vec{X}$ to $\vec{x}$.

Let us consider an example of a causal model due to Halpern and Pearl (2005, p. 856):

Example 9. Arsonists Example
Suppose that two arsonists drop lit matches in different parts of a dry forest, and both cause trees to start burning. Consider two scenarios. In the first, called the *disjunctive scenario*, either match by itself suffices to burn down the whole forest. That is, even if only one match were lit, the forest would burn down. In the second scenario, called the *conjunctive scenario*, both matches are necessary to burn down the forest; if only one match were lit, the fire would die down before the forest was consumed.

We can describe the essential structure of these two scenarios using a causal model with four variables:

- an exogenous variable U that determines, among other things, the motivation and state of mind of the arsonists. For simplicity, assume that $\mathcal{R}(U) = \{u_{00}, u_{10}, u_{01}, u_{11}\}$; if $U = u_{ij}$, then the first arsonist intends to start a fire iff $i = 1$ and the second arsonist intends to start a fire iff $j = 1$. In both scenarios $U = u_{11}$;[8]

[5] Although we find the interpretation of primitive value assignments as primitive events remarkable, we will relegate the discussion of this issue to another occasion, where we also discuss omissions. For some challenging observations on this interpretation, see Hall (2007).

[6] We deviate from Halpern and Pearl's presentation of causal models insofar as $\vec{X}$ denotes a set of variables and not a vector of variables. However, Halpern and Pearl (2005, p. 849, footnote 1) are "implicitly identifying the vector $\vec{X}$ with the subset of $\mathcal{V}$ consisting of the variables in $\vec{X}$". Halpern (2015, footnote 2) uses the vector notation, but sometimes views "$\vec{Z}$ as a set of variables."

[7] We follow Fenton-Glynn (2017) in conveying "=" a double role: the sign figures as (i) identity and (ii) an assignment operator such as Halpern and Pearl's (2005) "←" or Pearl's (2009) "do(·)". Although this double role is mathematically sloppy, it avoids an unnecessary multiplication of notation and is harmless as long as we keep in mind that $[\vec{Y} = \vec{y}]\phi$ expresses a conditional whose antecedent assigns the variables in $\vec{Y}$ the values $\vec{y}$.

[8] We will normally leave the context $\vec{u}$ implicit in the following examples (as is common practice in causal modeling). See, for example, Halpern and Pearl (2005) and Halpern and Hitchcock (2010).

- endogenous variables ML_1 and ML_2, each either 0 or 1, where $ML_i = 0$ if arsonist i does not drop the lit match and $ML_i = 1$ if he does, for $i = 1, 2$;
- an endogenous variable FB for forest burns down, with values 0 (it does not) and 1 (it does).

The two scenarios differ with respect to the structural equation for FB:

- Disjunctive scenario: $FB = F_{FB}(ML_1, ML_2) = max(ML_1, ML_2)$.
- Conjunctive scenario: $FB = F_{FB}(ML_1, ML_2) = min(ML_1, ML_2)$.

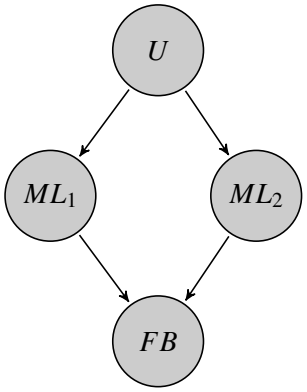

Figure 5.1: Causal network for the Arsonists Example. The arrows represent the dependences among the variables as encoded by the structural equations.

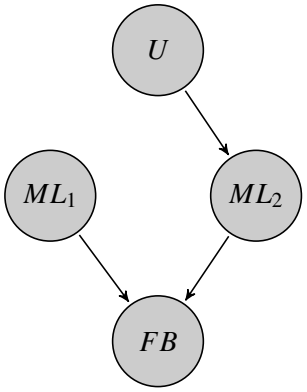

Figure 5.2: The causal network of the submodel for setting $ML_1 = 0$. The removal of the arrow between U and ML_1 represents that F_{ML_1} does not any more depend on the value of U.

Causal networks can be used to depict the recursive dependences of the structural equations, as in figures 6.1 and 6.2. A causal network is a directed acyclic graph, where the nodes correspond to the variables and there is an arrow from a node labeled X to one labeled Y iff X is a parent variable of Y iff F_Y depends on the value of X iff $X \prec Y$.[9] Analogous to the asymmetry of structural equations, causal networks have the property that variables only causally affect their descendants: if Y is not a descendant of X, then a change in the value of X has no effect on the value of Y.

Let us now check whether or not the conditional $[ML_1 = 0]FB = 1$ is true in the contextualized causal model $\langle M, u_{11}\rangle$ of the Arsonists Example. The intervention that sets $ML_1 = 0$ induces a submodel $M_{ML_1=0}$ of M. If the solution to the structural equations of $M_{ML_1=0}$ satisfies $FB = 1$, then $[ML_1 = 0]FB = 1$ is true in the contextualized causal model. If the conditional is indeed satisfied in the contextualized causal model, we write $\langle M, u_{11}\rangle \models [ML_1 = 0]FB = 1$.

In the disjunctive scenario, $\langle M, u_{11}\rangle \models [ML_1 = 0]FB = 1$ iff $\langle M_{ML_1=0}, u_{11}\rangle \models FB = 1$.[10] The structural equations for the submodel $M_{ML_1=0}$ are:

- $F_{ML_1}^{ML_1=0} = 0$. Hence, $ML_1 = 0$.
- $F_{ML_2}^{ML_1=0}(u_{11}) = 1$. Hence, $ML_2 = 1$.

[9] Note that the acyclicity follows from the assumption that the structural equations are recursive.

[10] Notice that the operator $[ML_1 = 0]$ relates $\langle M, u_{11}\rangle$ and $\langle M_{ML_1=0}, u_{11}\rangle$. In general, [] can be seen as both an operator on a (contextualized) causal model and as the antecedent of a conditional.

- $F_{FB}^{ML_1=0}(ML_1, ML_2) = max(ML_1, ML_2)$. Since $ML_2 = 1$, $FB = 1$.

The solution of the structural equations of $M_{ML_1=0}$ satisfies $FB = 1$, and thus $\langle M, u_{11}\rangle \models [ML_1 = 0]FB = 1$. Notice the difference between the structural equation $F_{ML_1}(u_{11}) = 1$ and $F_{ML_1}^{ML_1=0} = 0$: the former depends on the value of U, whereas the latter does not. After the intervention that sets $ML_1 = 0$, the variable ML_1 is treated as if it were an exogenous variable, viz. it is assigned a value by its structural equation that does not depend on other variables. The structural equations for the other variables, i. e. $\mathcal{V} \setminus \{ML_1\}$, remain unchanged: $F_{ML_2}(u_{11}) = 1 = F_{ML_2}^{ML_1=0}(u_{11}) = 1$ and $F_{FB}(ML_1, ML_2) = F_{FB}^{ML_1=0}(ML_1, ML_2) = max(ML_1, ML_2)$.

The conjunctive scenario differs from the disjunctive scenario only in the structural equation for FB: $F_{FB}^{ML_1=0}(ML_1, ML_2) = min(ML_1, ML_2)$. Since $ML_1 = 0$, $FB = 0$. Hence, $\langle M, u_{11}\rangle \not\models [ML_1 = 0]FB = 1$. We conclude that the conditional $[ML_1 = 0]FB = 1$ is true in the disjunctive scenario and not true in the conjunctive scenario.

Before turning to our extension, we observe a correspondence between an arbitrary set of primitive value assignments $\vec{X} = \vec{x}$ and the conjunction of all members of $\vec{X} = \vec{x}$, i. e. $\bigwedge \vec{X} = \vec{x}$. Given a contextualized causal model $\langle M, \vec{u}\rangle$, and an arbitrary Boolean combination of value assignments ϕ, $\langle M, \vec{u}\rangle \models [\vec{X} = \vec{x}]\phi$ iff $\langle M, \vec{u}\rangle \models [\bigwedge \vec{X} = \vec{x}]\phi$, where $\bigwedge \vec{X} = \vec{x} = (X_1 = x_1 \wedge ... \wedge X_i = x_i)$. Based on this correspondence, we will move back and forth between sets of primitive value assignments and their big conjunctions – without further mentioning. Moreover, by convention we take $\langle M, \vec{u}\rangle \models [\vec{X} = \vec{x}, \vec{Y} = \vec{y}]\phi$ to be synonymous with $\langle M, \vec{u}\rangle \models [\vec{X} = \vec{x} \cup \vec{Y} = \vec{y}]\phi$.

5.2.2 Agnostic Models

Our Ramsey Test requires a suspension of judgment. In order to define the suspension of judgment, we extend Halpern and Pearl's (2005) causal model semantics by a new type of model, which we call agnostic. Such an agnostic model is meant to represent the suspension of judgment. Intuitively, the actual context is suspended in an agnostic model such that the value assignments are (partially) indeterminate: the values of some variables are suspended, while the values of other variables are unchanged. The idea behind such a (partially) agnostic value assignment is that some variables of a causal model are assigned the set of all of their possible values. Formally, an agnostic value assignment is a function $\mathcal{A}$ that assigns each variable Y of a subset of a causal model's variables the set $\mathcal{R}(Y) = \{y_1, ..., y_n\}$ of its mutually exclusive and jointly exhaustive values. Notice that $\mathcal{A}$ assigns sets of values to endogenous and possibly exogenous variables.

An agnostic value assignment contrasts with a context of a causal model. A context $\vec{u}$ uniquely determines the value assignment $\vec{u} \cup \vec{v}$ of the variables $\mathcal{U} \cup \mathcal{V}$ with respect to $\langle M, \vec{u}\rangle$. A *partially* agnostic value assignment $\mathcal{A}$ renders indeterminate the value assignment $\vec{u} \cup \vec{v}$ of *some* of the variables in $\mathcal{U} \cup \mathcal{V}$ with respect to $\langle M, \vec{u}\rangle$.

Similar to Halpern and Pearl's (2005) [] operator based on the notion of a submodel, we now define a new operator $[]_s$ based on the notion of an agnostic model.

Definition 26. Agnostic Model $\langle M_{\vec{X}=\vec{x}}, \mathcal{A}\rangle$
Let $\langle M, \vec{u}\rangle$ be a contextualized causal model and $\vec{X} = \vec{x}$ a set of primitive value assignments for a (possibly empty) subset of the variables in $\mathcal{U} \cup \mathcal{V}$. We say that $\langle M_{\vec{X}=\vec{x}}, \mathcal{A}\rangle$ is the agnostic model of $\langle M, \vec{u}\rangle$ with the core $\vec{X} = \vec{x}$ iff

(1) $M_{\vec{X}=\vec{x}}$ is a submodel of M, and

(2) $\mathcal{A}(Y_i) = \mathcal{R}(Y_i)$ for all $Y_i \in (\mathcal{U} \cup \mathcal{V}) \setminus \vec{X}$.

An agnostic model $\langle M_{\vec{X}=\vec{x}}, \mathcal{A}\rangle$ with the core $\vec{X} = \vec{x}$ is agnostic with respect to the variables in the set $\vec{Y} = \mathcal{U} \cup \mathcal{V} \setminus \vec{X}$. The variables in $\vec{X}$ are set to the values $\vec{x}$ independent of the context value(s) for $\vec{U}$. Intuitively, the core $\vec{X} = \vec{x}$ contains the variable assignments that are protected from the suspension of judgment. In other words, we are not agnostic about the variables in the core. Hence, the variables of the causal model are partitioned into the variables $\vec{Y}$, on which we suspend judgment, and the variables $\vec{X}$, on which we do not suspend judgment.

We define satisfaction of a formula ϕ in an agnostic model as follows: $\langle M_{\vec{X}=\vec{x}}, \mathcal{A}\rangle \models \phi$ iff $\langle M_{\vec{X}=\vec{x}}, \vec{u} \cup \vec{v}\rangle \models \phi$ for any value assignment $\vec{u} \cup \vec{v}$ to the variables $\mathcal{U} \cup \mathcal{V}$ that respects the structural equations of $M_{\vec{X}=\vec{x}}$.[11] Note that $M_{\vec{X}=\vec{x}} = M$ in the special case where $\vec{X}$ is the empty set. We define the suspension operator $[]_s$ as follows:

$$\langle M, \vec{u}\rangle[\vec{X} = \vec{x}]_s = \langle M_{\vec{X}=\vec{x}}, \mathcal{A}\rangle \qquad \text{(Def-}[]_s\text{)}$$

The suspension operator $[]_s$ renders the values of the variables $\vec{Y} = \mathcal{U} \cup \mathcal{V} \setminus \vec{X}$ indeterminate.

To summarize, the suspension of judgement on the variables $\mathcal{U} \cup \mathcal{V} \setminus \vec{X}$ with respect to a contextualized causal model $\langle M, \vec{u}\rangle$ results in an agnostic model $\langle M_{\vec{X}=\vec{x}}, \mathcal{A}\rangle$. After this 'agnostic move', the model $\langle M_{\vec{X}=\vec{x}}, \mathcal{A}\rangle$ satisfies ϕ iff any value assignment $\vec{u} \cup \vec{v}$ respecting the structural equations of the submodel $M_{\vec{X}=\vec{x}}$ satisfies ϕ.

5.3 A Strengthened Ramsey Test for Causal Models

We briefly introduce the idea of the Ramsey Test and uncover its relation to Halpern and Pearl's (2005) causal model semantics. Subsequently, we define our strengthened Ramsey Test. By the template presented in the Introduction, this test gives us an analysis of causation.

5.3.1 The Ramsey Test and Causal Models

Ramsey (1929/1990) proposed a test to evaluate conditionals. When an agent has no firm opinion on a proposition A, she should believe a conditional 'if A then B' if she believes B as a result of adding A to her stock of beliefs. Stalnaker (1968, p. 102) extended Ramsey's test to cover the remaining epistemic attitudes with respect to the antecedent:

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true.

In brief, the recipe says that the conditional $A > C$ should be believed just in case C is believed as a result of assuming A.

Structurally, Stalnaker's Ramsey Test resembles the evaluation of conditionals in Halpern and Pearl's causal model semantics. For this to be seen, recall that the conditional $[Y = y]X = x$ is true, if X takes on the value x as a result of setting Y on y. Of course, the conditional $[Y = y]X = x$ is only true or false with respect to a contextualized causal model – very much like the Ramsey Test is relative to the beliefs of an epistemic state. The intervention in a causal model plays a role similar to the assumption of the antecedent in the Ramsey Test. Likewise, the value of the variable X after the intervention corresponds to the belief in the consequent after the Ramsey Test revision. In other words, the consequent of the Ramsey Test is comparable to the solution for the variable X of the

[11] Henceforth, we simply write $\vec{u} \cup \vec{v}$ for a value assignment $\mathcal{U} \cup \mathcal{V} = \vec{u} \cup \vec{v}$.

structural equations in the submodel $M_{Y=y}$. On top of this, the structural equations of the submodel provide a precise meaning to the phrase 'as a result of'. Hence, Halpern and Pearl's causal model semantics may be regarded as containing a fully worked out version of the Ramsey Test. In fact, they flesh out the (hypothetical) assumption of the Ramsey Test as a (hypothetical) intervention. In the next section, we devise a strengthened Ramsey Test for causal models.

5.3.2 A Strengthened Ramsey Test for Causal Models

Let us recall our exposition from the introduction of the strengthened Ramsey Test:

> First, suspend judgment about the antecedent and the consequent. Second, add the antecedent (hypothetically) to your stock of beliefs. Finally, consider whether or not the consequent is entailed by your beliefs.

We translate now this test into the framework of causal models. Let $Prim\langle M, \vec{u}\rangle$ denote the set of primitive value assignments that are satisfied in the contextualized causal model $\langle M, \vec{u}\rangle$. In symbols,

$$Prim\langle M, \vec{u}\rangle = \{X = x \mid \langle M, \vec{u}\rangle \models X = x\}.$$

Having this shorthand for the actual value assignments, how do we determine which beliefs or judgments are to be suspended in our strengthened Ramsey Test? Inspired by AGM belief revision, we define a contraction operator that turns a contextualized causal model into a partially agnostic model such that ϕ is not believed (any more) in the latter. We define the contraction operator via a remainder set of primitive value assignments:

Definition 27. Remainder Set $\langle M, \vec{u}\rangle\perp\phi$
Let $\langle M, \vec{u}\rangle$ be a contextualized causal model. ϕ is a Boolean combination of value assignments. We call $\langle M, \vec{u}\rangle\perp\phi$ the remainder set of $\langle M, \vec{u}\rangle$ with respect to ϕ:
$P \in \langle M, \vec{u}\rangle\perp\phi$ iff
(1) $P \subseteq Prim\langle M, \vec{u}\rangle$, and

(2) $\langle M, \vec{u}\rangle[P]_s \not\models \phi$, and

(3) there is no P' such that $P \subset P' \subseteq Prim\langle M, \vec{u}\rangle$ and $\langle M, \vec{u}\rangle[P']_s \not\models \phi$.

Condition (1) simply demands that the beliefs that need not be suspended are actual beliefs. Intuitively, Condition (2) demands that the contextualized causal model $\langle M, \vec{u}\rangle$ does not satisfy ϕ after suspending judgment on all the variables of M, except the variables in P. Condition (3) says that each member P of the remainder set is maximal in the sense that adding an actual belief would revoke the suspension of judgment with respect to ϕ. A remainder set $\langle M, \vec{u}\rangle\perp\phi$ contains thus all maximal subsets P of $Prim\langle M, \vec{u}\rangle$ that do not satisfy ϕ with respect to the respective agnostic models $\langle M_P, \mathcal{A}\rangle$.

Using the suspension operator and the definition of a remainder set, we define now the contraction operator $[]^-$. The contraction of the contextualized causal model $\langle M, \vec{u}\rangle$ by the Boolean combination ϕ is given by:

$$\langle M, \vec{u}\rangle[\phi]^- = \langle M, \vec{u}\rangle[\bigcap \langle M, \vec{u}\rangle\perp\phi]_s\,, \qquad \text{(Def } []^-\text{)}$$

where $\bigcap\langle M, \vec{u}\rangle\perp\phi$ designates the intersection of all the members of the remainder set $\langle M, \vec{u}\rangle\perp\phi$. $\bigcap\langle M, \vec{u}\rangle\perp\phi$ represents the primitive value assignments that survived the suspension. Henceforth, we abbreviate $\bigcap\langle M, \vec{u}\rangle\perp\phi$ by P^-.

We are now in a position to define our strengthened Ramsey Test conditional $\gg$. In brief, $\vec{X} = \vec{x} \gg \phi$ iff ϕ is entailed by the solution to the structural equations of the agnostic model $\langle M_{P^-}, \mathcal{A}\rangle$,

once $\vec{X}$ is set to $\vec{x}$, where $P^- = \bigcap\langle M, \vec{u}\rangle\bot(\vec{X} = \vec{x} \vee \phi)$. The evaluation of $\vec{X} = \vec{x} \gg \phi$ comprises two steps: (i) contracting the set $Prim\langle M, \vec{u}\rangle$ of primitive value assignments such that we become agnostic on whether or not $(\vec{X} = \vec{x}) \vee \phi$. (ii) Checking whether or not ϕ is entailed by the structural equations of the agnostic model $\langle M_{P^-}, \mathcal{A}\rangle$ after setting $\vec{X}$ to its original values $\vec{x}$. We define the strengthened Ramsey Test for the contextualized causal model $\langle M, \vec{u}\rangle$ as follows:

$$\langle M, \vec{u}\rangle \models \vec{X} = \vec{x} \gg \phi \text{ iff } \langle M, \vec{u}\rangle[\vec{X} = \vec{x} \vee \phi]^- \models [\vec{X} = \vec{x}]\phi \qquad (\text{SRT}_{SE})$$

We explain SRT_{SE} by the following proposition and corollary.

Proposition 9. Let $\langle M, \vec{u}\rangle$ be a contextualized causal model, $\vec{X} = \vec{x}$ a set of primitive value assignments and ϕ a Boolean combination of value assignments. Then $\langle M, \vec{u}\rangle \models \vec{X} = \vec{x} \gg \phi$ iff $\langle M_{P^-,\vec{X}=\vec{x}}, \vec{u} \cup \vec{v}\rangle \models \phi$ for all value assignments $\vec{u} \cup \vec{v}$ which respect the structural equations of $M_{P^-,\vec{X}=\vec{x}}$.

Proof. By SRT_{SE}, $\langle M, \vec{u}\rangle \models \vec{X} = \vec{x} \gg \phi$ iff $\langle M, \vec{u}\rangle[\vec{X} = \vec{x} \vee \phi]^- \models [\vec{X} = \vec{x}]\phi$. Abbreviating $\bigcap\langle M, \vec{u}\rangle\bot(\vec{X} = \vec{x} \vee \phi)$ by P^-, the definition of $[]_s$ gives us $\langle M_{P^-}, \mathcal{A}\rangle \models [\vec{X} = \vec{x}]\phi$. By the definition of [], we obtain $\langle M_{P^-,\vec{X}=\vec{x}}, \mathcal{A}\rangle \models \phi$. This means by the definition of satisfaction in an agnostic model nothing but $\langle M_{P^-,\vec{X}=\vec{x}}, \vec{u} \cup \vec{v}\rangle \models \phi$ for all value assignments $\vec{u} \cup \vec{v}$ that respect the structural equations of $M_{P^-,\vec{X}=\vec{x}}$. □

Corollary 1. Let $\langle M, \vec{u}\rangle$ be a contextualized causal model, P^- and $\vec{X} = \vec{x}$ sets of primitive value assignments and ϕ a Boolean combination of value assignments. Then $\langle M, \vec{u}\rangle[P^-]_s \models [\vec{X} = \vec{x}]\phi$ iff $\langle M, \vec{u}\rangle[P^- \cup \vec{X} = \vec{x}]_s \models \phi$.

Proof. Both sides of the equivalence are equivalent to $\langle M_{P^-,\vec{X}=\vec{x}}, \mathcal{A}\rangle \models \phi$. □

We should make our notational convention explicit that for any operator $[]^*$ on causal models, we have:

$$\langle M, \vec{u}\rangle[\ldots]^* \models \phi \text{ iff } \langle M, \vec{u}\rangle \models [\ldots]^*\phi.$$

In words, $[]^*$ can be seen as both an operator on a contextualized causal model and as the antecedent of a conditional. In particular, we can present the strengthened Ramsey Test as a nested conditional

$$\langle M, \vec{u}\rangle \models \vec{X} = \vec{x} \gg \phi \text{ iff } \langle M, \vec{u}\rangle \models [\vec{X} = \vec{x} \vee \phi]^-[\vec{X} = \vec{x}]\phi.$$

5.3.3 A Ramsey-Test Definition of Actual Causation

Based on the template of the introduction, we define actual causation with respect to a contextualized causal model as follows:

Definition 28. $\vec{X} = \vec{x}$ **is an Actual Cause of** ϕ **with respect to** $\langle M, \vec{u}\rangle$
Let $\langle M, \vec{u}\rangle$ be a causal model, $\vec{X} = \vec{x}$ a set of primitive value assignments and ϕ a Boolean combination of value assignments. $\vec{X} = \vec{x}$ is an actual cause of ϕ with respect to $\langle M, \vec{u}\rangle$ iff

C1 $\langle M, \vec{u}\rangle \models \vec{X} = \vec{x} \wedge \phi$, and

C2 $\langle M, \vec{u}\rangle \models \vec{X} = \vec{x} \gg \phi$.

According to C1, for $\vec{X} = \vec{x}$ to be an actual cause of ϕ, both $\vec{X} = \vec{x}$ and ϕ must occur in the actual context. Hence, the members of $\vec{X} = \vec{x}$ are actual value assignments. C2 demands that the cause $\vec{X} = \vec{x}$ and the effect ϕ form the antecedent and the consequent of a Strengthened Ramsey Test conditional that holds true in the respective causal model. That is, C2 demands that $\langle M, \vec{u}\rangle[\vec{X} = \vec{x} \vee \phi]^- \models [\vec{X} = \vec{x}]\phi$. The suspension of judgment on the actual belief $\vec{X} = \vec{x} \vee \phi$ results in believing the actual primitive value assignments of $P^- \subset Prim\langle M, \vec{u}\rangle$. Observe that the suspension of judgment does not always force us to give up all actual beliefs. For, the beliefs in P^- are protected from the suspension of judgment. This protection is implemented by setting the structural equations of the variables in P^- to their actual values, as expressed by M_{P^-}. In this way, actual beliefs surviving the suspension of judgment may influence our judgments on relations of actual causation.

5.3.4 Minimality

As we shall see below, we do not need any minimality condition to treat the canonical examples of overdetermination, preemption, and conjunctive scenarios. Such a minimality condition is here superfluous due to the simplicity of the examples and standard translations into causal models. Notably, causal models usually do not contain irrelevant variables. However, we can implement a minimality condition by rephrasing C2 as follows:

$$\text{C2'}\ \ \langle M, \vec{u}\rangle \models \vec{X'} = \vec{x'} \gg \phi \text{ for all } \vec{X'} = \vec{x'} \subseteq \vec{X} = \vec{x}.$$

C2' says that each subset of an actual cause needs to be an actual cause. If a subset $\vec{X'} = \vec{x'}$ is irrelevant, it follows that $\langle M, \vec{u}\rangle \not\models \vec{X'} = \vec{x'} \gg \phi$. Hence, C2' prevents that actual causes contain irrelevant conjuncts. A more thorough motivation and presentation of C2' will be given when requisite.

5.4 Applying the Definition of Actual Causation

We apply now our definition of actual causation to overdetermination, conjunctive scenarios, and preemption.

5.4.1 Overdetermination and Conjunctive Scenarios

Recall the Arsonists Example from Section 5.2.1. The causal model comprises three endogenous variables ML_1, ML_2, and FB, each taking either the value 0 or the value 1. In the actual context u_{11}, all three variables take the value 1. In the disjunctive scenario, each of ML_1 and ML_2 are individually sufficient for $FB = 1$. The disjunctive scenario is thus a scenario of symmetric overdetermination. The set of actual value assignments is $Prim\langle M, u_{11}\rangle = \{ML_1 = 1, ML_2 = 1, FB = 1\}$.

By Definition 28, $ML_1 = 1$ is an actual cause of $FB = 1$ with respect to the contextualized causal model $\langle M, u_{11}\rangle$ iff

C1 $\langle M, u_{11}\rangle \models ML_1 = 1 \wedge FB = 1$, and

C2 $\langle M, u_{11}\rangle \models ML_1 = 1 \gg FB = 1$.

C1 is obviously satisified. We verify condition C2. By SRT_{SE},

$$\langle M, u_{11}\rangle \models ML_1 = 1 \gg FB = 1 \text{ iff}$$
$$\langle M, u_{11}\rangle[ML_1 = 1 \vee FB = 1]^- \models [ML_1 = 1]FB = 1.$$

Abbreviating $\bigcap\langle M, \vec{u}\rangle\bot(ML_1 = 1 \vee FB = 1)$ by P^-, the definition of $[]_s$ gives us $\langle M_{P^-}, \mathcal{A}\rangle \models [ML_1 = 1]FB = 1$. Notice that the agnostic set of value assignments is $P^- = \emptyset$. In particular, we need to suspend judgment on ML_2. Otherwise, we could derive $FB = 1$ by the structural equation $F_{FB} = max(ML_1, ML_2)$, which would violate the suspension of judgment on $ML_1 = 1 \vee FB = 1$. Hence, in the agnostic model $\langle M_{P^-}, \mathcal{A}\rangle$ of $\langle M, u_{11}\rangle$ no actual value assignment is retained.

To be more precise, suppose for reductio that $P^- = \{ML_2 = 1\}$. Then, by Definition 41 and Def $[]^-$, $ML_2 = 1 \in P$ for all P in the set $\langle M, u_{11}\rangle\bot(ML_1 = 1 \vee FB = 1)$. But then $\langle M, u_{11}\rangle[P]_s \models ML_1 = 1 \vee FB = 1$ – in violation of condition (2) of Definition 41. Therefore, we would, after the suspension of judgment on $ML_1 = 1 \vee FB = 1$, obtain $\langle M_{P^-}, \mathcal{A}\rangle \models ML_1 = 1 \vee FB = 1$. That is $\langle M_{P^-}, \vec{u} \cup \vec{v}\rangle \models ML_1 = 1 \vee FB = 1$ for all possible value assignments $\vec{u} \cup \vec{v}$ which respect the structural equations of M_{P^-}. Observe that, for each $\vec{u} \cup \vec{v}$, setting ML_2 to the value 1 results in $FB = 1$, as the structural equation for FB remains ready to be triggered.

It remains to check whether or not $\langle M_{P^-}, \mathcal{A}\rangle \models [ML_1 = 1]FB = 1$. If we set ML_1 back to its original value 1, FB takes on the value 1 under each value assignment respecting the structural equations of $M_{P^-,ML_1=1}$. In symbols, $\langle M_{P^-,ML_1=1}, \mathcal{A}\rangle \models FB = 1$. We conclude that ML_1 is an actual cause of $FB = 1$ in the disjunctive scenario.

Let us now move on to the conjunctive scenario of the Arsonist's example. Here, both ML_1 and ML_2 need to take the value 1 for FB taking on the value 1. Accordingly, the structural equation for FB is $F_{FB} = min(ML_1, ML_2)$, while the scenario remains otherwise unchanged. Now, the suspension of judgment on $ML_1 = 1 \vee FB = 1$ does not require to retract the belief in ML_2 since ML_1 alone is not sufficient to entail $FB = 1$ via the structural equations. Formally, $\bigcap\langle M, \vec{u}\rangle\bot(ML_1 = 1 \vee FB = 1)$ is identical to $P^- = \{ML_2 = 1\}$. The reason is that there is a possible value assignment $\vec{u} \cup \vec{v} = \{ML_1 = 0, ML_2 = 1, FB = 0\}$ respecting the structural equations of M_{P^-} such that $\langle M_{P^-}, \vec{u} \cup \vec{v}\rangle \not\models FB = 1$. Hence, $\langle M, \vec{u}\rangle[P^-]_s \not\models FB = 1$.

It remains to check whether or not $\langle M_{P^-}, \mathcal{A}\rangle \models [ML_1 = 1]FB = 1$. If we set ML_1 back to 1 in the agnostic model $\langle M_{P^-}, \mathcal{A}\rangle$, FB takes on the value 1 under each value assignment respecting the structural equations of $M_{P^-,ML_1=1}$. We conclude that $ML_1 = 1$ is an actual cause of $FB = 1$ in the conjunctive scenario.

By the same reasoning pattern, ML_2 is an actual cause of $FB = 1$ in both the disjunctive and conjunctive scenario due to the symmetry of the Arsonists Example. Moreover, we leave it to the reader to verify that the conjunction $ML_1 = 1 \wedge ML_2 = 1$ is also an actual cause of $FB = 1$ in both scenarios.

5.4.2 Preemption

Let us consider the following case of preemption analyzed by Halpern and Pearl (2005, p. 861):

Example 10. Suzy & Billy Throw Rocks at a Bottle
Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle had it not been preempted by Suzy's throw.

In such scenarios the actual cause preempts a merely potential cause. Here Suzy's throw is the actual cause of the bottle's shattering, which preempts Billy's throw. The key characteristic in this preemption case is: if Billy throws, the bottle will shatter – no matter what else is *actually* going on. Even if we counterfactually assume that Suzy is not throwing her rock at the bottle, knowing that Billy throws allows us still to infer that the bottle shatters. Nevertheless, Billy's throw does intuitively

not count as an *actual* cause of the bottle's shattering (while Billy's throw does count as a potential cause). Other scenarios of preemption have a similar structure.

We formalize the example following Halpern and Pearl (2005, pp. 861-864). The example comprises five endogenous variables:

- ST for 'Suzy throws' with two values: 0 (Suzy does not throw), 1 (she does).
- BT for 'Billy throws' with two values: 0 (Billy does not throw), 1 (he does).
- SH for 'Suzy's rock hits the bottle' with two values: 0 (it does not), 1 (it does).
- BH for 'Billy's rock hits the bottle' with two values: 0 (it does not), 1 (it does).
- BS for 'the bottle shatters' with two values: 0 (it does not), 1 (it does).

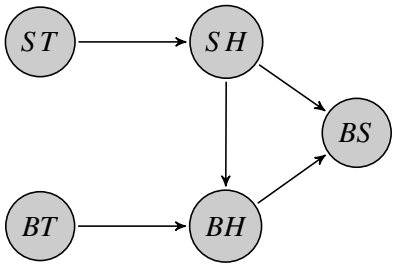

Figure 5.3: Causal network for the Suzy & Billy Throw Rocks at a Bottle Example.

We have the following structural equations:

- $F_{SH} = ST$.
- $F_{BH} = min(BT, \neg SH)$. Note in particular $BH = 1$ iff $BT = 1$ and $SH = 0$.
- $F_{BS} = max(SH, BH)$.

The actual context $\vec{u}$ determines that $ST = BT = 1$. The ensuing set of actual value assignments is $Prim\langle M, \vec{u}\rangle = \{ST = 1, SH = 1, BS = 1, BT = 1, BH = 0\}$.

We check first whether Suzy's throw ($ST = 1$) is an actual cause of the bottle's shattering ($BS = 1$). By Definition 28, $ST = 1$ is an actual cause of $BS = 1$ with respect to the contextualized causal model $\langle M, \vec{u}\rangle$ iff

C1 $\langle M, \vec{u}\rangle \models ST = 1 \wedge BS = 1$, and

C2 $\langle M, \vec{u}\rangle \models ST = 1 \gg BS = 1$.

Condition C1 is obviously satisfied. We verify C2. By SRT_{SE},

$$\langle M, \vec{u}\rangle \models ST = 1 \gg BS = 1 \text{ iff}$$
$$\langle M, \vec{u}\rangle[ST = 1 \vee BS = 1]^{-} \models [ST = 1]BS = 1.$$

Abbreviating $\bigcap\langle M, \vec{u}\rangle\perp(ST = 1 \vee BS = 1)$ by P^{-}, the definition of $[]_s$ gives us $\langle M_{P^{-}}, \mathcal{A}\rangle \models [ST = 1]BS = 1$. By Definition 41, $P = \{BT = 1, BH = 0\}$ is the only member of the remainder set $\langle M, \vec{u}\rangle\perp(ST = 1 \vee BS = 1)$, and so $P = P^{-}$. It is instructive to take a closer look at why this is

so. Clearly, P is a subset of $Prim\langle M, \vec{u}\rangle$. Furthermore, $\langle M, \vec{u}\rangle[P]_s \not\models ST = 1 \vee BS = 1$. The reason is that there is a possible value assignment $\vec{u} \cup \vec{v} = \{ST = 0, SH = 0, BS = 0, BT = 1, BH = 0\}$ that respects the structural equations of M_P such that, in this value assignment, $ST = 1 \vee BS = 1$ is false. (Respecting the structural equations means that the structural equations for BT and BH are set to 1 and 0, respectively, and the other structural equations remain intact). Hence, $\langle M_P, \mathcal{A}\rangle \not\models ST = 1 \vee BS = 1$. However, if an actual value assignment to one or more of the other variables would remain in P, the agnostic model $\langle M_P, \mathcal{A}\rangle$ would satisfy $ST = 1 \vee BS = 1$. Notice also that the set $\{BH = 0\}$ is a subset of $\{BT = 1, BH = 0\}$, and so it is excluded as a member of the remainder set $\langle M, \vec{u}\rangle\perp(ST = 1 \vee BS = 1)$ by the maximality condition of Definition 41.

It remains to check whether or not $\langle M_{P^-}, \mathcal{A}\rangle \models [ST = 1]BS = 1$, which is the case iff $\langle M_{P^-,ST=1}, \vec{u} \cup \vec{v}\rangle \models BS = 1$ for each possible value assignment $\vec{u} \cup \vec{v}$ respecting the structural equations of $M_{P^-,ST=1}$. The structural equations of $M_{P^-,ST=1}$ ensure that $ST = 1$ and so $SH = 1$, which in turn leads to $BS = 1$. We obtain $\langle M_{P^-,ST=1}, \mathcal{A}\rangle \models BS = 1$. Therefore, Suzy's throw is recognized as an actual cause of the bottle's shattering.

We demonstrate now that Billy's throw ($BT = 1$) is not an actual cause of the bottle's shattering ($BS = 1$). The reasoning is analogous to the above case, where Suzy throws. By Definition 28, $BT = 1$ is an actual cause of $BS = 1$ with respect to the contextualized causal model $\langle M, \vec{u}\rangle$ iff

C1 $\langle M, \vec{u}\rangle \models BT = 1 \wedge BS = 1$, and

C2 $\langle M, \vec{u}\rangle \models BT = 1 \gg BS = 1$.

Condition C1 is obviously satisfied. We falsify C2. By SRT_{SE},

$$\langle M, \vec{u}\rangle \models BT = 1 \gg BS = 1 \text{ iff}$$
$$\langle M, \vec{u}\rangle[BT = 1 \vee BS = 1]^- \models [BT = 1]BS = 1.$$

Abbreviating $\bigcap\langle M, \vec{u}\rangle\perp(BT = 1 \vee BS = 1)$ by P^-, the definition of $[]_s$ gives us $\langle M_{P^-}, \mathcal{A}\rangle \models [BT = 1]BS = 1$. This time $P^- = \{BH = 0\}$. There is only one maximal $P = \{BH = 0\}$ such that $\langle M, \vec{u}\rangle[P]_s \not\models BT = 1 \vee BS = 1$. The reason is that there is the possible value assignment $\vec{u} \cup \vec{v} = \{ST = 0, BT = 0, SH = 0, BH = 0, BS = 0\}$ which respects the structural equations of M_P. Hence, $\langle M, \vec{u}\rangle[P^-]_s \not\models BT = 1 \vee BS = 1$. Furthermore, P^- is maximal because including any other primitive value assignment (or assignments) of the set of actual assignments would result in an agnostic model that satisfies $BT = 1 \vee BS = 1$.

Let us now determine whether or not $\langle M_{P^-}, \mathcal{A}\rangle \models [BT = 1]BS = 1$. If we set $BT = 1$ in the agnostic model $\langle M_{P^-}, \mathcal{A}\rangle$, the model does not satisfy $BS = 1$. The reason is that the structural equation for BH is set to its actual value 0, and so the influence of BT on BS is blocked by BH. In other words, the intervention that sets BH on its actual value cancels BH's dependence on BT (and SH). Consequently, the value of BH cannot be determined by the value of BT (and/or SH).

More formally, $\langle M_{P^-}, \mathcal{A}\rangle \models [BT = 1]BS = 1$ iff $\langle M_{H^-,BT=1}, \vec{u} \cup \vec{v}\rangle \models BS = 1$ for each possible value assignment $\vec{u} \cup \vec{v}$ respecting the structural equations of $M_{P^-,BT=1}$. Since, however, $\vec{u} \cup \vec{v} = \{ST = 0, BT = 1, SH = 0, BH = 0, BS = 0\}$ respects the structural equations of $M_{P^-,BT=1}$, we have $\langle M_{P^-}, \mathcal{A}\rangle \not\models [BT = 1]BS = 1$. Hence, Billy's throw is not an actual cause of the bottle's shattering.

On our account, actuality influences whether or not there is an *actual* causal relation. The actual value assignment $BH = 0$ is exempt from the suspension of judgment, and thus remains in P^-. In other words, the 'actuality' $BH = 0$ survives the suspension of judgment by setting the structural equation for BH to the value 0. Consequently, BH takes on its actual value in the agnostic model $\langle M_{P^-}, \mathcal{A}\rangle$ independent of the values for BT and SH. This is an example of how certain parts of actuality may "intervene" to tell apart actual causes from merely potential causes.

5.5 Comparison to the Halpern-Pearl Definitions

We now compare the Halpern-Pearl definitions of actual causation to our Ramsey Test variant. In the introduction, we cited Eiter and Lukasiewicz's (2002) proof: according to Halpern and Pearl's (2005) definition the conjunction of two or more events never qualifies as an actual cause (at least when the causal model contains only finitely many variables). Recall the conjunctive scenario of the Arsonists Example, where both arsonists need to drop lit matches for the forest to burn down. Given the actual context, that is both arsonists intend to start a fire, each of them individually qualifies as an actual cause according to both the Halpern-Pearl definitions and our Ramsey Test variant.[12] However, both Halpern-Pearl definitions do not admit the conjunction of both arsonists dropping lit matches as an actual cause. But why should this conjunction *not* count as an actual cause of the forest's burning down? It seems that making an actual cause explicit in addition to another should not invalidate their conjunction as actual cause. Merely bringing an event of the actual context from the background into consideration should not invalidate a relation of actual causation, especially if this actually occurring event is a necessary factor for the effect to occur.

The modified HP definition considerably simplifies the definition given in Halpern and Pearl (2005). The modification licenses – in principle – conjunctive causes. To see why the conjunction of the arsonists dropping their lit matches is not an actual cause, we present the modified HP definition given in Halpern (2015):

Definition 29. $\vec{X} = \vec{x}$ **is an Actual Cause of** ϕ **with respect to** $\langle M, \vec{u}\rangle$
Let $\langle M, \vec{u}\rangle$ be a causal model, $\vec{X} = \vec{x}$ a set of primitive events and ϕ a Boolean combination of primitive events. $\vec{X} = \vec{x}$ is an actual cause of ϕ with respect to $\langle M, \vec{u}\rangle$ iff

AC1 $\langle M, \vec{u}\rangle \models \vec{X} = \vec{x} \wedge \phi$, and

AC2 There is a set $\vec{W}$ of variables in $\mathcal{V}$ and a setting $\vec{x'}$ of the variables in $\vec{X}$ such that:

$$\text{if } \langle M, \vec{u}\rangle \models \vec{W} = \vec{w}, \text{ then } \langle M, \vec{u}\rangle \models [\vec{X} = \vec{x'}, \vec{W} = \vec{w}]\neg\phi.$$

AC3 $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions AC1 and AC2.[13]

According to this definition, $ML_1 = 1 \wedge ML_2 = 1$ (or rather $\{ML_1, ML_2\}$) is not an actual cause of $FB = 1$ with respect to the conjunctive scenario. The reason is AC3. There are subsets of $\{ML_1, ML_2\}$, e.g. $\{ML_1\}$, that satisfy AC1 and AC2. For instance, let $\vec{W} = \{ML_2\}$. Consequently,

$$\text{if } \langle M, \vec{u}\rangle \models ML_2 = 1, \text{ then } \langle M, \vec{u}\rangle \models [ML_1 = 0, ML_2 = 1]FB = 0,$$

which is true because $\langle M_{ML_1=0,ML_2=1}, \vec{u}\rangle \models FB = 0$.

The modified definition, so reads Halpern's (2015) abstract, "gives reasonable answers (that agree with those of the [HP definition]) in the standard problematic examples". As we have just seen, the two definitions agree indeed in the conjunctive scenario. The conjunction of both primitive events, that are necessary for the effect to occur and individually recognized as actual causes, does not form an actual cause itself. We would not call this result 'reasonable'. As pointed out in the Introduction,

[12] According to our definition, why is $ML_1 = 1$ an actual cause of $FB = 1$ in the conjunctive scenario? Informally, because the actual context has it that $ML_2 = 1$ and the suspension of judgment does not require to give up this belief.

[13] The notation $[\vec{X} = \vec{x}]\phi$ can be read as a non-backtracking counterfactual conditional $X_1 = x_1 \wedge ... \wedge X_k = x_k \boxdot\!\!\rightarrow \phi$. If so inclined, the relation to Lewis's (1973c) analysis of causation becomes evident: if $\vec{X}$ had not been $\vec{x}$, ϕ would not have occurred. AC2 adds to this clause 'at least under certain contingencies'. The idea behind Definition 29 is thus roughly 'ϕ but for $\vec{X} = \vec{x}$ under the actual contingencies $\vec{W} = \vec{w}$'.

if lightning and a preceding drought are necessary factors for a forest fire to occur, the conjunction of these factors should qualify as an – if not 'the' – actual cause. All of the considered cases, the conjunctive scenario, the disjunctive scenario (or overdetermination), and preemption are 'standard problematic examples'. Interestingly, we will see below that the Halpern-Pearl definitions do not agree on overdetermination.

Let us consider our example of preemption again. We will find that the modified definition and our variant use a similar strategy, viz. an intervention of actuality. According to the modified definition, Suzy's throw ($ST = 1$) is an actual cause of the bottle's shattering ($BS = 1$) with respect to $\langle M, \vec{u}\rangle$. AC1 and AC3 are obviously satisfied. As to AC2, let $\vec{W} = \{BH\}$. Consequently,

$$\text{if } \langle M, \vec{u}\rangle \models BH = 0, \text{ then } \langle M, \vec{u}\rangle \models [ST = 0, BH = 0]BS = 0$$

is satisfied, as $\langle M_{ST=0,BH=0}, \vec{u}\rangle \models BS = 0$. Notice that choosing $\vec{W}$ to be $\{BH\}$ is tantamount to keeping BH fixed at its actual value 0. As the structural equation of BH is set to 0 in $\langle M_{ST=0,BH=0}, \vec{u}\rangle$, even if Billy throws ($BT = 1$), the bottle does not shatter. In our variant, the actual value assignment $BH = 0$ survives the suspension of judgment, and thus the same intervention of actuality applies. However, in contrast to the modified definition, we are not free to simply choose some actual value assignments. The interventions of actuality, so to speak, are determined by the suspension of judgment.

The modified definition implies that Billy's throw ($BT = 1$) is not an actual cause of the bottle's shattering ($BS = 1$) with respect to $\langle M, \vec{u}\rangle$. The reason is that AC2 is violated: there are no variables that can be kept fixed at their actual values such that setting BT to the value 0 results in BS taking on the value 0. In the actual context $\vec{u}$, Suzy throws, her rock hits, and so the bottle shatters whether or not Billy throws. This explanation of why we do judge Suzy's throw to be a cause of the bottle's shattering but not Billy's derives from the same intuition as in our definition: it was her rock that *actually* hit the bottle and not his. The difference is, again, that our variant gives us the pertinent piece of actuality ($BH = 0$), whereas the modified HP definition requires to check what happens when all subsets of variables are, respectively, kept at their actual values.

Let us turn our attention to the disjunctive scenario of the Arsonists Example, where each of the arsonists alone is sufficient for the forest to burn down. We show that, according to the modified definition, the conjunction $ML_1 \wedge ML_2 = 1$ qualifies as an actual cause of $FB = 1$. AC1 is satisfied, as $\langle M, \vec{u}\rangle \models ML_1 = 1 \wedge ML_2 = 1 \wedge FB = 1$. As to AC2, let $\vec{W} = \emptyset$. Consequently,

$$\langle M, \vec{u}\rangle \models [ML_1 = 0, ML_2 = 0]FB = 0 \text{ iff } \langle M_{ML_1=0,ML_2=0}, \vec{u}\rangle \models FB = 0.$$

This is satisfied, as $F_{FB}^{ML_1=0,ML_2=0} = max(ML_1, ML_2) = 0$. Interestingly, $\{ML_1, ML_2\}$ is minimal in the sense of AC3, as the two subsets $\{ML_1\}$ and $\{ML_2\}$ do not satisfy AC2. (The empty set does not even satisfy AC1.) We present the argument for $\{ML_1\}$. Let $\vec{W}$ be the $\emptyset$ or $\{ML_2\}$. According to AC2, either

$$\langle M, \vec{u}\rangle \models [ML_1 = 0]FB = 0,$$

or

$$\langle M, \vec{u}\rangle \models [ML_1 = 0, ML_2 = 1]FB = 0.$$

However, it is easy to see that neither $\langle M_{ML_1=0}, \vec{u}\rangle \models FB = 0$ nor $\langle M_{ML_1=0,ML_2=1}, \vec{u}\rangle \models FB = 0$, as $F_{FB}^{ML_1=0} = F_{FB}^{ML_1=0,ML_2=1} = max(ML_1, ML_2) = 1$. By the same reasoning pattern $\{ML_2 = 1\}$ does not satisfy AC2 due to the symmetry of the Arsonists Example.

The following tables summarize the differences between the considered definitions with respect to the Arsonists Example.

Conjunctive Scenario

Actual Cause of $FB = 1$	HP Definition	Modified HP Def.	RT Variant
$ML_1 = 1$	Yes	Yes	Yes
$ML_2 = 1$	Yes	Yes	Yes
$ML_1 = 1 \wedge ML_2 = 1$	No	No	Yes

The table compares the HP definition, the modified HP definition, and our Ramsey Test variant (RT variant) in the conjunctive scenario of the Arsonists Example. Remarkably, as shown above, Halpern's modification does not recognize the conjunction $ML_1 = 1 \wedge ML_2 = 1$ as an actual cause of $FB = 1$.

Disjunctive Scenario

Actual Cause of $FB = 1$	HP Definition	Modified HP Def.	RT Variant
$ML_1 = 1$	Yes	No	Yes
$ML_2 = 1$	Yes	No	Yes
$ML_1 = 1 \wedge ML_2 = 1$	No	Yes	Yes

The table compares the HP definition, the modified HP definition, and our Ramsey Test variant in the disjunctive scenario of the Arsonists Example. Unlike Halpern's (2015) statement, the modified definition does not agree with Halpern and Pearl's (2005) definition. This is troublesome because overdetermination is clearly a 'standard problematic example'. Furthermore, the modified definition qualifies the conjunction $ML_1 = 1 \wedge ML_2 = 1$ to be an actual cause of $FB = 1$ in the disjunctive scenario. Halpern's (2015) attempt to explain this conundrum is barely convincing. He resorts to the claim that if $\vec{X} = \vec{x}$ is an actual cause of ϕ, then each "conjunct in $\vec{X} = \vec{x}$ is called *part* of a cause of ϕ" (his emphasis). In the disjunctive scenario, however, each of $ML_1 = 1$ and $ML_2 = 1$ individually are intuitively judged to be actual causes of $FB = 1$, not merely part of one cause. This is why overdetermination is considered a major problem for Lewis's (1973c) original counterfactual account of causation.

A more general note on causal models is in order. We adopted the framework of causal models and so inherited their merits and drawbacks. The problems of spurious causation or joint effects will not arise due to the recursivity of the structural equations. Moreover, the recursivity lifts the need to obtain the directedness of causation, e.g. by assuming a temporal order.

On the other hand, models of structural equations suffer from a drawback: they encode *some* type-level causal relations that need to be antecedently given. In other words, generic knowledge about causal relations is presupposed, be it in form of sets of counterfactuals (e.g. Hitchcock (2001) and Woodward (2003)), or primitive causal mechanisms construed as law-like relationships that support a counterfactual interpretation (e.g. Pearl (2009) and Halpern and Pearl (2005)).[14]

Finally, our Ramsey Test analysis of actual causation is characterized by an epistemic interpretation of causal models. That is, we take the structural equations to express beliefs about elementary causal dependences. While for Halpern and Hitchcock (2015, p. 384) the structural equations may be seen as "describing objective features of the world", they agree that "judgments of actual causation are subjective". Note, furthermore, that their 'reasoning about causality' is suspiciously close to our 'judgments on causal relations'. "[U]ltimately", so Halpern and Pearl (2005, p. 878), "the choice of model is a subjective one". This is in line with the present Ramsey Test definition being relative

[14]This being said, the proponents of models of structural equations do not find the presupposition of type-level causal relations all too problematic. For causal models do not directly represent relations of actual causation. Rather, they encode generic knowledge about causal structures in terms of which actual causes can be identified, whereas the generic causal relations are primitives that cannot be further analysed in terms of actual causation. For a more detailed justification, see Halpern and Pearl (2005), Halpern and Hitchcock (2010), and Fenton-Glynn (2017).

to an epistemic state. Hence, nothing seems to preclude an interpretation of the model-relativity as agent-relativity.

5.6 Conclusion

We have proposed a Ramsey Test definition of actual causation in the framework of causal models. For this definition cases of overdetermination, conjunctive scenarios, and preemption are no problem. With respect to these standard examples, we have shown that the definition surpasses the Halpern-Pearl definitions. Underlying this claim are criteria that Halpern and Pearl (2005, p. 846) suggest for evaluating approaches of actual causation:

> The best ways to judge the adequacy of an approach are the intuitive appeal of the definitions and how well it deals with examples; we believe that this article shows that our approach fares well on both counts.

We believe that the present article has shown that our variant to the Halpern-Pearl definitions fares even better with respect to overdetermination and conjunctive scenarios.

Of course, we need to deal with the other problematic examples of the literature in order to properly compete with the Halpern-Pearl definitions. Fortunately, this research is already in progress, including a treatment of double prevention, switches, and omissions. Moreover, we will provide a more detailed comparison between the agnostic dependence used in our strengthened Ramsey Test and the notion of counterfactual dependence.

In the next chapter, we stay with causal models à la Halpern and Pearl (2005). In the tradition of Lewis (1986b), they exclude categorically that causes can be disjunctive events. More specifically, disjunctive events are excluded from being causes by Halpern and Pearl's (2005) definition: they require each candidate cause to have the form of a conjunction of primitive events. Lewis (1986b, p. 212) excludes disjunctive events from being candidates for actual causes simply because he does not know "how a genuine event could be the disjunction of two events both of which actually occur." Following this metaphysical verdict, Halpern and Pearl (2005, p. 853) state that the "only reasonable definition of 'A or B causes ϕ' seems to be that 'either A causes ϕ or B causes ϕ'. There are no truly disjunctive causes once all the relevant facts are known." Sartorio (2006) challenges the view that 'there exist no truly disjunctive causes when all the relevant facts are known'. She puts forth a switching scenario, in which 'the disjunction of two events both of which actually occur' plausibly seems to be an actual cause. We will work out how we can understand disjunctive causes in the sense of Sartorio's switch. Subsequently, we use this insight to extend Halpern and Pearl's (2005) definition of actual causation such that the definition covers this type of disjunctive causes.

Chapter 6

A Refined Halpern-Pearl Definition for Sartorio's Disjunctive Causes

In the previous chapter, we proposed an alternative to the Halpern-Pearl definitions of actual causation (Halpern and Pearl (2005); Halpern (2015)). We think that our variant has the potential to become a strong competitor to the Halpern-Pearl definitions when we refine it further. In general, however, it seems that the definition provided by Halpern and Pearl (2005) is currently still the benchmark for conditional analyses of causation in the tradition of Lewis (1973c). As is well known, Lewis (1986b, p. 212) does "not know how a genuine event could be the disjunction of two events both of which actually occur." His reasons for this statement derive from his metaphysics. More recently, Sartorio (2006) has shed doubt on this implication of Lewis's metaphysics. She argues for the existence of disjunctive causes by presenting an example of such. Now, for Halpern and Pearl's (2005) definition, disjunctions of primitive events are not even admissible candidates for actual causes. The reason for this exclusion, so Halpern and Pearl (2005, p. 853), be that if a disjunction is a cause at least one of its disjuncts must be a cause – we just do not know which one. Hence, "there be no truly disjunctive causes once all the relevant facts are known." (ibid.)

In the light of Sartorio's example of a disjunctive cause, we think Lewis's metaphysical reasons to exclude disjunctive causes might be somewhat premature. Hence, we aim in this chapter to refine Halpern and Pearl's (2005) definition of actual causation such that it allows for disjunctive causes of the type found in Sartorio's example. In order to treat such disjunctive causes within Halpern and Pearl's framework of causal models, we first extend their causal model semantics by disjunctive antecedents. Based on the extension, we will show that our refined Halpern-Pearl definition aligns with Sartorio's (2006) observation: a disjunctive cause does not imply that one of its disjuncts *must* also be a cause.

Sources. This chapter builds on Günther (2017b). Substantial content from this contribution to the *Proceedings of the 21st Amsterdam Colloquium* is reprinted by permission from Alexandre Cremers, Thom van Gessel, and Floris Roelofsen (editors): Institute of Logic, Language and Computation (ILLC), University of Amsterdam (2017).

6.1 Introduction

Halpern and Pearl (2005) define actual causation based on a causal model semantics of conditionals. The semantics is restricted to antecedents that do not contain disjunctions. "We might consider generalizing further to allow disjunctive causes", so Halpern and Pearl (2005, p. 853), but they discard the idea, because there be "no truly disjunctive causes once all the relevant facts are known".

Making no use of formal means such as causal models, Sartorio (2006) argues for the existence of disjunctive causes. She does so by putting forward a switching scenario, in which all the relevant facts are known. Sartorio's Switch provides motivation for Halpern and Pearl's (2005) definition of actual causation to be applicable to causes that have a particular disjunctive form. Hence, we first need to lift the restriction of causal models to non-disjunctive antecedents such that we can express arbitrary Boolean combinations in a conditional's antecedent.

In Section 6.2, we translate Sartorio's Switch into a causal model. En passant we introduce Halpern and Pearl's (2005) causal model semantics and definition of actual causation. In Section 6.3, we extend their causal model semantics by antecedents having a disjunctive form. This allows us to refine Halpern and Pearl's definition of actual causation such that it captures disjunctive causes of the type found in Sartorio's Switch.

6.2 Sartorio's Switch and Causal Models

Sartorio (2006) argues for the existence of disjunctive causes. She invokes roughly the following scenario to back up her claim.

Example 11. Sartorio's Switch (Sartorio (2006, p. 523–528))
Suppose a person is tied on a track and a train is running towards the person. Although the track branches and there is a switch determining on which of two tracks the train continues, the tracks reconverge before the place, where the person is captivated. Now, Sartorio adds details to this typical switching scenario. A person, called Flipper, flips the switch such that the train continues on the left track. Moreover, there is construction work carried out on the right track. Another person, called Reconnecter, reconnects the right track before the train would have arrived in case Flipper hadn't flipped the switch. The train travels on the left track and kills the trapped person.

Sartorio proposes that the disjunction 'Flipper flips the switch and/or Reconnecter reconnects' is the actual cause of the person's death, while both individually 'Flipper flips the switch' and 'Reconnecter reconnects' are not actual causes of the person's death.[1]

In her judgment, she complies with the simple counterfactual analysis of actual causation: event C is a cause of event E iff C and E occur, and E counterfactually depends on C, that is, if it were not the case that C occurs, E would not have occurred. Thereby, she takes counterfactual dependence to be a prima facie reason for and thus a proxy for causation.[2] Accordingly, 'Flipper flips the switch' (and 'Reconnecter reconnects') is not an actual cause of the person's death. For, if it were not the

[1] In this chapter, we are silent on the issue of whether or not actual causes are events or facts. We just assume that if there are disjunctive causes, they are events and/or facts. However, see Lewis (1986b, pp. 241-69) for an argument challenging the existence of disjunctive events, and thus disjunctive event causation.

[2] See Sartorio (2006, p. 529). Most recent counterfactual accounts of causation say that counterfactual dependence is sufficient for causation, but not necessary. See, for instance, Hitchcock (2001), Woodward (2003), Hall (2004), Hall (2007), Halpern and Pearl (2005), Halpern (2015). Hence, no counterfactual dependence between C and E seems to allow for causation between C and E, whereas the presence of counterfactual dependence seems to be a very strong reason for the presence of a causal relation.

case that 'Flipper flips the switch' (or 'Reconnecter reconnects' respectively), the person would die nevertheless. Additionally, the conjunction 'Flipper flips the switch and Reconnecter reconnects' is no actual cause of the person's death. For, if it were not the case, the person might die nevertheless, viz. in case one of Flipper and Reconnecter does what they do. However, the disjunction 'Flipper flips the switch or Reconnecter reconnects' is an actual cause of the person's death. For, if it were not the case, the person would not die. Sartorio (2006, p. 530) confirms that "the death happened because *at least one of them* did what they did."

Counterfactual accounts cash out the idea that a cause makes a difference.[3] In her argumentation for disjunctive causes, Sartorio endorses a principle of difference making for which she argued in Sartorio (2005): if an event is a cause of an effect, then its absence would not have been a cause of the same effect. Relying on this principle, she argues that Flipper's redirection is not a cause given that there was an alternative route that would have led to the same outcome (even though this route is not actualized). There is no net difference between flipping and not flipping the switch with respect to the person's death. Hence, the flipping fails to make a difference in the sense of the principle: if we were to count the flipping as a cause, we would have to count its absence as a cause too. Therefore, Sartorio (2006, p. 532) thinks that "the mere fact that there was an alternative route is sufficient to rob the event of the redirection of its causal powers." Flipper's redirection to the left track renders Reconnecters reconnection of the right track causally inefficacious, and, conversely, the reconnection renders the redirection causally inefficacious. The core of her reasoning goes as follows: "If either event had happened without the other, then that event would have been causally efficacious [...]. But, when both events happen, they deprive each other of causal efficacy." (ibid., p. 531) However, so argues Sartorio, the outcome must still depend on the existence of some viable causally efficacious path. Hence, the disjunctive fact that at least one path was causally efficacious is the cause of the outcome.

We translate now Sartorio's Switch into a causal model and check which formulas qualify as actual causes according to Halpern and Pearl's (2005) definition of actual causation.

6.2.1 Halpern and Pearl's Causal Model Semantics

Halpern and Pearl's (2005) causal model semantics of conditionals is defined with respect to a causal model over a signature.

Definition 30. Signature
A signature $\mathcal{S}$ is a triple $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$, where $\mathcal{U}$ is a finite set of exogenous variables, $\mathcal{V}$ is a finite set of endogenous variables, and $\mathcal{R}$ maps any variable $Y \in \mathcal{U} \cup \mathcal{V}$ on a non-empty (but finite) set $\mathcal{R}(Y)$ of possible values for Y.

Definition 31. Causal Model
A causal model over signature $\mathcal{S}$ is a tuple $M = \langle \mathcal{S}, \mathcal{F} \rangle$, where $\mathcal{F}$ maps each endogenous variable $X \in \mathcal{V}$ on a function $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} \setminus \{X\}} \mathcal{R}(Y)) \mapsto \mathcal{R}(X)$.

The mapping $\mathcal{F}$ defines a set of (modifiable) structural equations modeling the causal influence of exogenous and endogenous variables on other endogenous variables. The function F_X determines the value of $X \in \mathcal{V}$ given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. Note that $\mathcal{F}$ defines no structural equation for any exogenous variable $U \in \mathcal{U}$. We write $\vec{X}$ for a (finite) vector of variables $X_1, .., X_n$, and $\vec{x}$ for a (finite) vector of values $x_1, .., x_n$ of the variables. Hence, we abbreviate $X_1 = x_1, .., X_n = x_n$ by $\vec{X} = \vec{x}$.

[3] Lewis (1973c, p. 557), for example, writes "We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it.".

A structural equation such as $x = F_X(\vec{u}, y)$ says (in a context where the exogenous variables take the values $\vec{u}$), if Y *were* set to y (by means of an intervention), then X *would* take on the value x. Notice the difference to a direct intervention on X: an intervention that assigns a value $x' \neq x$ to X (by means external to the model) overrules the value x assigned by $F_X(\vec{u}, y)$. The difference points to a 'causal' asymmetry. In structural equations, the values of the variables on the right-hand side cannot be changed without changing the values of the variables on the left-hand side; but the values of the variables on the left-hand side can be changed (by means of an external intervention) without changing the values of the variables on the right-hand side.

Halpern and Pearl (2005, p. 849) confine themselves to models of recursive structural equations.[4] In such models, the causal dependences among the variables in $\mathcal{V}$ can be represented by a directed acyclic graph. Hence, there is a strict partial order such that, if $X \prec Y$, then the value of X may affect the value of Y, but the value of Y cannot have any effect on the value of X. A consequence of the restriction to models M of recursive equations is that, given a context $\vec{U} = \vec{u}$, there is always a unique solution to the equations in M; for we can always solve the equations in the order given by $\prec$.

In the spirit of Halpern and Hitchcock (2010, p. 397), we make two assumptions concerning the variables of a causal model: (i) no value of a variable X logically implies a value of another variable Y, and (ii) the different values of the same variable are mutually exclusive and jointly exhaustive.

Intuitively, a simple conditional $[Y = y]X = x$ is true in a causal model M given context $\vec{u} = u_1, ..., u_n$, if the intervention setting $Y = y$ results in the solution $X = x$ for the structural equations.[5] Such an intervention is defined by the notion of a submodel $M_{Y=y}$ of M.

Definition 32. Submodel
Let $M = \langle \mathcal{S}, \mathcal{F} \rangle$ be a causal model, $\vec{X}$ a (possibly empty) vector of variables in $\mathcal{V}$ and $\vec{x}, \vec{u}$ vectors of values for the variables in $\vec{X}, \vec{U}$. We call the causal model $M_{\vec{X}=\vec{x}} = \langle \mathcal{S}_{\vec{X}}, \mathcal{F}^{\vec{X}=\vec{x}} \rangle$ over signature $\mathcal{S}_{\vec{X}} = \langle \mathcal{U}, \mathcal{V} \setminus \vec{X}, \mathcal{R}\big|_{\mathcal{V}\setminus\vec{X}} \rangle$ a submodel of M. $\mathcal{F}^{\vec{X}=\vec{x}}$ maps each variable in $\mathcal{V} \setminus \vec{X}$ on a function $F_Y^{\vec{X}=\vec{x}}$ that corresponds to F_Y for the variables in $\mathcal{V} \setminus \vec{X}$ and sets the variables in $\vec{X}$ to $\vec{x}$.

We can describe the structure of Sartorio's Switch using a causal model including four binary variables:

- an exogenous variable T, where $T = 1$ if the train arrives and $T = 0$ otherwise;
- an endogenous variable F, where $F = 1$ if Flipper flips the switch and $F = 0$ otherwise;
- an endogenous variable R, where $R = 1$ if Reconnecter reconnects and $R = 0$ otherwise;
- an endogenous variable D, where $D = 1$ if the person dies and $D = 0$ otherwise.

Leaving the functions F_F, F_R and F_D implicit, the set of structural equations is given by:

- $F = T$
- $R = T$

[4] This being said, (Halpern and Pearl, 2005, pp. 883-884) provide a definition of actual causation for non-recursive models in their appendix.

[5] We follow Fenton-Glynn (2017) in conveying "=" a double role: the sign figures as (i) identity and (ii) an assignment operator such as Halpern and Pearl's (2005) "←" or Pearl's (2009) "do(·)". Although this double role is mathematically sloppy, it avoids an unnecessary multiplication of notation and is harmless as long as we keep in mind that $[\vec{Y} = \vec{y}]\phi$ expresses a conditional whose antecedent assigns the variables in $\vec{Y}$ the values $\vec{y}$.

- $D = max(F, R)$

In words, Flipper flips the switch ($F = 1$), if the train arrives ($T = 1$). Reconnecter reconnects ($R = 1$), if the train arrives. The person dies ($D = 1$), if at least one of $F = 1$ and $R = 1$ is the case.

Causal networks can be used to depict the recursive dependences of the structural equations. A causal network is a directed acyclic graph, where the nodes correspond to the variables and there is an arrow from a node labeled X to one labeled Y iff X is a parent variable of Y iff F_Y depends on the value of X iff $X \prec Y$.[6] Analogous to the asymmetry of structural equations, causal networks have the property that variables only causally affect their descendants: if Y is not a descendant of X, then a change in the value of X has no effect on the value of Y.

The recursive dependencies of the structural equations are depicted in Figure 6.1.

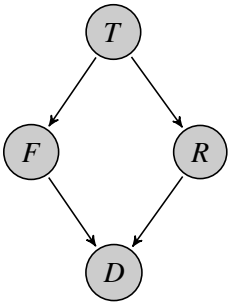

Figure 6.1: The causal network for Sartorio's Switch.

To illustrate the causal model semantics, let us check whether or not the conditional $[F = 1]D = 1$ is true in the causal model M of Sartorio's Switch (given the context $t = 1$). Intuitively, the intervention that sets $F = 1$ induces a submodel $M_{F=1}$ of M. If the solution to the structural equations of $M_{F=1}$ satisfies $D = 1$, then $[F = 1]D = 1$ is true in the causal model M under context $T = t$. In this case, we write $\langle M, t\rangle \models [F = 1]D = 1$.

In the scenario of Sartorio's Switch, $\langle M, t\rangle \models [F = 1]D = 1$ iff $\langle M_{F=1}, t\rangle \models D = 1$. The structural equations for the submodel $M_{F=1}$ are:

- $F = 1$
- $R = T$
- $D = max(F, R)$

We see that the solution to the structural equations of $M_{F=1}$ satisfies $D = 1$, and thus M satisfies the conditional $[F = 1]D = 1$ (given t). Notice the difference between the structural equation $F = T$ and $F = 1$: the former depends on T, whereas the latter does not. After the intervention that sets $F = 1$, the variable F is treated similar to an exogenous variable, i. e. it is assigned a value by its structural equation that does not depend on other (exogenous and/or endogenous parent) variables.[7] The structural equations for the variables in $\mathcal{V} \setminus \{F\}$ remain unchanged.

[6] Note that the acyclicity follows from the assumption that the structural equations are recursive.

[7] Intuitively, we may think of a value assignment $X = x$ in model M by an intervention as overruling the structural equation in M.

6.2.2 Halpern and Pearl's Definition of Actual Causation

The basic idea behind Halpern and Pearl's (2005) definition of actual causation is to extend Lewis's notion of causal dependence to a notion of contingent dependence. The simple counterfactual analysis mentioned above is Lewis's (1973c) definition of causal dependence: event E causally depends on event C iff (i) C and E occur, and (ii) E counterfactually depends on C. Furthermore, he identifies actual causation with the transitive closure of causal dependence. Hence, C is an actual cause of E iff there is a chain of causal dependencies from C to E. Halpern and Pearl extend this definition by (possibly non-actual) contingencies: C is an actual cause of E iff E causally depends on C *under certain contingencies*. Roughly, contingent dependence makes it possible that even if E does not counterfactually depend on C in the actual situation, E counterfactually depends on C under certain contingencies.[8]

Based on their causal model semantics for conditionals, Halpern and Pearl (2005, p. 853) propose the following definition of actual causation.

Definition 33. Actual Causation
$\vec{X} = \vec{x}$ is an actual cause of ϕ in $\langle M, \vec{u}\rangle$ iff the following three conditions hold:

AC1. $\langle M, \vec{u}\rangle \models (\vec{X} = \vec{x}) \wedge \phi$.

AC2. There exists a partition $\langle \vec{Z}, \vec{W}\rangle$ of $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $\langle \vec{x'}, \vec{w'}\rangle$ of the variables in $\langle \vec{X}, \vec{W}\rangle$ such that if $\langle M, \vec{u}\rangle \models Z = z^*$ for all $Z \in \vec{Z}$, then both of the following conditions hold:

(a) $\langle M, \vec{u}\rangle \models [\vec{X} = \vec{x'}, \vec{W} = \vec{w'}]\neg\phi$.

(b) $\langle M, \vec{u}\rangle \models [\vec{X} = \vec{x}, \vec{W'} = \vec{w'}, \vec{Z'} = \vec{z^*}]\phi$ for all subsets $\vec{W'}$ of $\vec{W}$ and all subsets $\vec{Z'}$ of $\vec{Z}$.

AC3. $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions AC1 and AC2.

AC1 requires both that the actual cause $\vec{X} = \vec{x}$ and its effect ϕ are true in the actual (contextualized) model. AC3 ensures that only the conjuncts of $\vec{X} = \vec{x}$ "essential" for changing ϕ in AC2(a) are part of a cause, or as Halpern and Pearl (2005, p. 853) say "inessential elements are pruned." To understand AC2, it is helpful to think of $\vec{X} = \vec{x}$ as the minimal set of conjuncts that counts as a cause of the effect ϕ, and to think of $\vec{Z} = \vec{z}$ as the active causal path(s) from $\vec{X}$ to ϕ.

AC2(a) is reminiscent of Lewis's (1973c) counterfactual dependence: ϕ would be false, if it were not for $\vec{X} = \vec{x}$. The condition says that there is a setting $\vec{X} = \vec{x'}$ changing ϕ to $\neg\phi$, if the variables not on the active causal path(s) take on certain values, i. e. $\vec{W} = \vec{w'}$. The difference to the counterfactual criterion is that ϕ's dependence on $\vec{X} = \vec{x}$ may be tested under certain contingencies $\vec{W} = \vec{w'}$, which are non-actual for $\vec{w'} \neq \vec{w}$. Note that those contingent 'but-for-tests' allow to identify more causal relationships than the simple counterfactual criterion.

AC2(b) restricts the contingencies allowed to be considered. The idea is that any considered contingency does not affect the active causal path(s) with respect to $\vec{X} = \vec{x}$ and ϕ. In other words, AC2(b) guarantees that $\vec{X}$ alone is sufficient to change ϕ to $\neg\phi$. The setting of a contingency $\vec{W} = \vec{w'}$ only eliminates spurious side effects that may hide $\vec{X}'s$ effect. The idea behind AC2(b) is implemented as follows: (i) setting a contingency $\vec{W} = \vec{w'}$ leaves the causal path(s) unaffected by the condition that changing the values of any subset $\vec{W'}$ of $\vec{W}$ from the actual values $\vec{w}$ to the contingent values $\vec{w'}$ has no effect on ϕ's value. (ii) At the same time, changing the values of $\vec{W'}$ may alter the values of the variables in $\vec{Z}$, but this alteration has no effect on ϕ's value.

[8] Note that Halpern and Pearl do not take the transitive closure for their definition of actual causation. In contrast to Lewis's dictum, they think (Halpern and Pearl, 2005, p. 844) that causation is not always transitive.

We apply now Halpern and Pearl's (2005) definition of actual causation to the causal model of Sartorio's Switch. The result is that each of $F = 1$ and $R = 1$ is an actual cause of $D = 1$. However, the conjunction $F = 1 \wedge R = 1$ and the disjunction $F = 1 \vee R = 1$ do not qualify as actual causes of $D = 1$.

We show that $F = 1$ is an actual cause of $D = 1$. (The argument for $R = 1$ is structurally the same as the causal model of Sartorio's Switch is symmetric with respect to F and R.) Let $\vec{Z} = \{F, D\}$, and so $\vec{W} = \{R\}$. The contingency $R = 0$ satisfies the two conditions of AC2: AC2(a) is satisfied, as setting $F = 0$ results in $D = 0$; AC2(b) is satisfied, as setting F back to 1 results in $D = 1$. The counterfactual contingency $R = 0$ is required to reveal the hidden dependence of D on F, or so argue Halpern and Pearl.

The conjunction $F = 1 \wedge R = 1$ is no actual cause of $D = 1$ due to the minimality condition AC3. Let $\vec{Z} = \{F, R, D\}$, and so $\vec{W} = \emptyset$. AC2(a) is satisfied, as setting $F = 0 \wedge R = 0$ results in $D = 0$. AC2(b) is satisfied trivially. However, two subsets of $\vec{X} = \{F, R\}$ satisfy the two conditions of AC2 as well, viz. $\vec{X'} = \{F\}$ and $\vec{X''} = \{R\}$. Therefore, $\vec{X} = \{F, R\}$ is not minimal and according to AC3 the conjunction $F = 1 \wedge R = 1$ is thus no actual cause of $D = 1$. Minimality is meant, so Halpern and Pearl (2005, p. 857), to strip "overspecific details from the cause."

The disjunction $F = 1 \vee R = 1$ does not count as actual cause of $D = 1$, simply because Halpern and Pearl's (2005) definition of actual causation does not admit causes in form of proper disjunctions, i. e. disjunctions having more than one disjunct. They do not "have a strong intuition as to the best way to deal with disjunction in the context of causality and believe that disallowing it is reasonably consistent with intuitions." (p. 858)

Sartorio (2006, p. 530) observes that "there is no general motivation for believing that, when (if) a disjunctive fact is a cause, at least one of its disjuncts must also be a cause." In other words, it is possible that a disjunctive fact is a cause while its disjuncts are not. This observation stands in sharp contrast to Halpern and Pearl's (2005) definition of actual causation, according to which both 'disjuncts' individually can qualify as actual causes, whereas the disjunction cannot. In the next section, we first define disjunctive antecedents for Halpern and Pearl's causal model semantics; subsequently, we extend their definition of actual causation to cover disjunctive causes as found in Sartorio's Switch.

6.3 An Extension of Causal Model Semantics by Disjunctive Antecedents

Recall Sartorio's Switch of Section 6.2. Sartorio argues that the person tied to the tracks dies because at least one of Flipper and Reconnecter does what they do. Therefore, the disjunctive fact that at least one track or path was causally efficacious is the cause of the outcome. Moreover, if only one of Flipper's and Reconnecter's events would occur, their *disjunction* would be causally inefficacious, but the single occuring event would be causally efficacious. We identify here two necessary conditions under which there are disjunctive causes: (i) there are at least two actually occuring and potentially efficacious events on different paths ("two tracks"), and (ii) there is an event that switches the paths without making a net difference with respect to the outcome ("flipping the switch").

Let us consider Sartorio's Switch using the variables of our causal model. In her switching scenario, Sartorio maintains that the disjunction $F = 1 \vee R = 1$ is an actual cause of $D = 1$. Recalling the necessary conditions for disjunctive causes, Sartorio's disjunctive conditional $[F = 1 \vee R = 1]D = 1$ means:

- the actual case $F = 1$ and $R = 1$ results in $D = 1$, *and*

- the counterfactual case $F = 1$ and $R = 0$ results in $D = 1$, *and*
- the counterfactual case $F = 0$ and $R = 1$ results in $D = 1$.

Disjunctive causation requires here that both of $F = 1$ and $R = 1$ are actual, and each of $F = 1$ and $R = 1$ alone would be sufficient to result in $D = 1$. In such a case, if one *or* the other is sufficient for the effect and both occur, then Sartorio judges the disjunction of both to be the cause. In this sense, Sartorio understands the disjunctive antecedent $F = 1 \vee R = 1$ as a summary of two actually occuring events $F = 1$ and $R = 1$, whose actual co-occurrence robs them of their individual causal efficacy. Moreover, in a context, in which only one of them is actual, this one individually would be an actual cause.

Halpern and Pearl's (2005) causal model semantics does not allow to evaluate the conditional $[F = 1 \vee R = 1]D = 1$. The reason is that the submodel $M_{F=1\vee R=1}$ is undefined, and thus they cannot evaluate disjunctions in the antecedent. Moreover, the structural equation for D of Sartorio's Switch does apply to values of F and R, but it does not apply to a disjunction such as $F = 1 \vee R = 1$. Hence, the value for D is not determined by the disjunction. Next, we propose a conservative extension of Halpern and Pearl's causal model semantics that allows us to evaluate antecedents that are disjunctive in Sartorio's sense.

6.3.1 Evaluating Disjunctive Antecedents

As we have just observed, Sartorio's disjunctive causes of the form $A = a \vee B = b$ require that the Boolean conjunction $A = a \wedge B = b$ actually obtains, and if one of $A = a$ or $B = b$ would obtain but not the other, the effect would still follow. We implement now this logic governing Sartorio's disjunctive causes by extending Halpern and Pearl's (2005) framework of causal models.

The idea behind evaluating a conditional with disjunctive antecedent is to check whether the consequent is true in *each* disjunctive situation of the antecedent. A disjunction $A = a \vee B = b$ is true in three possible situations:

(i) $A = a \wedge B = b$,

(ii) $A = a \wedge B = \neg b$, and

(iii) $A = \neg a \wedge B = b$.

We refer to (i)-(iii) as the disjunctive situations or possibilities of the formula $A = a \vee B = b$. We evaluate the antecedent $[A = a \vee B = b]$, for example, by carrying out three interventions rather than one. Each of these interventions corresponds to a disjunctive situation. The result is exactly one submodel per disjunctive situation.

In symbols, the idea behind evaluating a conditional with disjunctive antecedent can be expressed as follows:

$$\begin{aligned} &\langle M, \vec{u}\rangle \models [A = a \vee B = b]C \text{ iff} \\ &\langle M_{A,B}, \vec{u}\rangle \models C, \text{ and} \\ &\langle M_{A,\neg B}, \vec{u}\rangle \models C, \text{ and} \\ &\langle M_{\neg A,B}, \vec{u}\rangle \models C. \end{aligned} \tag{1}$$

Given $A = a \vee B = b$ is an actual disjunctive cause of C, the intervention corresponding to disjunctive situation (i) results in the submodel $M_{A=a,B=b}$, in which A and B take the same values as in the actual contextualized model $\langle M, \vec{u}\rangle$.

Our strategy to evaluate a conditional with disjunctive antecedent does not require to modify Halpern and Pearl's (2005) notion of a submodel. Rather, the evaluation requires to look at (possibly) more than one submodel, namely at exactly one submodel for each disjunctive situation. In general, we write ϕ_i, where $1 \leq i \leq n$, for the formula that expresses the i-th disjunctive situation of the formula ϕ (that contains only finitely many primitive events).[9]

For clarity, we define an extended causal language.

Definition 34. Extended Causal Language $\mathcal{L}$
The extended causal language $\mathcal{L}$ contains

- the two propositional constants $\top$ and $\bot$,
- a finite number of random variables $\vec{X} = X_1, ..., X_n$ associated with finite ranges $\mathcal{R}(X_1), ..., \mathcal{R}(X_n)$,
- the Boolean connectives $\wedge, \vee, \neg$ and the operator [], and
- left and right parentheses.

A formula ϕ of $\mathcal{L}$ is well-formed iff ϕ has the form

- $X = x$ for $x \in \mathcal{R}(X)$ (primitive event);
- if $\phi, \psi \in \mathcal{L}$, then $\neg\phi, \phi \wedge \psi, \phi \vee \psi \in \mathcal{L}$ (Boolean combinations of primitive events);
- if [] does not occur in $\phi, \psi \in \mathcal{L}$, then $[\phi]\psi \in \mathcal{L}$ (conditionals).

For the extended causal language, we define a valuation function. Recall that $\langle M, \vec{u}\rangle \models X = x$ is shorthand for $X = x$ is the solution to all of the structural equations in the recursive model M given context $\vec{u}$.

Definition 35. Valuation Function
A valuation function $v_{\langle M,\vec{u}\rangle}$ (abbreviated as v) is associated with any arbitrary model M and any arbitrary vector $\vec{u}$. $v_{\langle M,\vec{u}\rangle} : \mathcal{L} \mapsto \{1, 0\}$ assigns either 1 or 0 to all formulas of the extended causal language $\mathcal{L}$:

(a) $v(X = x) = \begin{cases} 1, & \text{if } \langle M, \vec{u}\rangle \models X = x \\ 0, & \text{otherwise} \end{cases}$

(b) $v(\neg\phi) = 1$ iff $v(\phi) = 0$

(c) $v(\phi \wedge \psi) = 1$ iff $v(\phi) = 1$ and $v(\psi) = 1$

(d) $v(\phi \vee \psi) = 1$ iff $v(\phi) = 1$ or $v(\psi) = 1$

(e) $v([\phi]\psi) = \begin{cases} 1, & \text{if } v(\psi) = 1 \text{ in each } \langle M_{\phi_i}, \vec{u}\rangle \\ 0, & \text{otherwise} \end{cases}$

, where M_{ϕ_i} is a submodel of M such that $\langle M, \vec{u}\rangle \models [\phi_i]\psi$, and ϕ_i is a non-disjunctive formula expressing one disjunctive possibility of ϕ.

[9] Note that the number i of ϕ_i depends on the number d of disjunctions occurring in ϕ. In general, $i \leq 2^{d+1} - 1$, i. e. there are at most $2^{d+1} - 1$ disjunctive situations of ϕ. When the disjuncts are mutually exclusive, there are less disjunctive situations, because some are impossible. Take for example $F = 1 \vee F = 0$ for the binary variable F. Here, there are only two disjunctive situations, because $F = 1 \wedge F = 0$ is impossible. The reason is, for any Boolean valuation function v, if $v(F = 1) = 1$ then $v(F = 0) = 0$ and if $v(F = 0) = 1$ then $v(F = 1) = 0$.

Clause (c) of the valuation function entails that $X_1 = x_1, .., X_n = x_n$ is the setting of the variables in the contextualized model $\langle M, \vec{u}\rangle$ iff $X_1 = x_1 \wedge ... \wedge X_n = x_n$ is true in $\langle M, \vec{u}\rangle$. Hence, a vector of primitive events $\vec{X} = \vec{x}$ corresponds to a conjunction of those primitive events $\bigwedge X_i = x_i$ for $1 \leq i \leq n$.

Let us now evaluate a conditional with disjunctive antecedent in the causal model of Sartorio's Switch. We check whether or not $\langle M, t = 1\rangle \models [F = 1 \vee R = 1]D = 1$. Let ϕ_1, ϕ_2, ϕ_3 express the disjunctive situations of $F = 1 \vee R = 1$. According to clause (e), we need to check whether $v(D = 1) = 1$ in each $\langle M_{\phi_i}, t = 1\rangle$ for $1 \leq i \leq 3$. Figure 6.2 depicts the causal network of the submodel M_{ϕ_1} for the disjunctive situation ϕ_1. M_{ϕ_2} and M_{ϕ_3} look the same for $\phi_2 = (F = 1) \wedge (R = 0)$ and $\phi_3 = (F = 0) \wedge (R = 1)$.

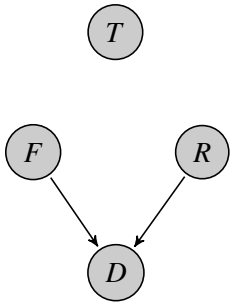

Figure 6.2: The causal network of M_{ϕ_1} for $\phi_1 = (F = 1) \wedge (R = 1)$.

As $D = max\{F, R\}$ remains unchanged in each $\langle M_{\phi_i}, t = 1\rangle$, we obtain for the three submodels:

(i) $\langle M_{\phi_1}, t = 1\rangle \models D = 1$

(ii) $\langle M_{\phi_2}, t = 1\rangle \models D = 1$

(iii) $\langle M_{\phi_3}, t = 1\rangle \models D = 1$

Hence, $v(D = 1) = 1$ in each $\langle M_{\phi_i}, t = 1\rangle$, and thus the model M satisfies the conditional $[F = 1 \vee R = 1]D = 1$ in context $t = 1$.[10]

6.3.2 A Refinement of Halpern and Pearl's Definition of Actual Causation

Now that we can evaluate disjunctive antecedents in extended causal models, we propose a refinement or amendment of Halpern and Pearl's (2005) definition of actual causation. As we have observed at the beginning of Section 6.3, Sartorio's disjunctive causes of the form $A = a \vee B = b$ require that the conjunction $A = a \wedge B = b$ actually obtains, and if one of $A = a$ or $B = b$ would obtain but not the other, the effect would still follow.

Definition 36. Actual Causation Refined
Let $\psi_{\vee_i} = (\vec{X}_i = \vec{x}_i)$ denote the i-th disjunct of the finite formula ψ which is in disjunctive normal form.[11] ψ is an actual cause of ϕ in $\langle M, \vec{u}\rangle$ iff the following three conditions hold:

[10] Our extended semantics may also be used to express incomplete information. Even if we only know that one of $F = 1$ and $R = 1$, but not which one, we can use the disjunction $F = 1 \vee R = 1$ of our semantics to obtain useful information as to whether or not the person dies. Here, however, we are not interested in incomplete information.

[11] The restriction of ψ to formulas of disjunctive normal form is no proper restriction, since all arbitrary Boolean formulas can be converted into an equivalent disjunctive normal form. See, e.g., Papadimitriou (1994, pp. 75-76).

AC1R. $\langle M, \vec{u}\rangle \models (\bigwedge \psi_{\vee_i}) \wedge \phi$ for all i.

AC2R. There exists a partition $\langle \vec{Z}, \vec{W}\rangle$ of $\mathcal{V}$ with $\vec{X}_i \subseteq \vec{Z}$ and some setting $\langle \vec{x'}_i, \vec{w'}\rangle$ of the variables in $\langle \vec{X}_i, \vec{W}\rangle$ such that if $\langle M, \vec{u}\rangle \models Z = z^*$ for all $Z \in \vec{Z}$, then both of the following conditions hold:

(a) $\langle M, \vec{u}\rangle \models [\bigwedge \vec{X}_i = \vec{x'}_i, \vec{W} = \vec{w'}]\neg\phi$ for all i.

(b) $\langle M, \vec{u}\rangle \models [\bigvee \vec{X}_i = \vec{x}_i, \vec{W'} = \vec{w'}, \vec{Z'} = \vec{z^*}]\phi$ for all subsets $\vec{W'}$ of $\vec{W}$ and all subsets $\vec{Z'}$ of $\vec{Z}$, and for all i.

AC3R. ψ is minimal; no subsets of the disjuncts $\psi_{\vee_i} = (\vec{X}_i = \vec{x}_i)$ satisfy conditions AC1R and AC2R, and no disjunction of the form $\bigvee(\vec{X}_i = \vec{x}_i) \vee \vec{Y} = \vec{y}$ with $\vec{Y} \subseteq \vec{Z}$, $\vec{Y} \cap \vec{X}_i = \emptyset$ (for all i) and $\vec{Y} \neq \phi$ satisfies AC1R and AC2R.

AC1R requires that each disjunct of the actual cause ψ and its effect ϕ are true in the actual contextualized model. Note that this is equivalent to the big *conjunction* of all disjuncts $\phi_{\vee_i}$ and the effect ϕ being true in the actual contextualized model. (The need for the big conjunction directly follows from Sartorio's first condition necessary for disjunctive causes.)

AC2R requires that each disjunct $\psi_{\vee_i} = (\vec{X}_i = \vec{x}_i)$ of ψ satisfies AC2. That is: (a) setting $\vec{X}_i = \vec{x'}_i$ (for all i) changes ϕ to $\neg\phi$, if the variables $\vec{W}$ not on the active causal path(s) take on certain values; (b) guarantees that the disjunction $\bigvee(\vec{X}_i = \vec{x}_i)$ alone is sufficient to change $\neg\phi$ to ϕ. Note that AC2R(b) is quite demanding: setting $\bigvee(\vec{X}_i = \vec{x}_i)$ results in one submodel for each disjunctive situation of ψ, and under all of these submodels ϕ needs to be satisfied.

AC3R extends the motivation behind AC3, which is to "prune inessential elements" from the actual causes. The extension demands that if we have another actually occurring disjunct that would alone be sufficient to result in the effect, we need to add it to the disjunctive cause. Correspondingly, we obtain that a formula of the form $(\vec{X} = \vec{x}) \wedge (\vec{Y} = \vec{y})$ for $\vec{X} \cap \vec{Y} = \emptyset$ is more specific and less minimal than $\vec{X} = \vec{x}$, which is in turn more specific and less minimal than $(\vec{X} = \vec{x}) \vee (\vec{Y} = \vec{y})$. Assume this disjunction is an actual cause of some effect. Then the disjunction strips the "overspecific detail" which specific disjunct is causally efficacious (both are!) from the actual cause.[12]

We show now that in the contextualized causal model $\langle M, t = 1\rangle$ the disjunction $F = 1 \vee R = 1$ is an actual cause of $D = 1$ according to our refined definition. AC1R is satisfied, as $\langle M, t = 1\rangle \models (F = 1 \wedge R = 1) \wedge \phi$. AC2R is satisfied as well. To see this, let $\vec{Z} = \{F, R, D\}$, and thus $\vec{W} = \emptyset$. Clearly, $F, R \subseteq \vec{Z}$. But then (a) $\langle M, t = 1\rangle \models [F = 0 \wedge R = 0]D = 0$. Furthermore, (b) $\langle M, t = 1\rangle \models [F = 1 \vee R = 1]D = 1$, as we have seen in the previous section. Finally, AC3R is satisfied: no subsets of the disjuncts $F = 1$ and $R = 1$ satisfy AC1R and AC2R. Moreover, there exists no further disjunct satisfying AC1R and AC2R, because the effect $D = 1$ is identical to ϕ and $\vec{Z} \setminus \{F, R\} = \{D\}$.

According to our refined definition, $F = 1$ does not count any more as an actual cause of $D = 1$. (The same holds mutatis mutandis for $R = 1$.) The reason is AC3R: $F = 1$ is not minimal. Why? Because there is a disjunction $F = 1 \vee R = 1$ with $R \subseteq \vec{Z}$, $R \cap F = \emptyset$ and $(R = 1) \neq (D = 1)$ satisfying AC1R and AC2R. Hence, $F = 1$ is "inessential" for $D = 1$ in the sense that it is not required for $D = 1$ to obtain. The reason is that the actual event $R = 1$ alone would also be sufficient for $D = 1$ to obtain.

Finally, note that the refined definition of actual causation is a proper generalization of Halpern and Pearl's definition: the refined definition reduces to Halpern and Pearl's (2005) if no disjunctive

[12] If we have a simple causal chain from a cause C over intermediates $B_1, B_2, ...$ to an effect E and nothing else, then the refined definition says that the big disjunction $\bigvee B_i \vee C$ is a cause of E. Hence, we can use the refined definition to detect causal chains. Alternatively, we could amend the refined definition further to exclude the intermediate B_i's. This amendment would use the fact that the B_i's are themselves effects of C.

formulas are considered. In this case, we retain all the characteristics of the Halpern-Pearl definition, in particular with respect to the problematic examples in the literature. Of course, unlike the Halpern-Pearl definition, the refined definition counts the disjunction of overdetermining causes as an actual cause, but not the indvidual overdetermining causes. We postpone an analysis of the difference between cases of overdetermination and cases like Sartorio's Switch to another occasion.

6.4 Conclusion

We have generalized Halpern and Pearl's (2005) causal model semantics to allow disjunctive causes of the type found in Sartorio's (2006) Switch. These disjunctive causes have an actual part, that is both disjuncts actually occur, and a counterfactual part, that is each disjunct would be sufficient for the effect to occur. Based on the causal model semantics extended by disjunctive antecedents, we refined Halpern and Pearl's definition of actual causation. Halpern and Pearl's original definition qualifies Flipper's flipping the switch as an actual cause of the captivated person's death and does not allow for disjunctive causes. In contrast, our refined definition does not count the individual disjuncts as actual causes but makes Sartorio's disjunction "at least one of Flipper flips the switch and Reconnecter reconnects" an actual cause of the person's death. Our refined definition, therefore, implements the observation of Sartorio (2006, p. 530) that "there is no general motivation for believing that, when (if) a disjunctive fact is a cause, at least one of its disjuncts must also be a cause."

We showed that a refined Halpern-Pearl definition can capture disjunctive causes in Sartorio's sense. We make now a big jump from analyses of causation to the question of mental causation. More specifically, we will have a look at mental causation from the viewpoint of Woodward's (2005) interventionism blended with the assumptions and methods of cognitive neuroscience.

Chapter 7

Interventionist Mental Causation and the Methods of Cognitive Neuroscience

How can the mind interact with the body? This is a big puzzle when you believe that mental states are fundamentally different from physical bodies. Descartes (1641/1978), for example, believed that there are two substances, viz. *res cogitans* and *res extensa*.[1] While the latter are spatially extended things, incapable of feeling or thought, the former are unextended things that think and feel such as immaterial souls. Despite believing in a radical distinction between these substances, Descartes accepted the common belief that mind and body causally interact. However, if minds and bodies belong to so radically different substances, it is not easy to see how they could interact. On his view, minds are immaterial and have no spatial location, so how could they causally affect bodies? This question expresses the problem of mental causation for Descartes.

A dualism of two distinct substances is not the main source of our contemporary worries about mental causation. Most of us, and in particular scientific-minded people, reject the immaterial mind as *res cogitans*. Rather the mind somehow depends on the brain: without brain no mind.[2] Like Descartes, we still take mind-world interactions for granted in everyday experience and in scientific practice. The pain you feel when you twist your knee is taken to cause you to open the freezer in search of an ice pack. Mental images, so tell us psychologists, enable us to navigate our surroundings in an intelligent way. Economists explain fluctuations in financial markets by citing the beliefs of traders about the price of oil next month. In each of these examples, some mental properties ('pain', 'mental images', 'beliefs') appear to cause complex and coordinated bodily motions ('open the freezer', 'navigate our surroundings', 'consumer decisions').

The question is now: how does the mind depend on the brain? All physicalist accounts claim that the mind supervenes on a physical entity.[3] A particularly strong mind-body supervenience is proposed by Smart's (1959) mind-brain identity thesis: mental states are identical to brain states. This reductionist physicalism is a limiting case as there is no difference between the mental and the physical left, and so we may say that the identity theory posits only one kind of states. Like the monistic identity theory, any functionalist view is – at least – committed to a supervenience relation. Systems that have a similar physical make-up, typical functionalists believe, must be alike regarding their mental states. The reason to expect this be that identically constituted physical systems satisfy

[1] Descartes (1641/1978) stipulates another substance, namely God. To be precise, he believes that there are two substances apart from God. For the sake of presenting a clear-cut dualism, we omit God from Descartes's metaphysics.

[2] In many societies, the 'no-brain, no mind' slogan is taken quite literally: a person is declared dead if her brain has stopped working.

[3] Of course, there are many ways how to spell out a supervenience relation. See McLaughlin (1995).

the same causal roles in all physical and behavioral contexts. The pain when you twist your knee, for example, causes certain overt responses: you moan, you aim to repair the damage, you belief that you are in pain, you desire to alleviate the pain, and so on. To be in pain is functionally equivalent to a state that has certain causal roles: it is caused by tissue damage and it causes certain overt responses.

Functionalists hold against the identity thesis that, for instance, being in pain cannot be identified with a particular neurological state. The reason is that there be not one neurological state that corresponds to pain but many. Let alone other creatures such as octopodes to which we attribute pain despite a vastly different physical make-up. Still, functionalism allows that mental states are 'higher-level' states which occur in virtue of, and are optionally reducible to, 'lower-level' states, their realizers. In a nutshell, functionalists claim that mental states are mutliply realizable, and thus there cannot be type identity between mental states and physical states.

Fodor (1974) put forth a strong challenge for any reductive physicalism. He argued that multiple realizability plausibly rules out the reduction of higher-level states to lower-level states. Fodor's functionalism is an example of state (or property) dualism: mental states (or properties) are not reducible to physical states (or properties). Other examples of non-reductive physicalism are emergentism, that is the claim that there are physically irreducible emergent states, and Davidson's (1963) 'anomalous monism'. And non-reductive physicalism seems to be what common-sense demands. Although mental states are not reducible to, and are not identical with, physical states, the former still depend on the latter. Non-reductive physicalism allows for a certain autonomy of the mental. Your mental states such as desires can move you to do something. In contrast, if reductive physicalism got it right, desires for chocolate could not induce you to ransack your child's Halloween bag, but only neural activity could. This stands in open contradiction to the mainstream opinion in the philosophies of psychology and neurophilosophy: their explanations hinge on the possibility of mental causation. If your mind and its states, such as your beliefs and desires, were causally isolated from your bodily behavior, then what goes on in your mind could not explain what you do, only other physical states could (see, for example, Davidson (1963) and Mele (1992)).

Nowadays, the focus has shifted away from mental states to mental properties. The problem is then often put as follows: How could mental properties be causally relevant to bodily behavior? The question might seem to be a bit odd. After all, we just have observed that non-reductive physicalism seems to allow for a plausible answer. However, Kim (2005) put forth an argument that excludes the causal efficacy of mental properties if we assume non-reductive physicalism. As he puts it, non-reductive physicalists subscribe to three tenets:

(i) Mental properties supervene on physical properties.

(ii) Mental properties cannot be reduced to physical properties.

(iii) Mental properties are causally efficacious.

Independent of non-reductive physicalism, Papineau (2000) gives us good reasons to accept the following, already plausible metaphysical postulate:

(P1) Every physical effect has a sufficient physical cause.

Now, one way to spell out Kim's (2005) causal exclusion argument runs roughly as follows. Suppose a mental property M is causally efficacious by causing another mental property M^*.[4] By (i), there is a physical property P^* on which M^* supervenes. Given that M^* has a supervenience base P^*, M^* is

[4]To be precise, it should say the instantiation of the mental property M causes the instantiation of another mental property M^*. We leave the instantiation part implicit in the argument.

there independent of M, unless M caused P^*. A way out of this tension is to say that M caused M^* by causing its supervenience base P^*. By (i) again, M has a supervenience base P. Since M causes P^* and P is sufficient for M, Kim (2005, p. 41) thinks there "are strong reasons" that P causes P^*. By (P1), there is in any case some P that is sufficient for P^*. So we obtain that M causes P^* and P causes P^*. By (ii), M is not identical with P. Crucially, Kim assumes that P^* cannot be causally overdetermined by M and P. By (P1), P^* is rather caused by P than by M. Hence, M is excluded by the causal efficacy of P.

In Kim's exclusion argument the mental property M was chosen arbitrarily. Therefore, it applies to any such property. Notice that this generalization would lead to 'systematic' overdetermination when we allow that P^* can be causally overdetermined by M and P. Hence, the causal efficacy of M is inconsistent with the joint acceptance of (i)-(iii) and (P1).

In general, causal exclusion arguments assume non-reductive physicalism and conclude that mental properties supervening on physical properties cannot cause physical or other mental properties, at least in the presence of further premises. By excluding the causal efficacy of mental properties, these arguments can be read as support for epiphenomenalism, the view that the mental cannot affect the physical, or else as an attack on non-reductive physicalism.

Causal exclusion arguments suggest that giving up the Cartesian conception of minds as non-physical substances in favour of a more materialist ontology does not make the problem of mental causation go away. Rather the widely-shared physicalist commitment that the mental somehow depends on the physical but cannot be reduced to it, as expressed by (i) and (ii), seems to be one of the new sources of the mental causation problem. Property dualism seems to be not much better off than substance dualism. Or, as Kim (2005, p. 158) puts it:

> [t]he demands of causality do not tolerate duality of properties any more than duality of substances, and both Cartesian substance dualism and contemporary property dualism run aground on the rocks of mental causation.

Interestingly, Kim appeals here to the 'demands of causality'. The concept of causation used in the causal exclusion arguments is, however, typically left implicit. Nevertheless, as Hitchcock (2012) points out, the validity of these arguments may depend on the specific account of causation assumed in the background.[5]

The metaphysics of event-causation suggests that events occur or not. In modern science, however, the relata of causation seem to be random variables rather than binary variables. This difference is grounded in a different interest of scientists: while philosophers are primarily interested in analysing causation as a (metaphysical) concept, scientists want a concept of causation that fits their experimental practice. Normally, the variables an empirical scientist measures are not binary but many-valued, and often enough these variables take probabilistic values. In an experiment, scientists typically manipulate the values of some random variables and measure how some dependent variables are affected. Cognitive neuroscientists are, among other things, interested in manipulating mental variables and measuring the ensuing effect on neural activity. Perhaps, the problem of mental causation is often enough treated as a problem of applied metaphysics.

[5]Recently, Gebharter (2017) investigated the validity of causal exclusion arguments using the concept of causation as provided by the theory of causal Bayes nets. Interestingly, he argues that supervenience relations within this theory would behave just like causal relations. As regards 'higher-level' special sciences, this seems somewhat mistaken. Psychology and cognitive neuroscience, for example, assume the 'formal' characteristic of causation that a cause precedes its effect in time. A mental property and its supervenience base, in contrast, seem to be instantiated simultaneously. If not for the metaphysics of causation, this difference between the two kinds of relations seems at least highly relevant for the scientific practice.

In this chapter, we will have a look at the actual scientific practice rather than the metaphysics of mental causation. Our aim is to determine what exactly is required for acceptable causation as regards the mind-brain interaction. More specifically, we aim to make the characteristics of causation, as assumed in cognitive neuroscience, explicit. Subsequently, we impose these 'demands of causality' upon the account of interventionism put forth by Woodward (2005) and Woodward (2015). Within the resulting framework, we investigate to what extent we are justified to derive causal relations between mental properties and properties of the brain, depending on which scientific methods are used in the neuroscientific studies.

7.1 Introduction

Should cognitive neuroscientists interpret the results of their studies as establishing genuine causal relations? Well, the answer depends (at least) on three factors.

1. What does 'causal relation' mean?
2. Which methods are employed in the studies?
3. How is the relation between the mental and the brain?

Woodward's (2005) interventionist notion of causation captures, or so we argue, the notion of causation employed in cognitive neuroscience quite well. Furthermore, we coarsely distinguish between two methods: one manipulates a mental property and measures a property of the brain, the other proceeds vice versa.

The search for the neural correlates of mental functions is premised upon the idea that there is some dependency between mental properties and properties of the brain.[6] For reasons of cautiousness, we assume a minimal dependency relation in this chapter, which we call minimal supervenience.[7] We represent properties by variables, as in the following definition.

Definition 37. Minimal Supervenience
A variable M minimally supervenes on variable P (for $M \neq P$) iff

(i) M and P occur synchronically, and

(ii) any change of the value of M necessarily changes the value of P.

Minimal supervenience is defined by the conditions that are common to most supervenience relations. Hence, it expresses necessary conditions for supervenience, which are – depending on the philosophical view – not necessarily sufficient. Nevertheless we use 'minimally supervenes' and 'supervenes' interchangeably, unless noted otherwise.

Finally, we assume that brain properties are causes of other brain properties. We obtain the picture of Figure 7.1, which resembles Kim's (2005) canonical diagram. Given the depicted assumptions, we argue that cognitive neuroscientists should interpret the findings of their studies as establishing genuine causal relations.

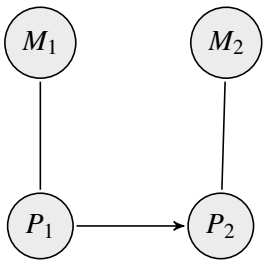

Figure 7.1: A causal network extended by minimal supervenience relations. The undirected edges stand for these supervenience relations. Note that there is a causal relation between brain property P_1 and brain property P_2.

[6] Cf. Squire et al. (2008, Ch. 53).

[7] Of course, a complete picture would require that we investigate what happens to the results obtained here if different relations between the mental and the brain are assumed, such as identity, functional reduction, causal relations, etc. We need to leave this more comprehensive investigation for future research.

In Section 7.2, we introduce the notion of causation assumed in cognitive neuroscience and the two methods we distinguish. In Section 7.3, Woodward's (2005) interventionist account of causation for models of exclusively causal relations is presented. In Section 7.4, we review Baumgartner's (2009) causal exclusion argument for the first method and Woodward's (2015) reply. In Section 7.5, we find that Baumgartner's argument does not apply with respect to the second method on Woodward's original account. We discuss the results in Section 7.6.

7.2 Causation and Methods in Cognitive Neuroscience

The research in cognitive neuroscience assumes a certain notion of causation. This notion is characterised by a temporal order: a cause precedes its effect in time. By Definition 37, condition (i), if a variable M supervenes on P, then neither M is a cause of P, nor P is a cause of M. For neither temporally precedes the other. Furthermore, causal relations are assumed to generate statistical dependencies. According to this theoretical assumption, if a cognitive neuroscientist manipulates the cause of an effect, she enforces a statistical dependency between them. Textbooks on experimental research widely agree that two conditions must be met to show a causal relation between the variable X and the variable Y:

(c_i) A systematic variation of X's value changes Y's value, and simultaneously

(c_{ii}) all other variables are controlled for to exclude other possible causes of Y.[8]

On the face of it, the notion of causation assumed in cognitive neuroscience, including the corresponding experimental practice, seems to be well-captured by Woodward's (2005) interventionist account of causation. We will have a closer look at this claim in the next section.

Many of the currently conducted studies in cognitive neuroscience adhere to one of two methods.

Method I: Manipulate a mental property of a participant and measure, subsequently, one of her brain properties.

Method II: Manipulate a brain property of a participant and inquire, subsequently, one of her mental properties.

In studies employing Method I, the mental property of a participant is manipulated by presenting different task conditions. Afterwards, the changes in the participant's neural activity are measured. Examples are studies in which changes in neural activity are measured by means of Functional Magnetic Resonance Imaging (fMRI) or Electro-Encephalogram (EEG). Method I studies are often interpreted to merely show a correlation rather than a genuine causal relation.

In studies employing Method II, a brain property of a participant is manipulated by brain stimulation techniques. Thereafter, the changes in the participant's mental property are inquired (or 'measured'). Examples are studies in which the electrical activity of the brain is directly manipulated by brain stimulation techniques such as Deep Brain Stimulation (DBS) and Transcranial Magnetic Stimulation (TMS). DBS has been used, e.g., to treat a variety of intractable pain syndromes, including neuropathic pain, phantom-limb pain, failed low back pain, and cluster-headache pain.[9] We take the pain relief to be a mental variable, which we can simply 'measure' by a participant's report. TMS

[8] See, for example, Carter and Shieh (2015); Gravetter and Forzano (2011); Squire et al. (2008); Windhorst and Johansson (1999)

[9] See Perlmutter and Mink (2006).

is a noninvasive procedure that uses magnetic fields to stimulate nerve cells in the brain. Like DBS, TMS can be used to improve symptoms of depression.[10] We take the mood reported by a participant to be an inquirable mental variable. In contrast to Method I studies, Method II studies are normally interpreted as establishing causal relations.[11]

Whether we are allowed to derive causal relations from studies in cognitive neuroscience depends on the question what causal relations are. We turn now to a recent answer due to Woodward (2005).

7.3 Woodward's Interventionist Account of Causation

Woodward (2005) provides the following definitions of direct and contributing cause, which we slightly rephrased.

Definition 38. Direct Cause (cf. Woodward (2005))
The variable X is a direct cause of the variable Y relative to a variable set $\mathcal{V}$ iff there is a possible intervention on X that changes Y's value, presupposed the values of all other variables V_i in $\mathcal{V}$ are kept fixed.[12]

We observe that an intervention on X that changes the value of Y corresponds to (c_i), and the presupposition that the values of all other variables V_i in $\mathcal{V}$ are kept fixed formally implements (c_{ii}). Hence, Woodward's definition of direct cause formally captures the necessary conditions of cognitive neuroscience's notion of causation.

Definition 39. Contributing Cause (cf. Woodward (2005))
The variable X is a contributing cause of Y relative to a variable set $\mathcal{V}$ iff

(i) there is a directed path from X to Y such that each link in this path is a direct causal relation and

(ii) there is some intervention on X that changes Y, presupposed the values of all variables in $\mathcal{V}$ that are not on the path are kept fixed.

From now on, we say that X is a cause of Y iff X is a contributing cause. Note that direct causes are a limiting case of contributing causes for a path of length 1.

Both of the definitions cite the term 'intervention'. In the context of experimental design, an intervention is a manipulation that changes the value of the (independent) variable X. Woodward defines interventions by means of an intervention variable.

Definition 40. Intervention Variable (cf. Woodward (2005))
Let I, X, Y be variables and $\{I, X, Y\} \subseteq \mathcal{V}$ be a set of variables. I is an intervention variable for X with respect to Y relative to $\mathcal{V}$ iff

(1) Setting $I = i$ determines $X = x$.[13]

(2) I breaks X's causal dependences on all variables in $\mathcal{V} \setminus I$.

(3) Any directed path in $\mathcal{V}$ from I to Y includes X.

[10] See George et al. (2003).
[11] See Sack (2006).
[12] For simplicity, we omitted "or Y's probability distribution" after "Y's value". We continue to do so as no argument here hinges on it.
[13] Corresponding to the previous footnote, we omit "or $P(X = x)$, where P is a probability measure" here.

(4) I is statistically independent of any variable $V \in \mathcal{V} \setminus I$ that is

(a) a cause of Y, and

(b) on a directed path not including X.

An intervention is thus, according to Woodward's formalism, a setting of the intervention variable to a value $I = i$ that changes the value of the variable X. By the conditions (1) and (2), I is the only and a direct cause of X. Furthermore, Woodward demands that such an intervention must satisfy the conditions (1)–(4) to establish a causal relation. We visualize an intervention variable and its properties in Figure 7.2 and Figure 7.3, respectively.

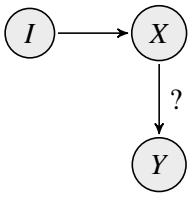

Figure 7.2: A directed acyclic graph $G = \langle \mathcal{V}, \mathcal{A} \rangle$ representing a causal network, whose sets of variables and arrows are $\mathcal{V} = \{I, X, Y\}$ and $\mathcal{A} = \{\langle I, X \rangle, \langle X, Y \rangle\}$, respectively. Each directed arrow represents a causal relation, or equivalently causal dependency. A directed arrow with question mark indicates which potential causal relation is investigated. Here the candidate under investigation is the relation between X and Y. By Definition 40, I is an intervention variable for X with respect to Y relative to $\mathcal{V}$.

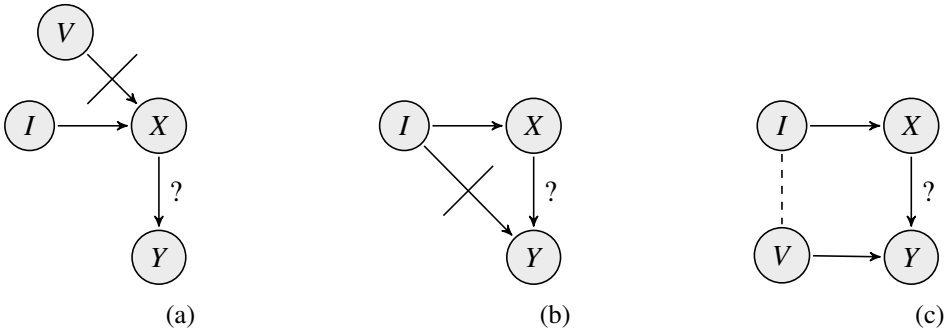

Figure 7.3: (a), (b), and (c) illustrate the conditions (2), (3), and (4) of Definition 40, respectively. A crossed arrow represents that a causal dependency cannot exist, if the respective conditions are satisfied. A dashed line stands for statistical independence. Note that no question mark arrows, crossed arrows and/or dashed lines are elements of basic causal networks. Those types of edge serve illustrative purposes only.

Definition 40 is meant to exclude potential confounding variables, which need to be controlled for in an experiment. This intention behind an intervention variable will be crucial to the debate on which variables ought to be controlled for in assessing causal relations, especially with respect to the causal efficacy of mental properties. If Definition 40 is modified with respect to the set of variables that should be kept fixed, we obtain a different (but still interventionist) notion of cause.

Woodward's framework is intended for models of exclusively causal relations, i. e. the variables in such a model may be causally related or correlated, but they do not stand in a supervenience relation. In such a model, each represented variable is independently fixable, i. e. it is possible to set the value of each variable independently of the other represented variables. In a model including a supervenience relation, independent fixability is violated. For an intervention on the supervening variable necessarily changes its subvenient variable. This prompts the question how to read the Definitions 38, 39, and 40 for models including causal and supervenience relations.

7.4 Method I, Baumgartner's Causal Exclusion Argument and Woodward's Reply

We model now Method I under Baumgartner's (2009) reading of Woodward's (2005) interventionist account of causation extended by minimal supervenience relations and show that Method I falls prey to Baumgartner's (2009) Causal Exclusion Argument.

We extend Woodward's interventionist account by including supervenience relations, as defined by Definition 37. The result are models containing two types of relations: supervenience relations between properties of an event (e.g. between M_1 and P_1) and causal relations between events (e.g. the event M_1/P_1 is a cause of M_2/P_2.) For such models, Baumgartner does not modify the Definitions 38, 39, and 40.

Method I studies examine the effect on brain state P_2 after manipulating the mental state M_1 (see Figure 7.4). Typically, a cognitive neuroscientist, let's call her Pat, manipulates the mental state M_1 of the participant in an experiment: the subject faces different task or stimulus conditions which are supposed to induce changes in M_1. Then, Pat investigates the effect of her manipulations by measuring the participant's brain state P_2 in all conditions. Using controls like randomization of subjects, our ideal neuroscientist Pat ensures that the conditions only differ on the manipulated mental variable. If Pat finds a significant difference in brain activity between the conditions, she derives that M_1 has had an effect on P_2.

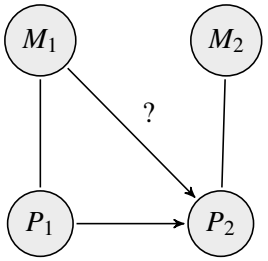

Figure 7.4: The relation investigated by studies of cognitive neuroscience using Method I.

However, Method I studies are normally interpreted to show a correlation only. Are we justified in deriving a causal relation from Method I studies within the framework of Woodward's extended interventionism? Baumgartner (2009) says no. More specifically, Baumgartner (2009) provides a causal exclusion argument that says: we are not allowed to derive a causal relation between M_1 and P_2, if we assume Woodward's interventionism extended by supervenience relations.

Baumgartner's (2009) causal exclusion argument runs as follows. Woodward's interventionism entails the implication if M_1 is a cause of P_2 relative to $\mathcal{V} = \{M_1, M_2, P_1, P_2\}$, then there is an

intervention variable I possible that changes the value of M_1 and is statistically independent of any variable $V \in \mathcal{V} \setminus I$ that is (a) a cause of P_2 and (b) on a directed path not including M_1 (compare p. 170). But such an I is impossible. Since M_1 minimally supervenes on P_1, any I that changes the value of M_1 also changes the value of P_1. Hence, I is statistically dependent on P_1. Moreover, P_1 is (a) a direct cause of P_2 by assumption and (b) on a directed path not including M_1. If I is an intervention variable, then – by condition (4) of Definition 40 – I is required to be statistically independent of P_1. Therefore, I is no intervention variable. By *modus tollens* on Baumgartner's implication, M_1 is not a cause of P_2.

In natural language, if Pat intervenes on a participant's mental state, she simultaneously intervenes on the physical state due to the supervenience relation between M_1 and P_1. Because of this non-causal relation, Pat cannot control the effect of P_1 on P_2 when manipulating M_1. By condition (i) of Definition 39, M_1 is also no contributing cause of P_2, since the link between M_1 and P_1 is no causal relation. Poor Pat, P_1 is a confounding variable for which she cannot control. Baumgartner's reading of Woodward's extended interventionism recommends that we should not derive a causal relation between M_1 and P_2 – quite in agreement with the current interpretation of Method I studies.

In reply to Baumgartner's argument, Woodward (2015) provides an interventionist framework for causal relations in the presence of supervenience relations. He emphasizes that Woodward's (2005) interventionist account of causation, as given by Definitions 38, 39, and 40, is intended for models of exclusively causal relations. If we want to include non-causal relations, such as the minimal supervenience of Definition 37, we need to modify the interventionist account. In light of Baumgartner's argument, Woodward's (2015) modification idea is to relax the requirement, or presupposition, to keep the variables fixed. In models including non-causal relations, only 'appropriate' variables need to be fixed for establishing a causal relation. When assessing a causal relation between a supervening variable and another variable, for example, it be not appropriate to keep fixed its supervenience base(s). In the modified framework, he claims, supervening properties can be causally efficacious.

To be more precise, Woodward (2015) modifies the notion of an intervention variable. Such an intervention variable I^* on X with respect to Y (relative to $\mathcal{V}$) (a) respects the supervenience relation between X and its supervenience base *SB(X)*, if there is one, and (b) requires that the conditions (1)-(4) of Definition 40 are restricted to causally related or correlated variables, but do not apply to supervenient and subvening variables. By (a), an I^*-intervention on X is thus at the same time an I^*-intervention on *SB(X)*, if such a supervenience base exists.

Consider Figure 7.4 again. Applying the modified Definition 38, we obtain that M_1 is a direct cause of P_2 relative to $\mathcal{V} = \{M_1, M_2, P_1, P_2\}$ iff there is a possible I^*-intervention on M_1 that changes the value of P_2, presupposed the values of all other appropriate variables in $\mathcal{V}$ are kept fixed. Since P_1 is the supervenience base of M_1, P_1 is not one of the appropriate variables to keep fixed. Hence, M_1 can be a cause of P_2 according to Woodward's modified account.

As we have seen, Baumgartner's (2009) causal exclusion argument entails that his reading of Woodward's (2005) original interventionism does not recommend to interpret the results of Method I studies as establishing causal relations. In contrast, Woodward's (2015) modified interventionism allows us to derive causal relations between mental states and brain states from Method I studies – somewhat at odds with the current interpretation in cognitive neuroscience.

7.5 Method II and Causal Exclusion

We revisit now Baumgartner's (2009) Causal Exclusion Argument for his and Woodward's (2015) reading of interventionism (both extended by minimal supervenience relations) with respect to Method

II studies. Those studies examine the effect on mental state M_2 after manipulating the brain state P_1 (see Figure 7.5). Our ideal cognitive neuroscientist Pat manipulates the physical state P_1 of the participant's brain by directly intervening upon the electrical activity in the brain using brain stimulation techniques. Then, Pat investigates the effect of her manipulations by inquiring the participant's mental state M_2. Using controls like sham stimulation, she ensures that the experimental conditions – often stimulation vs. no stimulation – only differ with respect to P_1.[14] If Pat finds a significant difference in the participant's mental state, she derives that P_1 has had an effect on M_2.

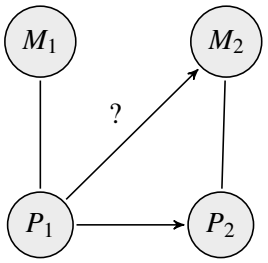

Figure 7.5: The relation investigated by studies of cognitive neuroscience using Method II.

Method II studies are often interpreted to show a proper causal relation.[15] Are we justified in deriving a causal relation from those studies according to Woodward's original account? As it turns out, yes.

We obtain an implication similar to Baumgartner's (2009): if P_1 is a direct cause of M_2 relative to $\mathcal{V} = \{M_1, M_2, P_1, P_2\}$, then there is a variable I possible that changes the value of P_1 and is statistically independent of any variable $V \in \mathcal{V}$ that is (i) a cause of M_2 and (ii) on a directed path not including P_1. This time such an I is possible. The reason is that there is no variable apart from P_1 that can be a cause of M_2 and is on a directed path not including P_1. Since there is a supervenience relation between M_2 and P_2, the only other candidate cause for M_2 is M_1. But Baumgartner's reading denies that M_1 can be a cause of M_2. For an intervention on M_1 does not allow to keep fixed its supervenience base P_1, as required by Baumgartner. Pictorially speaking, P_1 is on any directed path to M_2, if there is one.[16] Hence, a causal exclusion argument in Baumgartner's style does not apply with respect to Method II such that his reading of Woodward (2005) allows us to derive causal relations from studies using brain stimulation.[17]

Interestingly, Woodward's (2015) reading faces a putative problem with respect to Method II studies. The potential problem is that P_1 gets a competitor candidate cause of M_2, viz. M_1. Recall that Woodward's modification implies that supervenience bases need not be kept fixed. According to the modified account, we obtain the implication: if M_1 is a cause of M_2 relative to $\mathcal{V} = \{M_1, M_2, P_1, P_2\}$, then there is a variable I^* possible that changes the value of M_1 and is statistically independent of any appropriate variable $V \in \mathcal{V}$ that is (i) a cause of M_2 and (ii) on a directed path not including M_1. Since the supervenience bases of M_1 and M_2, i. e. P_1 and P_2 respectively, do not belong to the appropriate

[14] Sham stimulation is a generic term to indicate an inactive form of stimulation (e.g., a very brief or weak one) that is used in research to control for the placebo effect. The subject believes he/she is being stimulated normally, but there should not be any real effects.

[15] See, for example, Martin and Gotts (2005) and Romei et al. (2012).

[16] Notice that the possible intervention on P_1 does not necessarily change M_1 due to the supervenience relation, but this possible intervention may change P_2 by the assumed causal relation. However, P_2 is not a cause of M_2 due to their supervenience relation, and thus P_1 is a direct cause of M_2, given a change of M_2's value has been detected.

[17] This argument has been put forward by Dijkstra and de Bruin (2016) in a slightly different context.

variables, there is such an I^* possible that changes the value of M_1 and, perhaps, 'miraculously' results in a change of M_2. Hence, M_1 is an alternative candidate cause of M_2. The putative problem on the modified account is thus that we cannot know from a Method II study whether P_1 or M_1 is *the* cause of M_2.[18] We briefly discuss this issue in the next section.

7.6 Discussion

Consider again Figure 7.4. Woodward's (2015) account draws an arrow from M_1 to P_2 in addition to the assumed arrow from P_1 to P_2, if the value of P_2 changes when intervening on M_1. However, the arrow from M_1 to P_2 does not *mean* that M_1 has an effect on P_2 over and above P_1's effect on P_2. Similarly, if interventions on M_1 change M_2's value, then Woodward (2015) draws an arrow from M_1 to M_2 but also from M_1 to P_2, as P_2's value will change under an intervention on M_1 that changes M_2.

It seems that Baumgartner's (2009) Causal Exclusion Argument rests on the intuition that M_1 can only be causally efficacious, if it has a causal influence on M_2 or P_2 which is separate from P_1's causal influence on those variables. Due to this implicit premise, or so it seems, Baumgartner thinks P_1 should be kept fixed in order to test the 'independent' causal influence of M_1. Woodward (2015) simply does not share this intuition and Baumgartner does not provide any argument why we should adopt it.

The question which variables should be controlled for when assessing the causal efficacy of the mental is at the heart of Baumgartner's (2009) Causal Exclusion Argument. Should a cognitive neuroscientist adopt Baumgartner's intuition and control for P_1 when intervening on M_1? No, we shouldn't ask scientists for the impossible. Instead, we should encourage neuroscientists to shed light on the relation between the mental and the physical. The more they know the specific relation, the clearer they can see that an intervention on M_1, for instance, is 'automatically' an intervention on its supervenience base P_1. In this sense, an intervention on M_i should respect the supervenience relation between M_i and $SB(M_i) = P_i$, i. e. setting $M_i = m_i$ specifies $P_i = p_i$ that is consistent with $M_i = m_i$ according to the supervenience relation.

To figure out the details which mental properties supervene on which physical properties is of utmost importance, as the following consideration suggests. In Figure 7.1, M_1 supervenes on P_1. Hence, it is possible to change the value of P_1 without changing the value of M_1 (but not vice versa). Thus we may empirically figure out the values of P_1 under which M_1 remains invariant in order to test whether M_1 is a cause of M_2 in a sense, in which P_1 is not. If we intervene on P_1 such that M_1 keeps its value and M_2 as well, but intervening on M_1 changes M_2's value, then M_1 is a cause in a sense in which P_1 is not.[19]

Consider again Figure 7.5. Assume it is empirically found that P_1 is a cause of M_2 and further empirical investigations establish that M_1 is a cause of M_2. This is a putative problem for Woodward's (2015) account. The cause of M_2 is underdetermined: M_1 is a cause of M_2 and P_1 is a cause of M_2. However, this is only a putative problem. According to Definition 39, both M_1 and P_1 may qualify as *contributing causes*. Furthermore, Woodward (2005, 2015) explicitly states that arrows do not necessarily correspond to (physical) causal mechanisms which can be distinguished. Hence, the underdetermination problem is no problem for Woodward's formalism.

[18] Notice the possibility, however, that an intervention on M_1 changes the value of P_2 without changing the value of M_2. In such a case M_1 disqualifies as a cause of M_2, while M_1 is still a cause of P_2.

[19] Note that the possibility to experimentally distinguish the causal influence of P_1 and M_1 depends on the possibility to change the supervenience base P_1 without changing the supervenient variable M_1. These considerations become especially pertinent when relations of multiple realizability are considered. For an attempt in this direction, see List and Menzies (2009).

Apart from the misunderstanding of Woodward's formalism, the underdetermination raises the question whether there are reasons to distinguish between the supervenient variable M_i and its supervenience base P_i with respect to causal efficacy. After all, M_i/P_i is *one* event. Perhaps, causal relations hold only between events. Then the different 'properties' of the same event have the same causal effects. Importantly, this consideration illustrates that we should have the reasons in mind why we distinguish between mental and physical properties in cases of mental causation. Moreover, the consideration once more points to the importance to positively characterise the relation between the mental and the brain.

7.7 Conclusion

We imposed properties of causation, as assumed in cognitive neuroscience, upon Woodward's account of interventionism, and extended it by minimal supervenience relations. Within the resulting framework, we investigated to what extent we are justified to derive causal relations between mental properties and properties of the brain from methods used in studies of cognitive neuroscience.

The results of Baumgartner's (2009) reading of Woodward's (2005) interventionist framework extended by a minimal supervenience relation fits the current interpretations of the results in cognitive neuroscience. Correspondingly, studies that manipulate mental properties and measure properties of the brain do not allow us to derive causal relations. In contrast, Baumgartner's (2009) reading justifies to derive causal relations from studies which manipulate brain properties and inquire mental properties. These results agree with the current interpretations in cognitive neuroscience. If Baumgartner's (2009) reading is correct, this provides support for the interpretations. If the interpretations of cognitive neuroscientists are justified, this provides support for Baumgartner's (2009) reading.

Woodward (2015) modifies his original framework to include non-causal supervenience relations. On this account, we are allowed to derive genuine causal relations from both methodological kinds of study. We have seen that the idea behind Baumgartner's (2009) causal exclusion argument is if the property M_1 is a cause of P_2 (or M_2), then M_1 must have a causal impact on P_2 (or M_2) while keeping fixed P_1's causal impact on these variables. However, as Woodward (2015) notes, it seems inappropriate to control for the supervenience base $SB(M_1) = P_1$ when assessing the causal efficacy of M_1. We side with Woodward, as it doesn't seem to be a necessary condition for a cause that we can intervene upon it while keeping fixed its supervenience base. The reason is simply that this is impossible. If it were a necessary condition for causation, then we would never have mental causation, which is absurd.

So far we see no reasons why we should not adopt Woodward's (2015) extended interventionism, enriched by properties of causation as assumed in cognitive neuroscience, as the theoretical framework for determining whether the corresponding studies establish genuine causal relations. On the one hand, Baumgartner (2009) does not provide any positive argument against Woodward's (2015) account. On the other hand, our resulting framework is formally precise and provides unique answers, is easily applicable and fits nicely to the experimental practice of cognitive neuroscientists. Therefore, we recommend – against the current interpretations – that cognitive neuroscientists should dare to interpret both methodological kinds of studies as establishing genuine causal relations.

The results and our recommendation depend on our assumption of a minimal supervenience relation between mental properties and brain properties. While most cognitive neuroscientists agree that there is at least such a minimal dependence relation between mental processes and brain processes, there could also be a stronger relation such as multiple realizability, functional reduction, or even type identity. Of course, minimal supervenience provides no satisfaction in contrast to a more detailed ac-

count of the relation between the mental and the brain. The investigation of how the obtained results change when we spell out the mind-brain relation is left for future work.

Chapter 8

Conclusion

We have put forth a method of learning conditional information, a strengthened Ramsey Test semantics, an analysis of an (almost) asymmetric natural language usage of 'because', and two analyses of causation. We review the main results in turn, and point to just a few of the many limitations of our achievements.

In Chapter 2, we have considered the question how a rational agent learns "If A, then C?" We have proposed that she images on the minimally informative meaning of Stalnaker's (1968) conditional $A > C$ (given a default assumption). Unlike extant methods of learning conditional information, such as the ones found in Douven (2015) and Hartmann and Rad (2017), our method generates the correct predictions for Douven's (2012) benchmark examples. We have generalised Lewis's (1976) imaging to Jeffrey imaging. The generalisation makes our method applicable to the learning of uncertain conditional information, which we have illustrated with Van Fraassen's (1981) Judy Benjamin Problem.

We adapted our method of learning conditional information to a method of learning causal information. Causal information has been modelled as causal dependence in Lewis's (1973c) sense. We have simply replaced Lewis's two counterfactuals in his definition of causal dependence by the two corresponding Stalnaker conditionals. (We have also replaced the default assumption by a causal difference assumption.) Within the method of learning (uncertain) causal information, we have implemented Douven's (2012) idea: upon learning a conditional, the probability of the antecedent is determined by how well the antecedent explains the consequent. The two methods form a unified framework for the learning of certain and uncertain factual, conditional, and causal information. Furthermore, we have sketched an amendment to Stalnaker's semantics which gives us reasons to conjecture that our framework might also be applicable for the learning of subjunctive conditional information. We leave it to future research whether or not we can uphold our conjecture.

We have not yet provided an in-depth comparison of our framework to the account of Hartmann and Rad (2017). They make external information available to the epistemic agent, viz. information about independence relations as suggested by a Bayesian network. Thereby, they confer conditionals a causal reading. (The same applies to Douven's (2015) account, if explanatory considerations are closely tied to causal relations.) However, as Edgington (2008, p. 18) puts it:

> No conditional that does not explicitly use causal language like 'produce' or 'make', 'result' or 'outcome', forces a causal reading, though of course it is very often rightly presumed to be asserted on causal grounds. 'If A happens, B will happen, but A won't cause B to happen' is never contradictory.

In light of Edgington's view, we can turn the table on Hartmann and Rad (and presumably Douven) by

asking: how do their account(s) deal with conditionals that express merely conditional information?

In Chapter 3, we have strengthened Ramsey's (1929/1990) test by a suspension of judgment:

> First, *suspend judgment on the antecedent A and the consequent C*. Second, add *A* hypothetically to your stock of beliefs. Finally, test whether you can infer *C*.

Based on this evaluation recipe for conditionals, we have worked out a strengthened Ramsey Test semantics. Extant Ramsey Test semantics, like Stalnaker's (1968), Lewis's (1973a), and Gärdenfors's (1988), have not sufficiently considered that changes of belief and the plausible acceptance of conditionals come apart. A mere revision of ones belief state by the antecedent is not sufficient to establish a relation of relevance between the antecedent and the consequent. Such a relevance relation, however, seems to be called for when plausibly accepting a conditional.

Unlike Rott's (1986) Strong Ramsey Test, our strengthening requires a relation of relevance between antecedent and consequent that is *asymmetric* for a wide variety of cases, at least if we model an agent's epistemic state by a prioritised belief base. Relative to an agent's epistemic state after the suspension of judgment, the antecedent A is necessary and sufficient to infer the consequent C, whereas the consequent C is typically not sufficient to infer A. Only if both implications $A \to C$ and $C \to A$ are non-trivially believed after suspending judgment, our strengthened Ramsey Test conditional $\gg$ is symmetric. This result is desirable because we do not want to preclude the possibility that, for instance, both conditionals "If $A \wedge B$, then A and B" and "If A and B, then $A \wedge B$" are acceptable.

We have used our strengthened Ramsey Test semantics to provide an analysis of one usage of 'because' in natural language. Inheriting the asymmetry expressed by the strengthened conditional $\gg$, the analysis has allowed us to validate in a context, where the sun is shining, that 'there is a shadow because there is a tower', while it invalidates the converse 'there is a tower because there is a shadow'. Hence, our analysis of 'because' can be used to capture explanatory directions. We have generalised the tower-shadow scenario in an attempt to provide a notion of explanation that is likewise relative to an epistemic agent. The rough idea has been that A inferentially explains C just in case the agent believes generalisations G and literals L such that C follows logically from A, G, L, but C does not follow logically from G, L. Our notion of inferential explanation thus looks a lot like a regularity analysis of causation, as outlined in the Introduction. We have shown that an agent a accepts 'C because of A' iff A inferentially explains C relative to a.

Our work centering around the strengthened Ramsey Test is limited in many ways. We have not achieved a full positive characterisation of the logical theory governing our new conditional connective $\gg$. In future, we hope that we can provide a sound and complete axiom system for our strengthened Ramsey Test semantics. Moreover, our strengthened Ramsey Test is so far limited to the domain of non-probabilistic applications. A probabilification of the strengthened Ramsey Test faces the problem that there are many ways how to suspend judgment on a given probability function. (For the expert reader, this is similar to the problem of inverse Bayesian conditionalization.) We would need some criterion to single out a unique resulting probability function after the contraction required by our suspension of judgment. Alternatively, we might go imprecise in the sense that we consider the set of all possible probability functions that result from the required contraction. Perhaps, a middle-ground in between these alternatives is the only reasonable option.[1]

In Chapter 4, we have proposed an analysis of actual causation in terms of our strengthened Ramsey Test conditional. The previous chapters suggest that our strengthened Ramsey Test expresses an (almost) asymmetric relation of bringing about or production. We have supplemented

[1] For preliminary work towards probabilistic belief contraction, see for example Gärdenfors (1988) and Ramachandran et al. (2012).

our Strenghthened Ramsey Test conditional $\gg$ by the notion of a forward-directed proof and imposed the Humean constraint that a cause temporally precedes its effect. Thereby, we have solved the problems of joint effects, overdetermination and conjunctive scenarios, as well as (early and late) preemption. We have achieved to solve the problems posed by switches and scenarios of double prevention by adding a weak condition of difference-making. The condition requires that the absence of an actual cause does not also bring about its effect, which seems to be a truism about actual causation. Finally, we have supplemented the analysis by a best system account of generalisations in order to tackle the problem of spurious causation. If this way to deal with spurious causes works, we have provided a reductive analysis of actual causation. This reductiveness is arguably a major advantage over contemporary 'definitions' of actual causation in the causal modelling literature.[2]

In Chapter 5, we have carried over our analysis of causation to the framework of causal models due to Halpern and Pearl (2005). We have extended their causal models semantics in order to be able to define our strengthened Ramsey Test for causal models. As a result, our analysis simplified significantly on pain of losing its reductiveness. Unlike the definitions of actual cause put forth by Halpern and Pearl (2005) and Halpern (2015), our analysis satisfactorily solves the problems of overdetermination and conjunctive scenarios at the same time. In particular, their analyses do not admit that the conjunction of lightning and a preceding drought is an actual cause of a forest fire. Moreover, the two definitions (of Halpern, and Halpern and Pearl) do not agree on which events are causes in scenarios of overdetermination. A minor limitation of our analysis within the framework of causal models is that we have not yet treated the problematic cases of double prevention, switches, omissions, and others. However, such a treatment is already underway.

In Chapter 6, we have generalised Halpern and Pearl's (2005) causal model semantics to allow disjunctive causes of the type found in Sartorio's (2006) Switch. These disjunctive causes have an actual part, that is both disjuncts actually occur, and a counterfactual part, that is each disjunct *would be sufficient* for the effect to occur. Based on the causal model semantics extended by disjunctive antecedents, we refined Halpern and Pearl's definition of actual causation. Unlike Halpern and Pearl's original definition, our refined version allows for disjunctive causes in the sense specified above.

In Chapter 7, we have tackled the question "Should cognitive neuroscientists interpret the results of their studies as establishing genuine causal relations?" We have done so by enriching Woodward's (2005) interventionist account of causal explanation with the assumptions about causation typically made in cognitive neuroscience. Within the resulting framework, we have investigated to what extent we are justified to derive causal relations between mental properties and properties of the brain when two different classes of methods are used in studies of cognitive neuroscience. Unlike Baumgartner's (2009) reading of Woodward's (2005) account, our result has been that cognitive neuroscientists should dare to interpret both methodological kinds of studies as establishing genuine causal relations. The result is predicated on the assumption of a minimal supervenience relation between mental and brain properties. How the obtained results change when we spell out the mind-brain relation differently is left for future work.

Let us end by reconsidering the conditional from the Introduction: "If Paris shoots his fatal arrow, Achilles dies young." Achilles just killed Hector, and, as the legend has it, Achilles is fated to die young if he kills Hector. Fate, at least in the ancient Greek sense, does not determine every action, incident, and occurrence, but it does determine the outcome of life. So, if Paris had not shot the fatal arrow, Achilles would have died young anyways – just in another way. Lewis's (1973c) analysis of

[2]More generally, our analysis is meant to be a competitor to contemporary accounts of causation, such as the ones proposed in Halpern and Pearl (2005), Halpern (2015), Hitchcock (2001, 2007), Hall (2007), but also in Ramachandran (1997), Paul (2000), Yablo (2002), and Schaffer (2005).

causation in terms of counterfactual dependence does not count Paris's shooting as an actual cause. The reason is that there is a myriad of possible ways to seal Achilles's fate. Hence, we just cannot figure out an event strictly in between Paris's shooting and Achilles's death, after which Achilles would not have died young, because there is none. Fate finds always a way to fulfil destiny. Hence, Achilles untimely decease is, so to speak, preemptively overdetermined by his fate, and by whatever actual circumstance kills him.

The example illustrates the source of the problem for counterfactual accounts of actual causation: counterfactual difference-makers depend on events that would be actual causes of an effect, but in fact, are not. Such would-be or back-up difference-makers come to the surface in cases of preemption. Even if Paris had not been the actual cause, fate would still have had many of these would-be difference-makers left up her sleeves. The problem of the counterfactual criterion for actual causation is that it takes the would-be difference-makers into account, and in doing so, allows them to mask the difference-making of the actual causes. In contrast, our analysis does not suffer from this drawback. Our analysis requires either to suspend judgment on would-be difference-makers, or else allows to retain an actual belief blocking the efficacy of would-be difference-makers. In brief, taking the presumed cause and putative effect 'out of the picture' leads to a less informative but actual situation, in which we can test whether the presumed cause is an actual cause of the putative effect. Suspend judgment on Paris shooting the arrow and Achilles dying young. As a consequence, you have no information left regarding whether or not Achilles is fated to die young and whether or not Achilles killed Hector. Assume Paris shoots the arrow fatal to Achilles. Infer that Achilles dies young. Hence, Paris shooting the fatal arrow is an actual cause of Achilles's early decease, independent of any would-be difference-makers. This is just one difference between counterfactual dependence and the 'agnostic dependence' expressed by our strengthened Ramsey Test. We hope it became clear that a detailed comparison between agnostic and counterfactual dependence deserves future consideration. This is but one of the stories that remain to be told.

Appendix A

A Possible Worlds Model of the Jeweller Example

Following the presentation in Douven and Romeijn (2011), we consider the Jeweller Example.

Example 12. The Jeweller Example (Douven and Romeijn (2011, p. 654))
A jeweller has been shot in his store and robbed of a golden watch. However, it is not clear at this point what the relation between these two events is; perhaps someone shot the jeweller and then someone else saw an opportunity to steal the watch. Kate thinks there is some chance that Henry is the robber (R). On the other hand, she strongly doubts that he is capable of shooting someone, and thus, that he is the shooter (S). Now the inspector, after hearing the testimonies of several witnesses, tells Kate:

If Henry robbed the jeweller, then he also shot him. (1)

As a result, Kate becomes more confident that Henry is not the robber, while her probability for Henry having shot the jeweller does not change.

We model Kate's belief state as the Stalnaker model $\mathcal{M}_{St} = \langle W, R, \leq, \leq' V\rangle$ depicted in Figure A.1. W contains four elements covering the possible events of $R, \neg R, S, \neg S$, where R stands for "Henry is the robber", and S for "Henry has shot the jeweller". The example suggests that $0 < P(R) < 1$ and $P(S) = \epsilon$ for a small ϵ, and thus $P(\neg S) = 1 - \epsilon$. The prescribed intuitions are that $P^*(R) < P(R)$ and $P^*(S) = P(S)$. We know about Kate's degrees of belief before receiving the conditional information that $0 < P(w_1) + P(w_2) < 1$ and $P(w_1) + P(w_3) = \epsilon$, as well as $P(w_2) + P(w_4) = 1 - \epsilon$. Note that Kate is 'almost sure' that $\neg S$, and thus we may treat $\neg S$ as 'almost factual' information.

Kate receives certain conditional information. She learns the minimally informative proposition $[R > S] = \{w_1, w_3, w_4\}$ such that $P(R > S) = P^R(S) = 1$. By the law of total probability, $P(R > \neg S) = P^R(\neg S) = 0$. Taking her uncertain but almost factual information into account, Kate learns in total the minimally informative proposition $[(R > S) \wedge \neg S]$, which is identical to $\{w_4\}$. By $P(R > S) = 1$, $P((R > S) \wedge \neg S) = P(\neg S) = 1 - \epsilon$. Note the tension expressed in $P((R > S) \wedge \neg S) = 1 - \epsilon$. It basically says that S is almost surely not the case *and*, under the supposition of R, we exclude the possibility of $\neg S$. Intuitively, the thought expressed by this statement should cast doubt as to whether R is the case.

By $\neg((R > S) \wedge \neg S) \equiv (R > \neg S) \vee S$, we also know that $P(R > \neg S) \vee S) = \epsilon$. Note that the proposition $[(R > S) \wedge \neg S] = \{w_4\}$ (interpreted as minimally informative) specifies a similarity order $\leq$ such that $w_{(R>S)\wedge\neg S} = w_4$ for all w. In contrast, the proposition $[(R > \neg S) \vee S]$ is minimally informative in a strong sense, since it does not exclude any world w. Hence, the 'maximally inclusive'

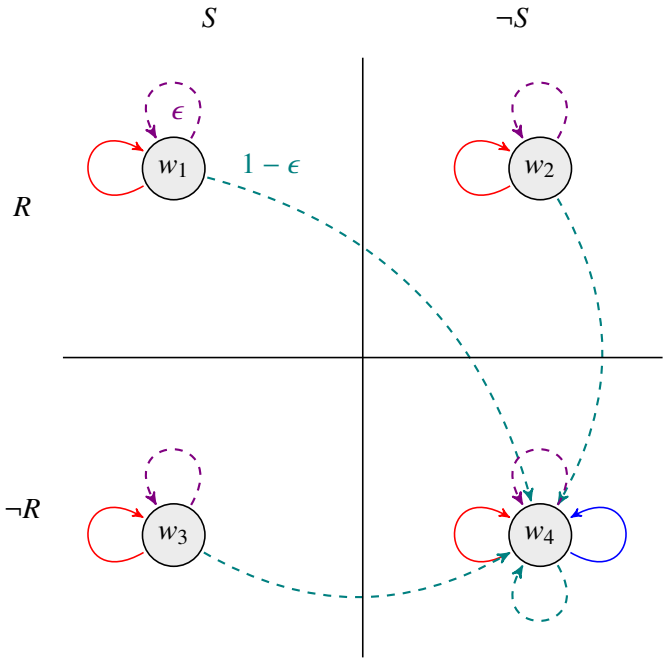

Figure A.1: A Stalnaker model for Kate's belief state in the Jeweller Example. The blue arrow indicates the unique $w_{((R>S)\wedge\neg S)}$-world under $\leq$. The red arrows indicate that each world is its most similar $\neg((R > S) \wedge \neg S)$-world under $\leq'$. The teal arrows represent the transfer of $(1-\epsilon)\cdot P(w)$, while the violet arrows represent the transfer of $\epsilon \cdot P(w)$.

proposition $[(R > \neg S) \vee S] = \{w_1, w_2, w_3, w_4\}$ specifies a similarity order $\leq' \neq \leq$ according to which $w_{(R>\neg S)\vee S} = w$ for each w.

We apply now Jeffrey imaging to the Jeweller Example, where $k = 1 - \epsilon$.

$$P^{(R>S)\wedge\neg S}_{1-\epsilon}(w') = P^*(w') = \sum_w (P(w) \cdot \left\{ \begin{array}{ll} 1-\epsilon & \text{if } w_{(R>S)\wedge\neg S} = w' \\ 0 & \text{otherwise} \end{array} \right\} + P(w) \cdot \left\{ \begin{array}{ll} \epsilon & \text{if } w_{(R>\neg S)\vee S} = w' \\ 0 & \text{otherwise} \end{array} \right\}) \tag{2}$$

We obtain the following probability distribution after learning:

$$\begin{array}{ll} P^*_{1-\epsilon}(w_1) = P^*_{1-\epsilon}(R \wedge S) = \epsilon \cdot P(w_1) & P^*_{1-\epsilon}(w_2) = P^*_{1-\epsilon}(R \wedge \neg S) = \epsilon \cdot P(w_2) \\ P^*_{1-\epsilon}(w_3) = P^*_{1-\epsilon}(\neg R \wedge S) = \epsilon \cdot P(w_3) & P^*_{1-\epsilon}(w_4) = P^*_{1-\epsilon}(\neg R \wedge \neg S) \\ & = (1-\epsilon) \cdot (P(w_1) + P(w_2) + P(w_3) + \\ & P(w_4)) + \epsilon \cdot P(w_4) \end{array} \tag{3}$$

The results almost comply with the prescribed intuitions. The intuition concerning the degree of belief in R is met: $P^*(R) < P(R)$, since $P^*_{1-\epsilon}(w_1) + P^*_{1-\epsilon}(w_2) < P(w_1) + P(w_2)$. The intuition concerning the degree of belief in S is 'almost' met: $P^*_{1-\epsilon}(S) \approx P(S)$, for $P(w_1) + P(w_3) = \epsilon$ and $P^*_{1-\epsilon}(w_1) + P^*_{1-\epsilon}(w_3) = P(w_1) \cdot \epsilon + P(w_3) \cdot \epsilon \approx \epsilon$. In words, the method gives us the result that Kate is now pretty sure that Henry is neither the shooter nor the robber.

Appendix B

Proofs of Chapter 3

Proposition 2. Let K be a non-absurd belief set and γ a non-tautology. Then (UPC) and $\alpha, \gamma \in K$ implies (Because$_R$).

Proof. The proof presents the simplification of (UPC) to (Because$_R$). Assume $K \neq K_\perp$ (K is not the absurd belief set) and γ is not a tautology. Further, suppose $\alpha, \gamma \in K$. Then, by Gärdenfors's contraction postulate (K^-4), $\gamma \notin K - \gamma$. Moreover, since $\gamma \in K$ and $K \neq K_\perp$, $\neg\gamma \notin K$. By (K^-3) this implies that $\neg\gamma \notin (K - \gamma)$. By Proposition 1, $\gamma \notin (K - \gamma)$ implies $\gamma \notin (K - \gamma) * \alpha$ or $\gamma \notin (K - \gamma) * \neg\alpha$, and $\neg\gamma \notin (K - \gamma)$ implies $\neg\gamma \notin (K - \gamma) * \alpha$ or $\neg\gamma \notin (K - \gamma) * \neg\alpha$. Hence, (i) $\gamma \notin (K - \gamma) * \alpha$ or $\gamma \notin (K - \gamma) * \neg\alpha$ and (ii) $\neg\gamma \notin (K - \gamma) * \alpha$ or $\neg\gamma \notin (K - \gamma) * \neg\alpha$. (i) implies that (iii), if $\gamma \in (K - \gamma) * \alpha$, then $\gamma \notin (K - \gamma) * \neg\alpha$. (ii) implies that (iv) if $\neg\gamma \in (K - \gamma) * \neg\alpha$, then $\neg\gamma \notin (K - \gamma) * \alpha$. From (iii), (iv), and (UPC), we can infer (Because$_R$). □

Proposition 3. Let α and γ be literals or conjunctions of literals. Further, $\alpha, \gamma \in K$. Suppose that $\alpha \to \gamma$ is a non-trivial implication in K and Assumption 1 holds for $\alpha \to \gamma$. Then, $\alpha \overset{a}{\Rightarrow} \gamma \in K$ and $\gamma \overset{a}{\Rightarrow} \alpha \in K$.

Proof. ($\alpha \overset{a}{\Rightarrow} \gamma \in K$) Suppose $\alpha \to \gamma$ is a non-trivial implication in K. Then, by Definition 13 and Assumption 1, $\alpha \to \gamma$ is a non-trivial implication in $K - \gamma$. By Definition 13, $(K - \gamma) - \neg\alpha \vdash \alpha \to \gamma$. We obtain, by the deduction theorem, $(K - \gamma) - \neg\alpha, \alpha \vdash \gamma$, which we can rewrite as $((K - \gamma) - \neg\alpha) + \alpha \vdash \gamma$. By the Levi identity, we obtain $\gamma \in (K - \gamma) * \alpha$. Using $\alpha, \gamma \in K$, we can infer therefrom that $\alpha \overset{a}{\Rightarrow} \gamma \in K$.

($\gamma \overset{a}{\Rightarrow} \alpha \in K$) Suppose $\alpha \to \gamma$ is a non-trivial implication in K. Then, by Assumption 1, $\alpha \to \gamma$ is a non-trivial implication in $K - \alpha$. By contraposition, $\neg\gamma \to \neg\alpha$ is a non-trivial implication in $K - \alpha$. By Definition 13, $\neg\gamma \to \neg\alpha$ is a non-trivial implication in $(K - \alpha) - \gamma$. Via the deduction theorem, we obtain $(K - \alpha) - \gamma, \neg\gamma \vdash \neg\alpha$. By the Levi identity, we obtain $\neg\alpha \in (K - \alpha) * \neg\gamma$. Using $\alpha, \gamma \in K$, we can infer therefrom that $\gamma \overset{a}{\Rightarrow} \alpha \in K$. □

Proposition 4. Assume a (Because$_R$) agent accepts all facts and the generalisation of the tower-shadow scenario, i. e. $t, s, sh, t \wedge s \to sh \in K$, where the order of epistemic entrenchment is $t, s, sh < t \wedge s \to sh$. Then, $t \overset{a}{\Rightarrow} sh \in K$ if $t \leq sh$ and $sh \overset{a}{\Rightarrow} t \in K$ if $sh \leq t$.

Proof. By (Because$_R$):

$$t \overset{a}{\Rightarrow} sh \in K \text{ iff } sh \in (K - sh) * t \text{ or } \neg sh \in (K - sh) * \neg t$$
$$\text{and } t, sh \in K.$$

$t, s, sh \in K$ holds by assumption. We show that (a) $t \overset{a}{\Rightarrow} sh \in K$ if $t \leq sh$. Let us assume $t \leq sh$. By (G-), (EE2), (EE1), this implies (i) $t \notin K - sh$. By the recovery postulate

$$\text{If } \alpha \in K, \text{ then } K \subseteq (K - \alpha) + \alpha \tag{K^-5}$$

$t \in (K-sh)*sh$. By the Levi identity, the deduction theorem, and (i), this implies that $sh \rightarrow t \in (K-sh)$. Hence, $\neg t \rightarrow \neg sh \in (K - sh)$. Using $t \notin K - sh$ and the Levi identity, we can infer therefrom that $\neg sh \in (K - sh) * \neg t$. This entails (a) in light of (Because$_R$).

The proof of (b) $sh \overset{a}{\Rightarrow} t$ if $sh \leq t$ is completely analogous to that of (a).

□

Proposition 5. Assume a (Because$_R$) agent accepts all the formulas in $K(H, <) = K(S)$ for $H = \{t, s, sh, t \wedge s \rightarrow sh\}$, where the order of epistemic priority is $t \sim s \sim sh < t \wedge s \rightarrow sh$. Then $t \overset{a}{\Rightarrow} sh \notin K_>(S)$ and $sh \overset{a}{\Rightarrow} t \in K_>(S)$.

Proof. $t \overset{a}{\Rightarrow} sh$ and $sh \overset{a}{\Rightarrow} t$ remain well-defined, when replacing K with $K(S)$:

$$t \overset{a}{\Rightarrow} sh \in K_>(S) \text{ iff } sh \in K((S - sh) * t) \text{ or } \neg sh \in K((S - sh) * \neg t) \text{ and } t, sh \in K(S)$$

$$sh \overset{a}{\Rightarrow} t \in K_>(S) \text{ iff } t \in K((S - t) * sh) \text{ or } \neg t \in K((S - t) * \neg sh) \text{ and } sh, t \in K(S).$$

$t \overset{a}{\Rightarrow} sh \notin K_>(S)$, where $S = (H, <)$: $t, sh \in K(S)$ is satisfied by assumption. By Definition 41, the remainder set $H \perp sh$ contains three sets, $H' = \{t \wedge s \rightarrow sh, s\}$, $H'' = \{t \wedge s \rightarrow sh, t\}$, and $H''' = \{s, t\}$. By Definition 12, $H' \leq H''$ and $H'' \leq H'$, but $H' \not\leq H'''$. Hence, by (Def σ), both H' and H'' are selected for the partial meet base contraction (PMBC) which yields $H - sh = \bigcap \sigma(H \perp sh) = \{t \wedge s \rightarrow sh\}$.

By (PMBR), $(H - sh) * t = \bigcap \sigma((H - sh) \perp \neg t) + t$. Since $\neg t \notin Cn(H - sh)$, by Definition 41, $H - sh$ is the unique member of $(H - sh) \perp \neg t$. By Definition 12, $(H - sh) \leq (H - sh)$ and by (Def σ), the partial meet base contraction (PMBC) yields $(H - sh) - \neg t = \bigcap \sigma((H - sh) \perp \neg t) = \{t \wedge s \rightarrow sh\}$. Notice that when $\neg t \notin H - sh$, then $(H - sh) - \neg t = H - sh$. By (H + α), $(H - sh) * t = \{t \wedge s \rightarrow sh\} \cup \{t\}$. Hence, $(S - sh) * t = ((H - sh) * t, <')$, where $<'$ is such that generalisations have strict priority over literals. Then $sh \notin K((S - sh) * t)$.

By (PMBR), $(H - sh) * \neg t = \bigcap \sigma((H - sh) \perp t) + \neg t$. By similar reasoning as above, $t \notin H - sh$ and thus $(H - sh) - t = H - sh$. By (H + α), $(H - sh) * \neg t = \{t \wedge s \rightarrow sh\} \cup \{\neg t\}$. Hence, $(S - sh) * \neg t = ((H - sh) * \neg t, <')$, where $<'$ is such that generalisations have strict priority over literals. Then $\neg sh \notin K((S - sh) * \neg t)$.

$sh \overset{a}{\Rightarrow} t \in K_>(S)$, where $S = (H, <)$: $t, sh \in K(S)$ is satisfied by assumption. By Definition 41, the remainder set $H \perp t$ contains only $H' = \{t \wedge s \rightarrow sh, s, sh\}$. By Definition 12, $H' \leq H'$ and by (Def σ), the partial meet base contraction (PMBC) yields $(H - t) = \bigcap \sigma(H \perp t) = \{t \wedge s \rightarrow sh, s, sh\}$. By similar reasoning, $(H - t) - sh = \{t \wedge s \rightarrow sh, s\}$. By (H + α), $((H - t) - sh) + \neg sh = \{t \wedge s \rightarrow sh, s, \neg sh\}$, which is by (PMBR) $(H - t) * \neg sh$. Hence, $(S - t) * \neg sh = ((H - t) * \neg sh, <')$, where $<'$ is such that generalisations have strict priority over literals. Then $\neg t \in K((S - t) * \neg sh)$.

□

Proposition 6. Assume a (Because$_P$) agent accepts all facts and the single, more entrenched generalisation of the tower-shadow scenario, i. e. $t, s, sh, t \wedge s \rightarrow sh \in K$, where the order of epistemic entrenchment is $t \sim s \sim sh < t \wedge s \rightarrow sh$. Then $t \overset{P}{\Rightarrow} sh \in K_>$ and $sh \overset{P}{\Rightarrow} t \in K_>$.

Proof. $t \overset{P}{\Rightarrow} sh \in K_>$: The agnostic belief set is $K' = K - (t \vee sh)$. By (G-), $t, s, sh \notin K'$, but by *recovery* (i) $(t \vee sh) \rightarrow sh \in K'$. By assumption, $\neg t \notin K$ and so by (K^-3), (ii) $\neg t \notin K'$. Using $t \vdash t \vee sh$ and the Levi identity, we can infer from (i) and (ii) that $sh \in K' * t$.

$sh \overset{P}{\Rightarrow} t \in K_>$: The agnostic belief set is, again, $K' = K - (t \vee sh)$ such that $t, s, sh \notin K'$, but by *recovery* $(t \vee sh) \rightarrow t \in K'$. Using $sh \vdash t \vee sh$, we infer that $t \in K' * sh$. □

Proposition 7. Assume a (Because$_P$) agent accepts all the formulas in $K(H, <) = K(S)$ for $H = \{t, s, sh, t \wedge s \rightarrow sh\}$, where the agent may assume whatever epistemic ordering $<$. Then $t \overset{P}{\Rightarrow} sh \in K_>(S)$, but $sh \overset{P}{\Rightarrow} t \notin K_>(S)$.

Proof. $t \overset{P}{\Rightarrow} sh \in K_>(S)$, where $S = (H, <)$; $<$ remains unspecified: $t, sh \in K(S)$ is satisfied by assumption. By Definition 41, the remainder set $H \perp (t \vee sh)$ contains only $H'' = \{t \wedge s \rightarrow sh, s\}$. Since $H'' \leq H''$ and by (Def σ), the partial meet base contraction (PMBC) yields the agnostic belief base (i) $H' = H - (t \vee sh) = \bigcap \sigma(H \perp (t \vee sh)) = \{t \wedge s \rightarrow sh, s\}$. By (PMBR), $H' * t = \bigcap \sigma(H' \perp \neg t) + t$. Using $\neg t \notin Cn(H')$, we obtain $H' * t = H' + t$. Together with (i), this implies that $H' * t = \{t \wedge s \rightarrow sh, s\} \cup \{t\}$. Hence, $sh \in K_>(S) - (t \vee sh) * t$ so that $t \gg sh \in K_>(S)$.

$sh \overset{P}{\Rightarrow} t \notin K_>(S)$, where $S = (H, <)$; $<$ remains unspecified: as just shown, (i) $H' = H - (sh \vee t) = \{t \wedge s \rightarrow sh, s\}$. By (PMBR), $H' * sh = \bigcap \sigma(H' \perp \neg sh) + sh$. Using $\neg sh \notin Cn(H')$, we obtain $H' * sh = H' + sh$. Together with (i), this implies that $H' * sh = \{t \wedge s \rightarrow sh, s\} \cup \{sh\}$. Hence, $t \notin K_>(S) - (sh \vee t) * sh$ so that $sh \gg t \notin K_>(S)$.

□

Proposition 8. Let α and γ be literals. Epistemic states are represented by prioritised belief bases with two levels: an upper level G of generalisations and a lower level L of literals, as explained in Section 3.2.6. A (Because$_{P'}$) agent accepts 'γ because of α' with respect to $(H, <)$ iff α inferentially explains γ – in the sense of Definition 16 – in the eyes of the agent accepting all members of H.

Proof. Suppose (i) γ because of α is verified by an epistemic state $(H, <)$ (in the sense of (Because$_{P'}$)). Let G be the set of generalisations of H, while L is the set of literals of H. Hence, (ii) there are $(H'', <'')$ and L^- such that $(H', <') = (H, <) - \bigvee L^-$ and $\alpha \gg \gamma \in K_>(H', <')$. Therefore, (iii) there is $(H'', <'') = (H', <') - \alpha \vee \gamma$ such that $H'' = G'' \cup L''$, $G'' = G \cap H''$, and $L'' = L \cap H''$. Hence, the pair (G'', L'') satisfies conditions (1), (2), and (4) for an agent who accepts all members of H. Moreover, (ii) and (iii) imply that $G'' \cup L'', \alpha \vdash \gamma$. Hence, Condition (5) is satisfied as well for (G'', L''). Finally, Condition (6) holds for (G'', L'') because of (ii) and (iii). (i) implies that Condition (3) of Definition 16 is satisfied for an agent who accepts all members of H. Hence, all conditions of this definition are satisfied for such an agent. Thus, α inferentially explains γ – in the sense of Definition 16 – in the eyes of an agent who accepts all members of H.

For the other direction, suppose (i) α inferentially explains γ in the eyes of an agent a, in the sense of Definition 16. Hence, there is a set G of generalisations and a set L of literals such that conditions (1) – (6) of Definition 16 are satisfied for a. (ii) $H := G \cup L \cup \{\alpha\}$. $<$ is such that generalisations are prioritised over literals. Obviously, (iii) $\alpha \in K(H)$. By Condition (5) of Definition 16, (iv) $\gamma \in K(H, <)$. We show that $\alpha \vee \gamma \notin Cn(G \cup L)$. Suppose, for contradiction, $\alpha \vee \gamma \in Cn(G \cup L)$. This implies that (v) $\neg\alpha \rightarrow \gamma \in Cn(G \cup L)$. By Condition (5) of Definition 16, we know that (vi) $\alpha \rightarrow \gamma \in Cn(G \cup L)$. Since $\alpha \vee \neg\alpha \in Cn(G \cup L)$, (v) and (vi) imply that $\gamma \in Cn(G \cup L)$. This contradicts Condition (6) of Definition 16. Hence, $\alpha \vee \gamma \notin Cn(G \cup L)$. Therefore, $\sigma((H, <) \perp \alpha \vee \gamma) := \{G \cup L\}$, where σ is defined by (Def σ) and Definition 12. Using Condition (5) of Definition 16, we can infer therefrom that $\alpha \gg \gamma \in K_>(H, <)$. Using (iii) and (iv), we can infer therefrom that 'γ because of α' is verified by the epistemic state $(H, <)$ (in the sense of (Because$_{P'}$)). □

Appendix C

Defining Belief Changes

In view of the Levi identity, we can focus on belief base contractions and set specific belief base revision schemes aside. A contraction by A can be defined using the notion of a *remainder set* $H\perp A$. Recall that the contraction of $K(S)$ by A yields a belief set $K(S')$ that does not contain A. The remainder set $H\perp A$, consequently, contains all maximal subsets of H that do not entail A. In formal terms:

Definition 41. $H\perp A$ (Hansson, 1999, p. 12)
Let H be a set of formulas and A a formula. $H' \in H\perp A$ iff
(C1) $H' \subseteq H$

(C2) $A \notin Cn(H')$

(C3) there is no H'' such that $H' \subset H'' \subseteq H$ and $A \notin Cn(H'')$.

Cn stands for the consequence operation of classical logic.

The next step is to define the contraction of the belief base H by A:

$$H \div A =_{df} \bigcap H\perp A. \qquad \text{(FMB Contraction)}$$

This way of defining a contraction is also referred to as *full meet base contraction*. $\bigcap H\perp A$ designates the intersection of the members of the remainder set $H\perp A$.

Full meet base contractions are motivated by two ideas. First, the *conservativity principle*: when forced to change our beliefs, we should retain as many as possible of the present beliefs (Harman, 1986, p. 46). This is the motivation for Condition (C3) in the above definition of $H\perp A$. Second, if we have no reason to prefer one way of retracting A over another, we retain only those beliefs that are preserved in all possible ways of retracting A. Since any member of $H\perp A$ represents a logically possible way of retracting A from H, we retain only those beliefs that are in the intersection of the members of $H\perp A$.

If combined with the idea of different levels of epistemic priority, full meet base contractions yield intuitive results. Such a combination is therefore assumed to define the suspension of judgement in our strengthened Ramsey Test.

To define full meet revisions for a prioritised belief base, we merely need to extend the definition of a remainder set $H\perp A$ to the prioritised case. Let

$$\mathbf{H} = \langle H_1, \ldots, H_n \rangle$$

be a prioritised belief base. That is, $H_1, \ldots, H_n$ are sets of formulas that represent explicit beliefs, and the indices represent an epistemic ranking of the beliefs. H_1 is the set of the most firmly established beliefs, the beliefs in H_2 have secondary priority, etc. We define the remainder set of such a prioritised belief base as follows:

Definition 42. $\mathbf{H}\perp A$
Let $\mathbf{H} = \langle H_1, \ldots, H_n\rangle$ be a prioritised belief base and A a formula. $\mathbf{H}' \in \mathbf{H}\perp A$ iff (i) $\mathbf{H}' = \langle H'_1, \ldots, H'_n\rangle$ and (ii) for all i $(1 \leq i \leq n)$, $H'_1 \cup \ldots \cup H'_i \in (H_1 \cup \ldots \cup H_i)\perp A$.

Any member $\mathbf{H}'$ of $\mathbf{H}\perp A$ is thus itself a prioritised belief base preserving the epistemic prioritisation of $\mathbf{H}$, i. e. any $\mathbf{H}'$ has the same structure of epistemic levels (1 to n) than $\mathbf{H}$. The union of the epistemic levels of such a base $\mathbf{H}'$ – the set $H'_1 \cup \ldots \cup H'_n$ – does not entail A by classical logic.[1]

The next step is to define the contraction of $\mathbf{H}$ by A in terms of the componentwise intersection of the members of the prioritised remainder set:

$$\mathbf{H} \div A =_{df} \bigcap \mathbf{H}\perp A \qquad \text{(Def } \div\text{)}$$

The componentwise intersection of $\mathbf{H}\perp A$ is defined in the obvious way: the set of the i-th level of $\bigcap \mathbf{H}\perp A$ $(1 \leq i \leq n)$ is given by the intersection of the i-th levels of the members of $\mathbf{H}\perp A$. Note that we use $\bigcap$ in a context-dependent manner; it means different things when applied to a set and to a sequence of sets.

It remains to define expansions for prioritised belief bases $\mathbf{H} = \langle H_1, \ldots, H_n\rangle$. Two types of definitions suggest themselves:

$$\mathbf{H} + A =_{df} \langle \{A\}, H_1, \ldots, H_n\rangle \qquad \text{(Def } +\text{)}$$

$$\mathbf{H} +_i A =_{df} \langle H_1, \ldots, H_i \cup \{A\}, \ldots, H_n\rangle \qquad \text{(Def } +_i\text{)}$$

For non-iterated revisions (where further revisions that follow upon a primary revision are not considered), the two definitions are equivalent as regards the resulting belief set. That is, $K(H + A) = K(H +_i A)$ for all levels i. As we do not use iterated revisions, we chose the first definition, for the sake of simplicity.

Once expansions and contractions have been explained for prioritised belief bases, we can define revisions using the Levi identity in a straightforward manner:

$$\mathbf{H} * A =_{df} (\mathbf{H} \div \neg A) + A = (\bigcap \mathbf{H}\perp \neg A) + A. \qquad \text{(Def } *\text{)}$$

The last step is to define the belief set K of a prioritised belief base $\mathbf{H}$ by the inferential closure of some suitable logic. We simply take the consequence operation of classical logic:

$$K(\mathbf{H}) =_{df} Cn(H_1 \cup \ldots \cup H_n). \qquad (K(\mathbf{H}))$$

Once $K(\mathbf{H})$, $\mathbf{H} + A$, and $\mathbf{H} * A$ are defined, we can readily define the revision and the contraction of the belief set $K(\mathbf{H})$ of the belief base $\mathbf{H}$:

$$K(\mathbf{H}) + A =_{df} K(\mathbf{H} + A)$$

$$K(\mathbf{H}) * A =_{df} K(\mathbf{H} * A).$$

[1] The definition of $\mathbf{H}\perp A$ is inspired by the construction of a preferred subtheory in Brewka (1991).

Appendix D

Proof of Chapter 4

We show that, within the original AGM belief revision theory, Lewis's (1973c) notion of counterfactual dependence is sufficient for causation as understood by our analysis when the cause temporally precedes its effect. Specifically, if $C, E \in K$ for contingent propositions C, E, and counterfactual dependence is transcribed into the AGM framework as $\neg C > \neg E \in K$, it follows that C is a cause of E according to our analysis, where the epistemic state S is a classic belief set K, as opposed to a prioritised belief base $\mathbf{H}$.

Proposition 9. Assume (i) $C, E \in K$ for contingent propositions C, E, and (ii) $\neg C > \neg E \in K$. Then C is a cause of E relative to K in the sense of the following conditions that mirror (C1), (C3), and (C4) of definition 22:

(C1') $C, E \in K$,

(C3') $C \gg E \in K$, and

(C4') $\neg C \gg E \notin K$.

Proof. (C1') is satisfied by assumption. First, we show that (C3') is satisfied. From (ii), we obtain by (RT_G): $\neg C > \neg E \in K$ iff $\neg E \in K * \neg C$. By the Levi identity, we have $\neg E \in (K - C) + \neg C$. But then $E \notin (K - C)$. For suppose it were. Then $(K - C) + \neg C$ would be inconsistent, because it would contain both E and $\neg E$. (The inconsistency is excluded by the assumption in (i) that C and E are contingent propositions.) Since $K - (C \vee E) \subseteq (K - C)$ and $E \notin (K - C)$, $E \notin K - (C \vee E)$. By the recovery postulate, however, $C \vee E \rightarrow E \in K - (C \vee E)$, and thus, by closure, $C \rightarrow E \in K - (C \vee E)$. Therefore, $E \in K - (C \vee E) + C$, and so $C \gg E \in K$.

It is left to show that (C4') is satisfied. Again, from (ii) by (RT_G) and the Levi identity, we obtain $\neg E \in (K - C) + \neg C$, and thus $E \notin (K - C) + \neg C$. Since $(K - (C \vee E)) + \neg C \subseteq (K - C) + \neg C$, we obtain that $E \notin (K - (C \vee E)) + \neg C$. Hence, $\neg C \gg E \notin K$. □

Observe that we have not considered here the requirement of forward-directedness and (C2), that is $t(C) < t(E)$. Otherwise, however, counterfactual dependence as encoded in $\neg C > \neg E \in K$ and actuality as encoded in (C1') are sufficient for causation according to our analysis within the AGM framework. If we switch to our framework, using prioritised belief bases, (C1), (C3), and (C4) are satisfied just in case the generalisation $C \rightarrow E$ is in $K(\mathbf{H})$ and $\neg C \rightarrow E$ is not. Of course, if the 'counterfactual' $\neg C \rightarrow \neg E$ is in $K(\mathbf{H})$ as a generalisation, $\neg C \rightarrow E$ is not.

Bibliography

Adams, E. W. Subjunctive and Indicative Conditionals. *Foundations of Language*, 6:89–94, 1970.

Adams, E. The Logic of Conditionals. *Inquiry : An Interdisciplinary Journal of Philosophy*, 8(1-4): 166–197, 1965.

Adams, E. W. Probability and the Logic of Conditionals. In Hintikka, J. and Suppes, P., editors, *Aspects of Inductive Logic*, volume 43 of *Studies in Logic and the Foundations of Mathematics*, pages 265 – 316. Elsevier, 1966.

Adams, E. W. *The Logic of Conditionals*. D. Reidel Publishing Co., Dordrecht, 1975.

Alchourrón, M. A., Gärdenfors, P., and Makinson, D. On the Logic of Theory Change: Partial Meet Contraction Functions and Their Associated Revision Functions. *Journal of Symbolic Logic*, 50: 510–530, 1985.

Andreas, H. and Günther, M. A Ramsey Test Analysis of Causation for Causal Models. *British Journal for the Philosophy of Science*, December 2018.

Andreas, H. and Günther, M. On the Ramsey Test Analysis of 'Because'. *Erkenntnis*, Jun 2018.

Andreas, H. and Günther, M. Causation in Terms of Production. *Philosophical Studies*, Mar 2019.

Arlo-Costa, H. and Egré, P. The Logic of Conditionals. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

Baumgartner, M. Interventionist Causal Exclusion and Non-Reductive Physicalism. *International Studies in the Philosophy of Science*, 23(2):161–178, 2009.

Bennett, J. *A Philosophical Guide to Conditionals*. Oxford University Press, Oxford, 2003.

Bennett, J. Counterfactuals and Possible Worlds. *Canadian Journal of Philosophy*, 4(December): 381–402, 1974.

Bradley, R. Restricting Preservation: A Response to Hill. *Mind*, 121(481):147–159, 2012.

Bradley, R. Radical Probabilism and Bayesian Conditioning*. *Philosophy of Science*, 72(2):342–364, 2005.

Bradley, R. A Defence of the Ramsey Test. *Mind*, 116(461):1–21, 2007.

Bradley, R. Supporters and Underminers: Reply to Chandler. *Mind*, 126(502):603–608, 2017.

Brewka, G. Belief Revision in a Framework for Default Reasoning. In *Proceedings of the Workshop on The Logic of Theory Change*, pages 602–622, London, 1991. Springer.

Carter, M. and Shieh, J. *Guide to Research Techniques in Neuroscience*. Elsevier Science, 2015.

Chandler, J. Transmission Failure, AGM-Style. *Erkenntnis*, 78(2):383–398, 2013.

Chandler, J. Preservation, Commutativity and Modus Ponens: Two Recent Triviality Results. *Mind*, 126(502):579–602, 2017.

Chellas, B. F. *Modal Logic: An Introduction*. Cambridge University Press, 1980.

Cohen, J. and Callender, C. A better best system account of lawhood. *Philosophical Studies*, 145(1): 1–34, 2009.

Crestani, F. Logical Imaging and Probabilistic Information Retrieval. In Crestani, F., Lalmas, M., and van Rijsbergen, C. J., editors, *Information Retrieval: Uncertainty and Logics: Advanced Models for the Representation and Retrieval of Information*, pages 247–279. Springer US, Boston, MA, 1998.

Dalal, M. Investigations into a Theory of Knowledge Base Revisions: Preliminary Report. In *Proceedings of the 7th National Conference on Artificial Intelligence (AAAI-88), St. Paul*, pages 475–479. 1988.

Davidson, D. Actions, Reasons, and Causes. *Journal of Philosophy*, 60(23):685–700, 1963.

Davis, W. A. Indicative and Subjunctive Conditionals. *Philosophical Review*, 88(4):544–564, 1979.

Descartes, R. *Meditationes De Prima Philosophia*. Librairie Philosophique J. Vrin, Paris, France, 1641/1978.

Dijkstra, N. and de Bruin, L. Cognitive Neuroscience and Causal Inference: Implications for Psychiatry. *Frontiers in Psychiatry*, 129, 2016.

Douven, I. Learning Conditional Information. *Mind & Language*, 27(3):239–263, 2012.

Douven, I. *The Epistemology of Indicative Conditionals: Formal and Empirical Approaches*. Cambridge University Press, 2015.

Douven, I. and Dietz, R. A Puzzle About Stalnaker's Hypothesis. *Topoi*, 30(1):31–37, 2011.

Douven, I. and Pfeifer, N. Formal Epistemology and the New Paradigm Psychology of Reasoning. *Review of Philosophy and Psychology*, 5:199–221, 2014.

Douven, I. and Romeijn, J.-W. A New Resolution of the Judy Benjamin Problem. *Mind*, 120(479): 637–670, 2011.

Douven, I. and Verbrugge, S. The Adams family. *Cognition*, 117(3):302–318, 2010.

Edgington, D. On Conditionals. *Mind*, 104(414):235–329, 1995.

Edgington, D. I–The Presidential Address: Counterfactuals. *Proceedings of the Aristotelian Society*, 108(1):1–21, 2008.

Eiter, T. and Lukasiewicz, T. Complexity results for structure-based causality. *Artificial Intelligence*, 142(1):53 – 89, 2002.

Evans, J. S. B. T. and Over, D. E. *If*. Oxford University Press, Oxford, 2004.

Fenton-Glynn, L. A Proposed Probabilistic Extension of the Halpern and Pearl Definition of 'Actual Cause'. *British Journal for the Philosophy of Science*, 68:1061–1124, 2017.

Fine, K. Review of Lewis's Counterfactuals. *Mind*, 84(335):451–458, 1975.

Fodor, J. A. Special Sciences (Or: The Disunity of Science as a Working Hypothesis). *Synthese*, 28 (2):97–115, 1974.

Fraassen, B. C. V. Singular Terms, Truth-Value Gaps, and Free Logic. *Journal of Philosophy*, 63(17): 481–495, 1966.

Fuhrmann, A. and Hansson, S. O. A Survey of Multiple Contractions. *Journal of Logic, Language and Information*, 3(1):39–75, 1994.

Gärdenfors, P. Conditionals and Changes of Belief. In Niiniluoto, I. and Tuomela, R., editors, *The Logic and Epistemology of Scientific Change*, volume 30 of *Acta Philosophica Fennica*, pages 381–404. 1978.

Gärdenfors, P. *Knowledge in Flux*. MIT Press, Cambridge, MA, 1988.

Gärdenfors, P. and Makinson, D. Revision of Knowledge Systems Using Epistemic Entrenchment. In Vardi, M., editor, *TARK' 88 - Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 83–95. Morgan and Kaufmann, Los Altos, 1988.

Gärdenfors, P. Belief Revisions and the Ramsey Test for Conditionals. *The Philosophical Review*, 95 (1):81–93, 1986.

Gebharter, A. Causal Exclusion and Causal Bayes Nets. *Philosophy and Phenomenological Research*, 95(2):353–375, 2017.

George, M. S., Nahas, Z., Lisanby, S. H., Schlaepfer, T., Kozel, F. A., and Greenberg, B. D. Transcranial magnetic stimulation. *Neurosurgery Clinics*, 14(2):283–301, 2003.

Goodman, N. The Problem of Counterfactual Conditionals. *Journal of Philosophy*, 44(5):113–128, 1947.

Gravetter, F. and Forzano, L. *Research Methods for the Behavioral Sciences*. Cengage Learning, 2011.

Grove, A. Two Modellings for Theory Change. *Journal of Philosophical Logic*, 17:157–170, 1988.

Günther, M. Learning Conditional and Causal Information by Jeffrey Imaging on Stalnaker Conditionals. *Organon F*, 24(4):456–486, 2017a.

Günther, M. Disjunctive Antecedents for Causal Models. In Cremers, A., van Gessel, T., and Roelofsen, F., editors, *Proceedings of the 21st Amsterdam Colloquium*, pages 25–35, Amsterdam, The Netherlands, 2017b. Institute of Logic, Language, and Computation (ILLC), University of Amsterdam.

Günther, M. Learning Conditional Information by Jeffrey Imaging on Stalnaker Conditionals. *Journal of Philosophical Logic*, 47(5):851–876, Oct 2018.

Hájek, A. Probabilities of Counterfactuals and Counterfactual Probabilities. *Journal of Applied Logic*, 12(3):235 – 251, 2014. Special Issue on Combining Probability and Logic to Solve Philosophical Problems.

Hall, N. Structural Equations and Causation. *Philosophical Studies*, 132(1):109–136, 2007.

Hall, N. Two Concepts of Causation. In John Collins, Ned Hall, and Laurie Paul, editors, *Causation and Counterfactuals*, pages 225–276. The MIT Press, 2004.

Halpern, J. Y. A Modification of the Halpern-Pearl Definition of Causality. 2015.

Halpern, J. Y. and Hitchcock, C. Actual Causation and the Art of Modeling. In Dechter, R., Geffner, H., and Halpern, J. Y., editors, *Heuristics, Probability, and Causality: a Tribute to Judea Pearl*, pages 383–406. College Publications, London, 2010.

Halpern, J. Y. and Hitchcock, C. Graded Causation and Defaults. *British Journal for the Philosophy of Science*, 66(2):413–457, 2015.

Halpern, J. Y. and Pearl, J. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *British Journal for the Philosophy of Science*, 56(4):843–887, 2005.

Hansson, S. O. *A Textbook of Belief Dynamics. Theory Change and Database Updating*. Kluwer, Dordrecht, 1999.

Hansson, S. O. Belief Contraction Without Recovery. *Studia Logica*, 50(2):251–260, 1991.

Hansson, S. O. In Defense of the Ramsey Test. *Journal of Philosophy*, 89(10):522–540, 1992.

Harman, G. *Change in View*. MIT Press, Cambridge, Massachusetts, 1986.

Hartmann, S. and Rad, S. R. Learning Indicative Conditionals. *Unpublished manuscript*, pages 1–28, 2017.

Hill, B. Defending the Ramsey Test: What is Wrong with Preservation? *Mind*, 121(481):131–146, 2012.

Hitchcock, C. The Intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy*, 98(6):273–299, 2001.

Hitchcock, C. Prevention, Preemption, and the Principle of Sufficient Reason. *Philosophical Review*, 116(4):495–532, 2007.

Hitchcock, C. Theories of Causation and the Causal Exclusion Argument. *Journal of Consciousness Studies*, 19(5-6):40–56, 2012.

Hume, D. *A Treatise of Human Nature*. Oxford: Clarendon Press, 1739/1978.

Hume, D. *An Enquiry Concerning Human Understanding*. Printed for A. Millar, London, 1748.

Husserl, E. Ideen zu einer reinen Phänomenologie und Phänomenologischen Philosophie I. In *Jahrbuch für Philosophie und phänomenologische Forschung*, volume 1, pages 1–323. Max Niemeyer, Halle a.d.S., Germany, 1913.

Jeffrey, R. C. *The Logic of Decision*. Mc Graw-Hill, New York, 1965.

Jeffrey, R. C. Statistical Explanation Vs. Statistical Inference. In Rescher, N., editor, *Essays in Honor of Carl G. Hempel*, pages 104–113. Reidel, 1969.

Kant, I. *Kritik der reinen Vernunft*. Philosophische Bibliothek. Felix Meiner, Hamburg, 4 (1998) edition, 1781/87.

Kant, I. *Prolegomena zu einer jeden künftigen Metaphysik, die als Wissenschaft wird auftreten können*. Philosophische Bibliothek 540. Felix Meiner, Hamburg, 4 (2001) edition, 1783.

Khoo, J. Backtracking Counterfactuals Revisited. *Mind*, 126(503):841–910, 2017.

Kim, J. *Physicalism, Or Something Near Enough*. Princeton monographs in philosophy. Princeton University Press, Princeton and Oxford, 2005.

Kim, J. Causes and Counterfactuals. *Journal of Philosophy*, 70(17):570–572, 1973.

Kraus, S., Lehmann, D., and Madigor, M. Nonmonotonic Reasoning, Preferential Models and Cumulative Logics. *Artificial Intelligence*, 44:167–207, 1990.

Kripke, S. A. Semantical Analysis of Modal Logic I. Normal Propositional Calculi. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 9(56):67–96, 1963.

Kyburg, H. E. Salmon's Paper. *Philosophy of Science*, 32(2):147–151, 1965.

Leitgeb, H. On the Ramsey Test Without Triviality. *Notre Dame Journal of Formal Logic*, 51(1): 21–54, 2010.

Leitgeb, H. A Probabilistic Semantics for Counterfactuals. Part A. *The Review of Symbolic Logic*, 5 (1):26–84, 2012a.

Leitgeb, H. A Probabilistic Semantics for Counterfactuals. Part B. *The Review of Symbolic Logic*, 5 (1):85–121, 2012b.

Levi, I. *Mild Contraction: Evaluating Loss of Information Due to Loss of Belief*. Oxford University Press, Oxford, 2004.

Levi, I. *For the Sake of the Argument: Ramsey Test Conditionals, Inductive Inference and Nonmonotonic Reasoning*. Cambridge University Press, Cambridge, 2007.

Levi, I. Iteration of Conditionals and the Ramsey Test. *Synthese*, 76(1):49–81, 1988.

Lewis, D. *Philosophical Papers II*. Oxford University Press, Oxford, 1986a.

Lewis, D. Counterfactual Dependence and Time's Arrow. *Noûs*, 13(4):455–476, 1979.

Lewis, D. Postscripts to 'Causation'. In Lewis, D., editor, *Philosophical Papers Vol. II*. Oxford University Press, 1986b.

Lewis, D. Events. In Lewis, D., editor, *Philosophical Papers Vol. II*, pages 241–269. Oxford University Press, 1986c.

Lewis, D. Humean Supervenience Debugged. *Mind*, 103(412):473–490, 1994.

Lewis, D. Causation as Influence. *Journal of Philosophy*, 97(4):182–197, 2000.

Lewis, D. K. *Counterfactuals*. Blackwell Publishers, Massachusetts, 1973a.

Lewis, D. K. Counterfactuals and Comparative Possibility. *Journal of Philosophical Logic*, pages 418–446, 1973b.

Lewis, D. K. Causation. *Journal of Philosophy*, 70(17):556–567, 1973c.

Lewis, D. K. Probabilities of Conditionals and Conditional Probabilities. *The Philosophical Review*, 85(3):297–315, 1976.

Lewis, D. *On the Plurality of Worlds*. Blackwell, Oxford, 1986d.

List, C. and Menzies, P. Nonreductive Physicalism and the Limits of the Exclusion Principle. *Journal of Philosophy*, 106(9):475–502, 2009.

Makinson, D. On the Status of the Postulate of Recovery in the Logic of Theory Change. *Journal of Philosophical Logic*, 16(4):383–394, 1987.

Martin, A. and Gotts, S. J. Making the Causal Link: Frontal Cortex Activity and Repetition Priming. *Nature Neuroscience*, 8(9):1134–1135, 2005.

McLaughlin, B. P. Varieties of Supervenience. In Savellos, E. E. and Yalcin, U., editors, *Supervenience: New Essays*, pages 16–59. Cambridge University Press, 1995.

Mele, A. R. *Springs of Action: Understanding Intentional Behavior*. Oxford University Press, Oxford and New York, 1992.

Oaksford, M. and Chater, N. *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford cognitive science series. Oxford University Press, Oxford, New York, 2007.

Papadimitriou, C. H. *Computational Complexity*. Addison-Wesley Publishing Company, Inc., Massachusetts, 1994.

Papineau, D. 10 The Rise of Physicalism. In Stone, M. W. F. and Wolff, J., editors, *The Proper Ambition of Science*, pages 2–174. Routledge, 2000.

Paul, L. A. Aspect Causation. *The Journal of Philosophy*, 97(4):235–256, 2000.

Paul, L. A. and Hall, N. *Causation: A User's Guide*. Oxford University Press, Oxford, 2013.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

Perlmutter, J. S. and Mink, J. W. Deep Brain Stimulation. *Annual Review of Neuroscience*, 29: 229–257, 2006.

Quine, W. V. O. *Methods of Logic*. A Holt-Dryden book. Holt, 1959.

Ramachandran, M. A Counterfactual Analysis of Causation. *Mind*, 106(422):263–277, 1997.

Ramachandran, R., Ramer, A., and Nayak, A. C. Probabilistic Belief Contraction. *Minds and Machines*, 22(4):325–351, 2012.

Ramsey, F. P. General Propositions and Causality. In Braithwaite, R. B., editor, *Foundations of Mathematics and other Logical Essays*, pages 237–257. Humanities Press, New York, 1950.

Ramsey, F. P. Universals of Law and of Fact. In *Foundations*, pages 145–163. Routledge & Kegan Paul, London, 1928/1978.

Ramsey, F. P. General Propositions and Causality. *Philosophical Papers*, pages 145–163, 1929/1990.

Romei, V., Thut, G., Mok, R. M., Schyns, P. G., and Driver, J. Causal Implication by Rhythmic Transcranial Magnetic Stimulation of Alpha Frequency in Feature-Based Local vs. Global Attention. *European Journal of Neuroscience*, 35(6):968–974, 2012.

Rott, H. Reapproaching Ramsey: Conditionals and Iterated Belief Change in the Spirit of AGM. *Journal of Philosophical Logic*, 40:155–191, 2011.

Rott, H. Ifs, Though, and Because. *Erkenntnis*, 25(3):345–370, 1986.

Rott, H. Two Dogmas of Belief Revision. *Journal of Philosophy*, 97:503–522, 2000.

Rott, H. Preservation and Postulation: Lessons From the New Debate on the Ramsey Test. *Mind*, 126 (502):609–626, 2017.

Rott, H. and Pagnucco, M. Severe Withdrawal (and Recovery). *Journal of Philosophical Logic*, 28 (5):501–547, 1999.

Sack, A. T. Transcranial Magnetic Stimulation, Causal Structure, Malfunction Mapping and Networks of Functional Relevance. *Current Opinion in Neurobiology*, 16(5):593 – 599, 2006.

Sartorio, C. Causes As Difference-Makers. *Philosophical Studies*, 123(1):71–96, Mar 2005.

Sartorio, C. Disjunctive Causes. *Journal of Philosophy*, 103(10):521–538, 2006.

Schaffer, J. Contrastive Causation. *The Philosophical Review*, 114(3):327–358, 2005.

Sebastiani, F. Information Retrieval, Imaging and Probabilistic Logic. *Computers and Artificial Intelligence*, 17(1):1–16, 1998.

Smart, J. Sensations and Brain Processes. *Philosophical Review*, 68(April):141–56, 1959.

Spohn, W. Causation: An Alternative. *British Journal for the Philosophy of Science*, 57(1):93–119, 2006.

Squire, L., Bloom, F., Spitzer, N., Squire, L., Berg, D., du Lac, S., and Ghosh, A. *Fundamental Neuroscience*. Fundamental Neuroscience Series. Elsevier Science, Amsterdam, Boston, Heidelberg, London, New York, Oxford, San Francisco, Singapore, Sydney, Tokio, 2008.

Stalnaker, R. A Theory of Conditionals. In Rescher, N., editor, *Studies in Logical Theory (American Philosophical Quarterly Monograph Series)*, pages 98–112. Blackwell, Oxford, 1968.

Stalnaker, R. *Ways a World Might Be: Metaphysical and Anti-Metaphysical Essays*. Oxford scholarship online. Clarendon Press, 2003.

Stalnaker, R. *Inquiry*. Cambridge University Press, Massachusetts, 1984.

Stalnaker, R. C. A Theory of Conditionals. In Sosa, E., editor, *Causation and Conditionals*, pages 165–179. OUP, 1975.

Stalnaker, R. C. *A Defense of Conditional Excluded Middle*, pages 87–104. Springer Netherlands, Dordrecht, 1981.

Stalnaker, R. C. and Thomason, R. H. A Semantic Analysis of Conditional Logic. *Theoria*, 36(1): 23–42, 1970.

Unterhuber, M. *Possible Worlds Semantics for Indicative and Counterfactual Conditionals? A Formal Philosophical Inquiry Into Chellas-Segerberg Semantics*. Ontos (now De Gruyter), Frankfurt, Paris, Lancaster, New Brunswick, 2013.

Van Benthem, J. Logic and Reasoning: Do the Facts Matter? *Studia Logica*, 88:67–84, 2008.

Van Benthem, J. and Smets, S. Dynamic Logics of Belief Change. In Ditmarsch, H. V., Halpern, J. Y., der Hoek, W. V., and Kooi., B., editors, *Handbook of Logics for Knowledge and Belief*, chapter 7, pages 299–368. College Publications, 2015.

Van Fraassen, B. C. Rational Belief and Probability Kinematics. *Philosophy of Science*, 47(2):165–187, 1980a.

Van Fraassen, B. C. A Problem for Relative Information Minimizers in Probability Kinematics. *The British Journal for the Philosophy of Science*, 32(4):375–379, 1981.

Van Fraassen, B. C. *Laws and Symmetry*. Oxford University Press, Oxford, 1989.

Van Fraassen, B. C., Hughes, R. I. G., and Harman, G. A Problem for Relative Information Minimizers, Continued. *The British Journal for the Philosophy of Science*, 37(4):453–463, 1986.

Van Fraassen, B. *The Scientific Image*. Clarendon Library of Logic and Philosophy. Clarendon Press, Oxford, 1980b.

Williams, P. M. Bayesian Conditionalisation and the Principle of Minimum Information. *British Journal for the Philosophy of Science*, 31(2):131–144, 1980.

Windhorst, U. and Johansson, H. *Modern Techniques in Neuroscience Research*. Lab Manuals Series. Springer, 1999.

Woodward, J. *Making Things Happen : A Theory of Causal Explanation*. Oxford University Press, Oxford, 2003.

Woodward, J. *Making Things Happen: A Theory of Causal Explanation*. Oxford Studies in the Philosophy of Science. Oxford University Press, 2005.

Woodward, J. Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research*, 91(2):303–347, 2015.

Yablo, S. De Facto Dependence. *The Journal of Philosophy*, 99(3):130–148, 2002.

Zhao, J., Crupi, V., Tentori, K., Fitelson, B., and Osherson, D. Updating: Learning versus Supposing. *Cognition*, 124(3):373–378, 2012.

CV

Research Profile

AOS: Epistemology, Causation

AOC: Logic, Philosophy of Language, Philosophy of Cognitive Science, General Philosophy of Science

Education

Now **PhD in Philosophy**
"Learning, Conditionals, Causation"
MCMP & GSN, LMU Munich
Hannes Leitgeb, Stephan Hartmann, Stephan Sellmaier, Holger Andreas

2015 **Master in Logic and Philosophy of Science**
"Non-Monotonic Logics, Artificial Neural Networks, and Complexity Considerations in the Horn Case"
MCMP, LMU Munich
Olivier Roy, Hannes Leitgeb, Gregory Wheeler

2012 **Bachelor in Philosophy and Cognitive Science**
"The Methodology of Scientific Research Programmes as Modification of Popper's Falsificationism"
Albert Ludwig University Freiburg i. Br.
Maarten J. F. M. Hoenen

References

Prof. DDr. Hannes Leitgeb
Chair of Logic and Philosophy of Language
Head of the Munich Center for Mathematical Philosophy
LMU Munich
+49 (0) 89 / 2180 - 6171
Hannes.Leitgeb@lmu.de

Prof. Dr. Stephan Hartmann

Chair of Philosophy of Science
Head of the Munich Center for Mathematical Philosophy
LMU Munich
+49 (0) 89 / 2180 - 3320
S.Hartmann@lmu.de

Prof. Dr. Olivier Roy
Chair of Philosophy 1
Department Chair of Philosophy
University of Bayreuth
+49 (0) 921 55-4151
Olivier.Roy@uni-bayreuth.de

Prof. Dr. Igor Douven
CNRS Research Professor
Université Paris IV – La Sorbonne
+32-(0)4-2864025
Igor.Douven@sorbonne-universite.fr

Prof. Dr. Gregory Wheeler
Professor of Theoretical Philosophy & Computer Science
Frankfurt School of Finance and Management
G.Wheeler@fs.de

Prof. Dr. Dr. hc. Julian Nida-Rümelin, Former Minister of State
Chair of Philosophy and Political Theory
Department Chair of Philosophy
LMU Munich
+49 (0) 89 2180-6180
Julian.Nida-Ruemelin@lrz.uni-muenchen.de

List of Publications

2019 "Causation in Terms of Production" (with Holger Andreas)
Philosophical Studies (forthcoming)

2018 "A Ramsey Test Analysis of Causation for Causal Models" (with Holger Andreas)
The British Journal for the Philosophy of Science (forthcoming)

2018 "On the Ramsey Test Analysis of 'Because"' (with Holger Andreas)
Erkenntnis (forthcoming)

2018 "Learning Conditional Information by Jeffrey Imaging on Stalnaker Conditionals"
Journal of Philosophical Logic, (47):851–876.

2017 "Learning Conditional and Causal Information by Jeffrey Imaging on Stalnaker Conditionals"
Organon F, 24 (4):456-486.

2017 "Disjunctive Antecedents for Causal Models"
Proceedings of the 21st Amsterdam Colloquium, 25-34.

Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation “Learning, Conditionals, Causation” selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation “Learning, Conditionals, Causation” is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

Mario Günther

München, den/Munich, date: April, 16, 2019 Unterschrift/Signature

Declaration of Author Contributions

I hereby declare that this thesis has not been submitted for any other degree or professional qualification.

Chapter 2 is based on two papers. Günther (2018) is published in the *Journal of Philosophical Logic*, and Günther (2017a) in the journal *Organon F*.

Chapter 3 is based on a joint paper with Holger Andreas. Andreas and Günther (2018) is published in the journal *Erkenntnis*. Both authors contributed equally to this manuscript.

Chapter 4 is based on a joint paper with Holger Andreas. Andreas and Günther (2019) has been accepted by the journal *Philosophical Studies*. Both authors contributed equally to this manuscript.

Chapter 5 is based on a joint paper with Holger Andreas. Andreas and Günther (2018) has been accepted by the journal *The British Journal for the Philosophy of Science*. Both authors contributed equally to this manuscript.

Chapter 6 is based on Günther (2017b).

Munich, date Prof. DDr. Hannes Leitgeb Prof. Dr. Holger Andreas

Mario Günther