Submitted December 11th, 2013.

Supervisors:

Prof. dr. Joke Meheus, Universiteit Gent
Dr. Bert Leuridan, Universiteit Gent

Reading Committee:

Prof. dr. Atocha Aliseda, Universidad Nacional Autónoma de México
Prof. dr. Igor Douven, Rijksuniversiteit Groningen
Dr. Michela Massimi, The University of Edinburgh
Prof. dr. Maarten Van Dyck, Universiteit Gent
Prof. dr. Erik Weber, Universiteit Gent

This thesis was typeset in LaTeX.
Cover Image: Daya Bay Neutrino Experiment
(Courtesy of Brookhaven National Laboratory)
Cover Design: Gitte Callaert

UNIVERSITEIT
GENT

Faculteit Letteren en Wijsbegeerte

Tjerk Gauderis

Patterns of Hypothesis Formation

At the crossroads of Philosophy of Science, Logic, Epistemology, Artificial Intelligence and Physics

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wijsbegeerte

Promotoren: Prof. Dr. Joke Meheus en Dr. Bert Leuridan

to my daughter Ada,
who explains it all

acknowledgments

This piece of work would never have emerged in its present form, were it not for the support of many. In particular, I want to thank my love Nel, for being there and complementing me; my daughter Ada for bringing her liveliness into our home: her look on the world was and is a constant inspiration; my supervisor Joke for allowing me to pursue my own path in these muddy waters; my co-supervisor Bert for his detailed and constructive remarks on every part of this thesis; the Centre for Logic and Philosophy of Science for giving me a four years grant; its current and former employees for their many useful comments and ideas over the years; my friends, they mean more to me than they realize; and finally, the world itself, for being such an intriguing and fascinating place to dwell in.

Contents

1 Introduction

The Muses love alternatives.

— Virgil, *Eclogues*, Book III

How do scientists form new hypotheses?

Are there general patterns of hypothesis formation?

If so, is the list of patterns humans are capable of limited?

Are there better or worse patterns to form new hypotheses?

Can we simulate processes of hypothesis formation?

How do scientists choose if different patterns are available?

How do hypotheses relate to scientific models?

How do scientists deal with mutually exclusive hypotheses?

Can we rationally endorse mutually exclusive alternatives?

Are hypotheses part of our daily life?

Can we live with uncertainties?

These are the general questions that motivated and led to this dissertation, a tour which took me through various philosophical subdisciplines to obtain as rich as possible a view of the topic of hypothesis formation in science.

In this general introduction, I will situate my main topic, give an overview of the existing literature and address some common themes. More specific introductions to the philosophical methods I will use can be found in the separate introductions to the various parts.

1.1 General Topic of the Dissertation

The topic of this PhD thesis is the formation and use of hypotheses in science, with a focus on the physical sciences. According to the view I develop, scientific hypotheses can be defined as follows:

> *Scientific hypotheses* are statements about the empirical world with an unknown or underdetermined truth status that are advanced as a tentative answer to a particular research question.

The different aspects of this definition of scientific hypotheses are argued for and assumed at various places throughout the dissertation, and a full elaboration of this definition can be found in Section 9.2. In the remainder of this section, I want to point out some delineating choices and general features to set the scene. More detailed argumentation will be given in due course.

Doxastic Attitude In forming a hypothesis an agent adopts at the same time a specific attitude or cognitive relation towards that hypothesis. As scientific reasoning should not be considered to be isolated from general reasoning in every day life, this attitude, which will be determined as a doxastic attitude *sui generis*, can be studied from a more general perspective in the field of epistemology.

Truth-purposiveness In this thesis, I deal only with *truth-purposive* hypotheses, hypotheses which the agent considers as neither true nor false, but as unknown or even underdetermined. The agent's main purpose in forming such hypotheses is either to determine their truth value or to use their possible truth for other epistemic purposes. This class of hypotheses has to be sharply distinguished from *truth-denying hypotheses*, hypotheses the agent considers to be false, but which can be useful for her, for instance, in setting up a thought experiment.

Reference to the Empirical World My topic is further limited by the fact that I consider only hypotheses that make reference to the empirical world. In other words, I do not focus on mathematical conjectures or hypothesized conceptual relations. While the analysis of the epistemological and logical parts of this dissertation may probably be extended to include this kind of hypotheses, my analysis of the use of hypotheses in scientific practice is tailored specifically for hypotheses that refer to the empirical world.

The Importance of a Research Question Not just any conjectural statement can be considered as a hypothesis. Scientific hypotheses are formed in response to and in relation with a certain *research question*, which acts as a *trigger* for them. Although hypotheses are sometimes presented more or less independently at later stages of research, they cannot be fully understood if they are disconnected from their triggers.

Hypothesis Formation and Rationality In this thesis I assume throughout the possibility of rational hypothesis formation, both in an absolute and a relative sense. More precisely, I assume that it can be rationally justified to suggest a hypothesis in answer to a particular trigger, and I assume that it can be rationally justified to prefer one method of hypothesis formation above another one in a particular case. Yet I do not assume that there is always a difference: two methods can be (equally) rationally justified in a particular case. Also, rational hypothesis formation does not guarantee higher truth probabilities for individual hypotheses. It is assumed, however, that rational hypothesis formation serves the agent's epistemic interests better (largely because of a better understanding and coverage of the space of possibilities).

Scientific Discovery and Abductive Reasoning In the literature, hypothesis formation in science is generally brought into connection with the concepts of 'scientific discovery' and 'abductive reasoning'. However, as the precise meaning of these notions can be ambiguous, I will first give an overview of the literature on hypothesis formation, abduction and scientific discovery and clarify how the main questions of this dissertation connect to the different research challenges presented in this literature. This will then allow me to specify how the relation between 'hypothesis formation', 'abductive reasoning' and 'scientific discovery' will be conceived in this thesis.

1.2 Overview of the Literature on Hypothesis Formation

This part is based on the second section of Gauderis (2013b). The rest of that paper is included as Chapter 7 of this dissertation.

In this section, I summarize the status quo of research on hypothesis formation in the philosophy of science by reflecting on how the main questions of today were shaped over the course of the twentieth century. As already suggested in the previous section, this overview relates to both the literature on scientific discovery and the literature on abductive reasoning.

In the first half of the twentieth century, at the height of logical positivism, philosophers generally held that the mind's ability to generate new hypotheses is situated outside the realm of rational thinking.[1] This idea was crystalized in Reichenbach's very influential distinction between the *context of discovery* and the *context of justification* (Reichenbach, 1938; Schickore and Steinle, 2006; Laudan, 1980; Kitcher, 2013), which underlies and gives expression to the longstanding prejudice that scientific discovery (and hypothesis formation) were not proper matters of interest for the philosophy of science (e.g. Popper, 1959; Laudan, 1980, p. 180).

From the different perspective of American pragmatism, Charles S. Peirce had, however, already a few decades earlier advocated the idea that explanatory hypothesis formation is a distinctive and important form of rational inference, for which he used the notion 'abduction' (Peirce, 1958, CP 5.172).[2] This suggestion was picked up by Hanson, the pioneer of a generation of philosophers who based their reflections on a thorough familiarity with the history of science (e.g. Hanson, 1958, p. 3). By discussing how Kepler inferred the hypothesis of elliptical planetary motion from the observations made by Tycho Brahe, Hanson argued that Kepler's "keen logical sense" is shown in the sound reasons he cited for every step he made, steps whose explanatory character prevents us from classifying them as merely inductive generalizations from the available data (1958, pp. 84-85). Although Hanson has been rightly criticized for underestimating the role of theoretical and other constraints in scientific discovery (Nickles, 1980, p. 23; Darden, 1991, p. 10) and for confusing the actual generation of hypotheses with their preliminary evaluation (Schaffner, 1980, p. 179), his observation that scientific hypothesis formation is a reasonable affair has been confirmed over and over again by philosophers of science who have

[1] It was not that these early philosophers of science claimed that full-fledged theories could originate from bold leaps of the imagination. Rather, as Meheus (1999) shows, they generally acknowledged and sometimes even discussed the use of rational search methods in scientific discovery. Only, for them, these methods relied essentially on the input of early hypotheses and particular interpretations, which could not themselves be derived by rational processes.

[2] Peirce distinguished abduction, the formation of explanatory hypotheses, not only from deduction but also from induction, the inference from cases to generalized statements. Although the distinction between abduction and induction can be put into question in its specifics (Aliseda, 2006, pp. 33-34), Peirce was certainly the first to argue for the rationality of explanatory reasoning (that is not based on generalizations of cases) since the general acceptance of the fallible nature of science in the mid-1800s.

extensively studied real historical discoveries (e.g. Franklin, 1993, p. 124; Darden, 1991, p. 3; Nersessian, 2008, p. 5).

Despite this widespread consensus, however, that scientific discovery processes and hypothesis formation can be addressed rationally, two things remain clear: (1) that the notion of 'abduction' or 'abductive reasoning' has not been broadly accepted, and, more importantly, (2) that the study of this particular type of inference has not been assigned a central place in research on scientific processes and discoveries. I distinguish three main reasons for this turn of events.

First, Peirce's original ideas on abduction, which were recorded over a period of decades, underwent several changes and lack a coherent interpretation (Anderson, 1986; Kapitan, 1992, p. 15; Plutynski, 2011). As a result, the concept of abduction refers to at least three different types of inference in the present literature: the generation of new (explanatory) hypotheses (e.g. Gabbay and Woods, 2006; Campos, 2011); the inference to the (truth of the) best explanation (e.g. Harman, 1965; Lipton, 1991; Douven, 2011); and the selection of the hypothesis that is most worthy of pursuit (McKaughan, 2008). Such a situation clearly leads to mutual misunderstanding if one neglects to specify the type of reasoning one is discussing.

Second, even if one does agree on the type of inference in question, there are even more interpretations concerning how broadly abductive reasoning should be conceived. Is it a particular and rather constrained formal reasoning pattern, as it is generally conceived in logics and AI (e.g. Flach and Kakas, 2000a; Gabbay and Kruse, 2000)? Or is it an all-encompassing scientific method, as Hanson (1958, 1961) and some proponents of the IBE view (e.g. McMullin, 1992) held it to be? In such a climate, it is also difficult to clearly draw the line between abduction and induction (Aliseda, 2006, pp. 33-34) and hardly any effort has been put into clarifying the relation between abductive reasoning and other well-studied practices in scientific discovery, such as the construction and refinement of models and the formation of new concepts. Aliseda, in her monograph on abduction, summarizes the situation as follows:

> Many authors write as if there were pre-ordained, reasonably clear notions of abduction and its rivals, which we only have to analyze to get a clear picture. But these technical terms may be irretrievably confused in their full generality, burdened with the

> debris of defunct philosophical theories. (Aliseda, 2006, p. 34)

Third, Hanson presented the rationality of discovery as if there existed a unified method of discovery, i.e. abduction, which could be linked to the old notion of a "logic of discovery" (Hanson, 1958, 1961). But for the next generation of scholars, it became clear that the old Baconian dream of a single subject-independent algorithmic procedure of discovery had been debunked by the complex subtlety of present-day theories, the long and arduous processes that led to them, and the important role of (previous) theories in scientific reasoning. Therefore, they distanced themselves from the Hansonian views in favor of a multitude of discovery methods; some even argued against the formal logical treatment of any particular pattern (Nickles, 1980, p. 23-28). This strong criticism led many to regard the literature on abduction as a more logically and epistemologically oriented side branch with a somewhat exegetical nature (e.g. Niiniluoto, 1999; Hintikka, 1998) only loosely connected to the mainly historically-oriented stem of research on scientific discovery in the philosophy of science. Still, in recent years the literature on abduction has, to a certain extent, bridged this gap by acknowledging the multitude of "patterns of abduction" and attempting to provide a classification of such formal patterns that can be used to address real historical cases (Schurz, 2008a).

In the 1960s and 70s, the rise of historicism in the philosophy of science showed that scientific confirmation is not the neat logical process it was once taken to be, and hence not so easily separable from the process of discovery (Nickles, 1980, p. 2). Around the same time, research in psychology and artificial intelligence showed that there are better and worse heuristics for problem solving. This opened the way, at least in principle, for the construction of normative theories for specific problem solving activities such as scientific discovery (Simon, 1973). These new insights led to the emergence of a group of philosophers whose research focused primarily on the process of scientific discovery, the so-called "friends of discovery". By around 1990, this loosely-knit group agreed on the following ideas, which still stand today: (1) that scientific research is a gradual step-by-step process of constant refinement (Darden, 1991, p. 11; Langley et al., 1987, pp. 57-59; Shah, 2007) and that, as such, there is no strict distinction between the context of discovery and the context of justification (Nickles, 1980, pp. 8-18; Hoyningen-Huene, 2006); (2) that discovery can be seen as a problem solving activity, and thus can be addressed rationally (Simon, Langley and Bradshaw, 1981; Nickles, 1978); (3) that there are no definite algorithms or logics of discovery, but only a plethora of heuristics, strategies and methods that are context- and subject-matter-dependent (Achinstein, 1980; Nickles, 1990; Darden, 1991, p. 11); (4) that both the hypothetico-deductive and inductive views of the scientific method are obsolete, as the first retains the old distinction between discovery and justification (Darden, 1991, pp. 9-17) while the second neglects the importance of theoretical constraints in scientific problem solving (Nickles, 1980, p. 35).

Where does this leave research on scientific discovery, hypothesis formation and abduction today? Although most agree on the central insights listed above, many different directions are now pursued, each with its own methodology. (1) Some philosophers have tried to further naturalize the insights about discovery processes by linking the patterns found in studying historical discoveries to the psychology literature (e.g. Nersessian, 2008; Thagard, 2012; Magnani, 2001). (2) Others have attempted to specify and classify the various particular patterns or strategies employed in scientific discovery in a more stringent way (e.g. Darden, 1991; Schurz, 2008a; Hoffmann, 2010), which has led to in-depth studies of lesser-known patterns or strategies (e.g. Darden and Craver, 2002; but also Gauderis and Van De Putte, 2012, see Chapter 6). (3) Given the general understanding of science as a step-by-step process (e.g. Blachowicz, 1998), some have attended closely to the construction and refinement of models in science, which they hold to be key instruments of investigation (Morgan and Morrison, 1999). Due to the heterogeneity of the class of models, this research varies wildly and is often subject-dependent, as the literature on mechanistic models in biology illustrates; it is also partially linked to the psychology literature via the research on model-based reasoning (e.g. Nersessian, 2008). Finally, (4) computational philosophers of science have continued to refine artificially intelligent agents in an effort to determine, in the spirit of Simon, which heuristics and weak problem-solvers might be efficient instruments of discovery (for an overview, see Darden, 1997).

Although the field has become disciplinarily fragmented in recent years, the research challenges that bind these different strands are very similar:[3] namely (1) to explicate the various heuristic and often context-dependent "patterns of discovery"[4] and the relation between them, (2) to provide

[3] Several scholars, such as Thagard and Darden, have contributed to more than one of these lines of research. In general, the various strands give each other's results a sympathetic reading.

[4] Hanson's phrase (1958) fits remarkably well here, though it does not refer exactly to the same thing; Hanson himself was more concerned with the 'discovery of (conceptual) patterns'

rational and normative guidance on their use, and (3) to pursue the possibilities of computational discovery and its relation to human discovery. All three of these challenges are daunting in scope, yet, given the results obtained in past decades, should not be considered unaddressable. In this thesis, I aim to contribute to each of these three challenges.

1.3 Four Parts, Four Approaches

This dissertation assesses the topic of hypothesis formation in science from four different angles, resulting in four main parts each using the methodology of a different subdiscipline of philosophy. The main questions that I address in each of these parts are motivated both by my own empirical study of the process of hypothesis formation in science and by the research challenges stated in the literature and the various branches into which it split.

Part I: Epistemological Considerations In this first part, I seek to understand the cognitive attitude towards a hypothesis from a broader perspective. The main question I try to answer in this epistemological part, which connects hypothesis formation in science with human reasoning in general, is whether the attitude of entertaining a hypothesis can be understood in terms of common epistemological doxastic concepts such as ‘belief’, ‘degrees of belief’ and ‘acceptance’. A better understanding of the rational attitude towards hypotheses is a first step towards the goal of providing rational and normative guidance on their formation (the second research challenge).

Part II: Logical Patterns In this part, which contributes to the first research challenge of explicating the various patterns of hypothesis formation, my main goal is to characterize some of the major patterns of hypothesis formation by means of formal logics, while assessing the extent to which a formal approach of these patterns is possible. At the same time, an effort is made to translate some of these insights to the context of artificial intelligence, and so contribute to the third research challenge of computational discovery.

than with the various ‘patterns of discovery’. My own use of the notions ‘patterns of discovery’ and ‘patterns of hypothesis formation’ is intended in the same sense as Schurz’s ‘patterns of abduction’ (2008a), though in a somewhat more generalized sense (see Section 1.4).

Part III: A Historical Case While the previous part focused mainly on the characteristics of some individual patterns, this historical part aims to provide some initial, largely descriptive, insights into why agents prefer one pattern over another in a particular context, by means of an in depth study of a case from the history of modern physics in which several physicists employed different patterns of hypothesis formation to solve a single anomaly. Detailed cases such as this one provide the groundwork and benchmarks for formulating normative guidance on the selection between the various patterns of discovery (the second research challenge).

Part IV: Thinking about Models In this final part, I will connect the research on hypothesis formation with the part of the literature that attempts to understand scientific discovery in terms of the construction and use of scientific models, by addressing the issue of how hypotheses and models are interwoven in scientific discovery. In this way, by reconnecting two diverging directions in the literature on scientific discovery, I am able to provide a richer perspective that benefits all three of the research challenges.

1.4 'Scientific Discovery' and 'Abductive Reasoning'

The overview of the literature has made it clear that the notions 'scientific discovery' and certainly 'abduction' have been given several interpretations in the literature. In this section, I want to specify how I will use these notions in this dissertation and how they connect to my main topic of hypothesis formation in science.

Hypothesis Formation in Science is a Subtopic of Scientific Discovery Having shrugged off the old connotations of a 'logic of discovery', scientific discovery is considered in the current literature (which is firmly based on actual scientific cases) as a step-by-step process that encompasses not only hypothesis formation but also hypothesis selection or the evaluation of their pursuit worthiness, selection of research questions, data gathering, scientific experimentation, model construction and refinement, etc.

Hence, the topic of this thesis, hypothesis formation in science, does not aim to capture the full process of scientific discovery but only the subprocess of hypothesis formation. Although it has been clearly shown that the distinction between the context of discovery and the context of justification is impossible to draw sharply (see Section 1.2), it is still common to distinguish between these two contexts in the literature, sometimes with the inclusion of a third intermediate *context of pursuit* (Laudan, 1977). Using this distinction, the topic of this dissertation would be a clear fit in the context of discovery, while occasionally addressing the context of pursuit. Yet the various case studies will make clear that even for hypothesis formation we cannot avoid having to deal with justificational aspects, even if we do not focus on the confirmation of hypotheses at all.

Given this qualification, it should be clear that my aim is to focus only on 'patterns of discovery' that are, in the first place, 'patterns of hypothesis formation'.

Hypothesis Formation encompasses Abductive Reasoning In the literature, hypothesis formation and abductive reasoning – defined by Peirce in its broadest sense as "the process of forming explanatory hypotheses" (1958, CP 5.172) – are often considered to be more or less the same. This can lead to ambiguity, because, as explained in Section 1.2, the notion of abduction has been given several interpretations in the literature and it is unclear how broadly this type of reasoning should be conceived. Yet various parts of this dissertation are contributions to the literature on abduction, so I cannot evade using the notion.

Therefore, let me start by specifying how I will use the notion 'abduction' (and the related notion 'abductive reasoning') to avoid any chance of equivocation.

First, with the notion 'abduction', I always mean the inference of generating or formulating explanatory hypotheses; I will, therefore, not consider 'abduction' as a more justificational inference, such as the *Inference to the Best Explanation* or the selection of the hypothesis most worthy of pursuit. I do assume, however, that some processes of explanatory hypothesis formation are rationally justified while others are not. Yet the result of a rationally justified hypothesis formation process is to be considered as a hypothesis: it involves no inference to its truth or to that it is the most worthy of pursuit.

Second, I will take the notion of 'abduction' as the formation of explanatory hypotheses in its broadest sense, i.e. as any instance of hypothesis formation for which the hypothesis is (part of) an explanation for the trigger or research question, without any further restriction concerning the inference schema according to which the hypothesis should be formed. Using this interpretation, I consider the often quoted Peircean schema of abduction of 1903:

> The surprising fact, C, is observed;
> But if A were true, C would be a matter of course,
> Hence, there is reason to suspect that A is true.
> (Peirce, 1958, CP 5.189)

in the first place as an inference of or an argument for the rationality of adopting A as a hypothesis in case it is explanatory, rather than a specific inference schema to generate the new hypothesis A. This interpretation would accord with what Peirce writes in the previous paragraph:

> Long before I first classed abduction as an inference it was recognized by logicians that the operation of adopting an explanatory hypothesis – which is just what abduction is – was subject to certain conditions. Namely, the hypothesis cannot be admitted, even as a hypothesis, unless it be supposed that it would account for the facts or some of them." (Peirce, 1958, CP 5.188)

It is, however, not my aim to provide an exegesis of Peirce's writings, especially as it is generally acknowledged that they contain multiple interpretations of abduction (see Section 1.2). I do agree that this scheme of abduction can also be useful if it is interpreted as an inference schema to form new hypotheses, an interpretation that is more in line with Peirce's earlier writings of 1878, in which he defined abduction[5] as the "inference of a *case* from a *rule* and a *result*" (Peirce, 1958, CP 2.623) and illustrated it with the following famous example:

> Suppose I enter a room and there find a number of bags, containing different kinds of beans. On the table there is a handful of white beans; and, after some searching, I find one of the bags contains white beans only. I at once infer as a probability, or as a fair guess, that this handful was taken out of that bag. This sort of inference is called *making an hypothesis*. [...]

[5] At the time, he actually called this inference "hypothesis", and later he has also called it "retroduction". Peirce insisted, however, that he always meant the same thing, even if, in retrospect, his interpretations really did evolve over time.

> *Rule* – All the beans from this bag are white.
> *Result* – These beans are white.
> ∴ *Case* – These beans are from this bag.
> (Peirce, 1958, CP 2.623)

If we would adopt this interpretation of abduction (as it is common in several logical and AI approaches), it is clear that abduction is a less broad concept that encompasses only instances of explanatory hypothesis formation in which the hypothesis or the elements of which it consists are already known by the agent.

As this thesis deals with hypothesis formation in science in general, it is more suitable for my purposes to consider the broadest of these two interpretations, i.e. to take abduction as any inference of explanatory hypothesis formation, including, therefore, instances of *creative abduction* in which a new idea, concept, or object is hypothesized.

Given these two conceptual choices concerning the meaning of the notion 'abduction', abductive reasoning (restricted to the context of science) can be considered as a subtopic of my topic: it deals with hypothesis formation processes that are explanatory in nature.

It could be argued that in a scientific context, any hypothesis raised in response to a research question or trigger can be considered as explanatory (at least in a broad sense). I leave it open whether this is the case, as this is a conceptual issue concerning how broadly the notion of 'explanatory' should be conceived. Yet I do consider my topic to be broader than what is dealt with in the literature on abduction. The main reason is that my starting point is hypothesis formation as it can be observed in actual scientific case studies. While studying such cases (see e.g. Part III), one encounters certain types of hypotheses that are never considered in the literature on abduction, such as suggesting that a particular law or generalization should be reconsidered or hypothesizing that two variables are somehow connected. Leaving it open whether such hypotheses should be considered explanatory or not, I also want to include patterns of hypothesis formation that lead to these kinds of hypotheses in my analysis, even if they are generally not included in lists of patterns of abduction in the literature. Hence, I conceive of 'patterns of hypothesis formation' as a notion encompassing the various 'patterns of abduction'.

Still, there is no doubt that abductive reasoning constitutes the major part of hypothesis formation in science. Many parts of this dissertation, especially the more logical chapters, will thus also be contributions to the literature on abduction, though always using the notion ‘abduction’ in the above delineated sense.

1.5 Some Common Themes

In this section, I discuss some common themes that connect the various parts and methodologies used in this dissertation.

Empirical Philosophy Although I do reason at times from *a priori* arguments, the starting point of my investigation is an empirical phenomenon, i.e. how hypotheses are formed in actual scientific practice. The constraints resulting from these empirical observations serve as benchmarks for my philosophical endeavors.

This kind of empirically flavored philosophy always presents an obvious tension, and this will be apparent in several places: the tension between description and normativity. Clearly, the main aim of philosophy has always been normative: as this thesis is meant to qualify as a dissertation in philosophy, it is my goal to draw normative conclusions concerning how hypotheses should be formed in science. On the other hand, by focusing in the first place on the empirical phenomenon of hypothesis formation in science, my analysis also has a strongly descriptive nature.

It would be presumptuous to claim that I have an answer or method that resolves this tension. The reader will find that some parts lean more towards description, others more towards prescription. The only guideline I could handle in charting the path of my reasoning is coherence between normative efforts and descriptive claims. I leave it to the reader to judge whether I have succeeded in this.

Philosophical Interdisciplinarity To assess this single empirical phenomenon, I have combined methods from various philosophical subdisciplines: epistemology, logic, history and philosophy of science. The advantage is clear. I am quite convinced that in this dissertation I can present a richer and broader view of hypothesis formation in science than I could have presented by focusing on a single method only. On the other hand, it has a clear disadvantage too: if in general focusing on one method and topic already leaves ample of opportunity for further in-depth research, using various methods leaves a multitude of possibilities for further and deeper exploration: I could have characterized more patterns of hypothesis formation, I could have analyzed a second major case study to confirm the results of my first one, I could have related my research more thoroughly to other existing questions in the literature, etc.

Obviously, there is no straightforward answer to this trade-off; it depends on the topic, and it is even, to a certain extent, a matter of taste. Still, in response to those who favor a single approach, I can point out that for each philosophical discipline, I have contributed to the existing literature in the form of stand-alone articles that are published or currently under review. And further, I have to content myself with the idea that time is finite, and so are dissertations.

Actual Case Studies from Physics As my starting point is actual hypothesis formation in science, this thesis contains a substantial number of actual case studies. Because a sufficient level of acquaintance with a field is required to understand the finesses of scientists' individual reasoning, these are all taken from my personal area of expertise, i.e. theoretical physics and astronomy.[6] Therefore, my conclusions are in principle tailored only to hypothesis formation in the physical sciences. I leave it to the reader to judge how far my conclusions reach in other fields.

Three Perspectives on Hypothesis Formation The process of hypothesis formation in science can be (and has been) studied from a vast number of perspectives. These many perspectives can in principle be assembled into three broad categories according to their main point of focus: the scientific community, the individual researcher, or the individual reasoning steps. Each of these three (broad) perspectives highlights different aspects of the hypothesis formation process.

Using the analogy of a zooming camera, we can first fully zoom out and study the *macro structure* in which a hypothesis is formed: the field, paradigm and research group the scientist is working in. This level determines both the problems or research questions and the general framework to solve these. Next we can zoom in to the level of the individual researcher. At this level we can distinguish personal preferences, general experience and metaphysical assumptions that steer the hypothesis formation process in certain directions. Finally, we can fully zoom in to the *micro*

[6] I have made an effort to keep everything understandable for the average philosopher with an interest in science.

structure of individual reasoning steps. At this level we can observe the actual steps of hypothesis formation; yet without looking at the sense of direction provided by the higher levels we would be unable to understand how these steps accumulate.

It is clear that every perspective requires its own methods of investigation. While logics and other formal approaches are, for instance, well-suited to study individual reasoning steps in a generalized way, history and philosophy of science are more suited to understand how the various perspectives relate in a single case. Throughout the various parts of this dissertation, I attempt, by using various methodologies, to address more than one of these perspectives, though the main focus is on the perspectives of individual reasoning steps and individual researchers.

Qualitative Approach All of the approaches in this thesis are qualitative. While the use of quantitative techniques such as Bayesian reasoning cannot be underestimated for hypothesis evaluation, their use is generally not crucial to understand the process of hypothesis generation.

1.6 A Note on the Structure of this Dissertation

This thesis is based on a collection of eight research articles, five of which are published or accepted for publication, two currently under review (December 2013), and one still in preparation. Each of these articles, of which one has a co-author, constitutes the main part of one of the following eight chapters.

Yet, although these eight articles are meant to be stand-alone papers, I have made a substantial effort to convert them into a single coherent volume. To this end, I have added this general introduction, a specific introduction for each main part, a general conclusion, and some additional sections to keep the papers up to date. Further, some sections and smaller passages are revised for general consistency of the thesis, and a uniform style sheet has been applied. Yet, given the stand-alone character of the various chapters, the reader might encounter at times a small amount of overlap.

Part I

Epistemological Considerations

motivation

In this first part, I will take some distance from hypotheses as they are understood specifically in the context of science in order to look in general at what it actually means for an agent to entertain a hypothesis, a query that brings me to the philosophical discipline of epistemology.

My main goal is to understand how the attitude towards a hypothesis, i.e. *entertaining a hypothesis* can be analyzed in epistemological terms. More particularly, I want to specify how this attitude relates to other propositional attitudes such as belief, having a degree of confidence and acceptance.

The core assumption at the heart of this part of the dissertation is that entertaining a hypothesis is a propositional attitude, more specifically a doxastic attitude, which could be understood from a broader perspective than that of scientific reasoning alone.

This assumption should be rather easy to digest. It is clear that once an agent has formulated a new hypothesis, she has a certain attitude towards it: she thinks that this proposition might be true, or at least relevant for her further actions. In other words, she thinks that this idea can possibly lead her to certain true beliefs. As such, it is clear that there is no *a priori* reason to state that this attitude could only be held towards propositions in a scientific context. Also in daily life, people are confronted with questions for which they have no direct answers, yet for which they might suggest possible hypotheses that lead them eventually to such answers.

If we divide the class of propositional attitudes (as is common in the literature) into a class of belief-like attitudes and a class of desire-like attitudes, it is clear that entertaining a hypothesis is a kind of belief-like attitude, also called a *doxastic attitude* (see Sections 2.1 and 2.2 for a more precise definition). In formulating a hypothesis we intent to represent the world as it is; we do not, as in the case of desire-like attitudes, intent to adapt the world to our hypotheses.

strengths and weaknesses of the method

Strengths The main strength of studying hypotheses from an epistemological point of view is that it allows us to study an aspect of scientific methodology from a much broader perspective and in connection with other forms of human reasoning. This gives us certain insights into what scientists are actually doing and how they perceive hypotheses that would be hard to obtain by focusing only on scientific practice itself.

Also, the method of conceptual analysis and counterexample, which constitutes the backbone of analytic epistemology, allows us to analyze to a very high degree of detail the differences between the various doxastic attitudes, which gives us in turn a sufficiently refined picture to craft a normative stance concerning the rationality of adopting these attitudes.

Weaknesses However, the method of conceptual analysis is also often criticized for considering only very stylized and simple (toy) examples. The use of such abstract examples to purify the attitudes and boil them down to their essence gives at times a legitimate reason to question whether this analysis still applies to real life examples, certainly in such a complex environment as scientific practice. Mindful of this threat, I have tried, in crafting my account of the doxastic attitude towards hypotheses, to make it compatible with actual cases, such as the one studied in Part III.

A final drawback for my purposes is the fact that the field of epistemology has paid hardly any attention to the attitude towards hypotheses, as the field is centered around the core attitudes of belief and knowledge. Yet as it turns out that the attitude of entertaining a hypothesis is not reducible to an attitude of belief (see Section 3.2), I have had first to deconstruct the attitude of belief in order to carve out an epistemological niche for the attitude towards a hypothesis.

overview of my contributions

In chapter 2, I construct a new conceptual framework to classify various doxastic concepts. This framework, which distinguishes between an agent's *theoretical doxastic attitude*, i.e. her credence in a proposition, and her *practical doxastic attitude*, i.e. her policy on trusting that proposition, can then be used to assess the doxastic concept 'belief'. In the literature on doxastic attitudes, the notion 'belief' is used in two different ways: on the one hand as a *coarse-grained notion* to indicate any attitude of assent towards a proposition; on the other, as a *fine-grained notion* that tries to capture the folk notion of belief. I show how this folk notion of belief is actually ambiguous, and how it can be deconstructed in terms of the doxastic reference framework I sketched.

In chapter 3, my attention shifts to the attitude of *entertaining a hypothesis* or *hypothesizing*. First, I show that if the *triad* of doxastic attitudes 'belief - disbelief - withholding judgment' should be regarded as capable of classifying in a coarse-grained manner every uncertainty present in human reasoning, it needs to be supplemented with a fourth attitude, one of hypothesizing, which cannot be reduced to the other three attitudes. Next, I use the conceptual framework developed in the previous chapter to define such an attitude in a more fine-grained manner.

2 Two Distinct Doxastic Attitudes

> *"Arkady Nikolaitch was telling us just now that you accept no authorities? Don't you believe in them?"*
>
> *"Why should I accept them? And what am I to believe in? If they tell me the truth, I agree, that's all."*
>
> — Turgenev, *Fathers and Sons*, 1862

This chapter is based on the paper "On Theoretical and Practical Doxastic Attitudes", which is currently under review (Gauderis, 2014a). I am indebted to Jan Willem Wieland, Bert Leuridan, Tim De Mey and several anonymous referees for their helpful comments on earlier drafts.

In this paper, two different doxastic attitudes are distinguished, i.e. the theoretical and the practical doxastic attitude, which roughly refer to an agent's credence in a proposition and her policy on trusting it in her practical reasoning. This framework is then used to dispel some ambiguities between various uses of the doxastic concept 'belief' and to clarify its distinction with other doxastic concepts such as 'degree of belief' and 'acceptance'.

The original content of the paper is retained, except for some stylistic adaptations.

2.1 The Notion 'Doxastic Attitude'

The notion of a *doxastic attitude* entered the general epistemology literature in the late 1970s, especially via the works of Goldman (e.g. 1978a; 1978b; 1979), who used it to describe in a generic way the propositional attitude of either *belief* or *disbelief*. Since the 1980s, the notion has become more widely used for this purpose, though one generally now adds a third option of *withholding belief* or *suspending judgment* (e.g. Feldman and

Conee, 1985; Steup, 1988; Chisholm, 1989; Sosa, 1991; Feldman, 2003; Steup, 2008). In this way, doxastic attitudes have come to be understood as the three possible attitudes an agent can intellectually adopt towards a proposition after considering it. This view has also been called *Triad* (Turri, 2012, p. 355).

As the notion 'doxastic attitude' gained currency, several authors started to also use it to describe a broader class of belief-like attitudes similar but not identical to the attitude of belief. From his Bayesian stance, Kaplan started this evolution by calling *degrees of confidence* – also often referred to as *degrees of belief* – doxastic attitudes (1981, p. 310). The attitude of *acceptance*, which was introduced in the literature by Van Fraassen (1980, p. 4), also generally came to be regarded as a doxastic attitude (e.g. Weintraub, 1990, p. 165). Kapitan (1986, p. 235) called the attitudes of *presuming*, *feeling* and *taking for granted* lower-level doxastic attitudes; unlike 'belief', these notions do not imply the agent's ability to articulate their content explicitly. Williams (1989, p. 124) even extended the idea further by calling *hypothesizing* and *suspecting* doxastic attitudes.

Already in 1983, Searle argued for the need to consider these belief-like attitudes, for some purposes, as a single category, and grouped them under the label BEL, in contrast with desire-like attitudes, which he called DES (Searle, 1983, pp. 29-36). Williams (1989, p. 124) made the same distinction, but named his groups 'doxastic attitudes' and 'orectic attitudes'. Leaving aside the question of whether all propositional attitudes can be reduced to (a combination of) elements of these two groups, it is commonly accepted in contemporary epistemology that 'belief' and 'desire' are two basic exemplars, each of them representative of and (for many purposes) interchangeable with a large group of similar propositional attitudes (Oppy, 1998). It is also common practice to call the group of belief-like attitudes 'doxastic attitudes' (Engel, 2012; Goldman, 2010, pp. 2, 26).

As we can observe, the notion 'belief' has been used in two different ways in the literature on doxastic attitudes. On the one hand, 'belief' is used as a coarse-grained technical concept designating any doxastic attitude that has an affirmative stance towards its content. This is the case, for instance, in the Triad position, mentioned above, according to which, as Turri (2012, p. 361) explains it, an agent chooses to take an attitude of assent, dissent or neutrality towards a given proposition. If one chooses an attitude of assent, it is called 'belief', irrespective of the intensity, degree, purpose or circumstances of this assent. For many analytical purposes, this abstraction from situational details can safely be made.

On the other hand, in the exploration of the various doxastic attitudes or belief-like attitudes, 'belief' is also employed as a fine-grained concept designating a specific doxastic attitude intuitively assumed to be more or less equivalent to a folk psychological notion of belief. This is clearly not the same use of 'belief' as in its coarse-grained meaning, as this fine-grained meaning is used to explain the other doxastic attitudes and contrast them with 'belief' precisely in terms of differences in intensity, degree, purpose or circumstances. Furthermore, as a general taxonomy of doxastic attitudes is lacking,[1] the other belief-like attitudes are often defined in terms of or with respect to such a specific fine-grained notion of belief, which is then regarded as a primitive and the most central doxastic attitude.[2]

While this double meaning of 'belief' should not itself, if properly conceived, pose a genuine problem, a tendency to conflate these two distinct uses in the literature has obscured the fact that the fine-grained notion of 'belief' is, unlike the rather precise and technical coarse-grained notion, utterly ambiguous and its specific distinctiveness in relation to other fine-grained doxastic attitudes is far from clear. As I will show, the example uses of the notion 'belief' in, for instance, the literature on 'acceptance' and in the literature on 'degrees of belief', seem to point to two different fine-grained notions.

I address these problems by proposing a taxonomy for specific doxastic attitudes that is not dependent on any specific fine-grained notion of 'belief'. I base this taxonomy on the idea that each agent actually has two quite distinct doxastic attitudes towards a given proposition, a *theoretical* and a *practical* one, corresponding respectively to her credence in the proposition and her policy on accepting it. This framework, in which the primitive doxastic concepts are 'degrees of belief' and 'acceptance', enables me to analyze other specific fine-grained doxastic concepts in terms of these two, including the intended meaning of a fine-grained notion of 'belief', i.e. a meaning that tries to capture the folk notion of belief. It will turn out that the folk notion of belief is a complex notion that specifies to a certain degree both an agent's theoretical and her practical doxastic attitude towards that proposition. The observed ambiguity in the use of a fine-grained notion of belief can therefore be attributed to the tendency

[1] Although a first attempt, from a somewhat different angle, can be found in Engel (2012).

[2] An exception to this is the literature on 'degrees of belief', which often takes the latter as the central notion, and defines the notion 'belief' in terms of it (see Section 2.6).

of different authors to stress one or the other part of this dual meaning of 'belief'.

After defining and explaining this doxastic framework in Sections 2.2 and 2.3, and using it to structure the various doxastic concepts in Section 2.4, I will use this framework in the final sections to re-assess two important debates in the literature on doxastic attitudes: namely the distinction between 'belief' and 'acceptance' (Section 2.5) and the distinction between '(plain) belief' and 'degrees of belief' (Section 2.6).

This elaboration will allow me to defend my reductionist stance to keep the notion of 'belief' philosophically only in its coarse-grained technical sense (as exemplified in the Triad view), while reducing it to an appropriate expression in terms of 'degrees of belief' and 'acceptance' in cases that require analysis of a particular and more specific notion of belief.

2.2 Doxastic Attitudes and Doxastic Concepts

I will start by addressing a minor conceptual issue to prevent confusion later on. In the literature, the notion 'doxastic attitude' is actually used in two senses. On the one hand, one can speak of the doxastic attitude of an agent towards p: although it gives us no further information about the nature of this attitude, because it is generic, it refers to the agent's attitude itself. On the other hand, one can speak of, for example, 'assuming' as a doxastic attitude. In this case, it refers to the type of an agent's doxastic attitude. I will avoid this confusion by using the notion *'doxastic concept'* for the different types, and, henceforth, *'doxastic attitude'* only for the generic attitude itself. In these terms, we can say, for example, that the nature of an agent's doxastic attitude towards p can be specified by choosing an appropriate doxastic concept such as 'accepting', 'assuming', 'being certain', etc.[3] Moreover, I will restrict my use of the term *'concept'* to this technical sense and use the term *'notion'* for general purposes.

To evade reference to the notion of belief, let me define doxastic attitudes in terms of the notion of *direction of fit*. This notion, first applied in the context of propositional attitudes by Searle (1983, p. 7), is a commonly

[3] It has been suggested to me that the type-token distinction could be used to capture this difference, but I am afraid that this might cause confusion here: on the one hand, a 'doxastic concept' is a specific interpretation of a generic 'doxastic attitude' (hinting that 'doxastic concepts' are tokens of the type 'doxastic attitude'); on the other hand, 'doxastic concepts' are still abstract types of attitudes, while the generic notion 'doxastic attitude' is often used to refer to the (unspecified) token attitude of a particular agent.

acknowledged way to distinguish doxastic attitudes from other propositional attitudes, because the direction of fit is regarded as the main difference between ‘belief’ and ‘desire’, the two basic (coarse-grained) exemplars of propositional attitudes (Williams, 1989, p. 124; Oppy, 1998).

In adopting a propositional attitude with a *mind-to-world direction of fit* (for instance, an attitude of belief), an agent aims to match the content of her attitude to the external world. In case of a mismatch, it is the content of the attitude that should be adapted. Accordingly, these attitudes can be judged to be true or false. In adopting a propositional attitude with a *world-to-mind direction of fit* (for instance, an attitude of desire), the agent aspires to match the world to the content of her attitude. In case of a mismatch, this cannot be remedied by changing the content of the attitude; it is, in a sense, the world that should be different. Accordingly, these attitudes can only be judged to be fulfilled or unfulfilled.

I define *doxastic attitudes* (and, hence, *doxastic concepts*) to be propositional attitudes (or concepts) that satisfy the following criteria:

(a) they have a mind-to-world direction of fit;
(b) they have no world-to-mind direction of fit;
(c) they are defined only in terms of criteria that are internal with respect to the agent holding the attitude.

I have added conditions (2) and (3) to the colloquial definition of a doxastic attitude in terms of ‘direction of fit’ in order to exclude both propositional attitudes with a double direction of fit (e.g. ‘fearing that *p*’, which involves both thinking that *p* is credible (mind-to-world) and wanting that $\neg p$ is the case (world-to-mind)) as well as attitudes that depend somehow on external criteria such as ‘knowing that *p*’ (for which it is commonly accepted that this implies, at least, that *p* is true; a criterion that is independent of the agent).

2.3 The Theoretical and the Practical Doxastic Attitude

By considering the various doxastic concepts, one can observe that in fact they specify two different doxastic attitudes. This has already been noted by scholars working on the notion of acceptance (e.g. Engel, 2012, pp. 20-21). Given a proposition p and an agent S, I define these two attitudes as follows:

(TDA) the *theoretical doxastic attitude*: the credence S gives to p or the confidence S has in the truth of p.

The nature of an agent's theoretical doxastic attitude towards p can be found out by asking her: "How likely is it, do you think, that p is true?" Her response can vary from the expression of a gut feeling to a fully reasoned answer. In any case, the agent's attitude will be the result of an assessment of the truth of p, based on what she regards as relevant evidence for it, and its expression can range gradually from an absolute disbelief in p to a total conviction concerning p's truth.

(PDA) the *practical doxastic attitude*: the policy S has on trusting p and relying on its content.

The nature of an agent's practical doxastic attitude can be found out by asking her: 'In which type of circumstances would you let your reasoning and actions depend on this proposition, and in which not?' Her response can vary from a vague reference to some archetypical contexts to a precise demarcation criterion in terms of a specific property of the circumstances. Accordingly, S's attitude will be the result of an assessment by her of the practical consequences of relying on the truth of p, and can range from a willingness to assume p only in hypothetical arguments to accepting p under any circumstances.

In the event that the context or circumstances are given, let us call them C, the practical doxastic attitude reduces to the following derivative attitude:

(PDAC) the *practical doxastic attitude in a context*: the policy S has on trusting p in the particular context C, i.e. whether or not she relies on p in the context C.

This time, an agent's attitude will be the result of a yes-or-no decision as to whether she is willing to let her reasoning and actions depend on p in some given particular situation. As such, the premises for practical reasoning are constituted by the agent's practical doxastic attitudes in the context at hand.

Let me add five further clarifications. Where confusion might arise concerning which variant of the doxastic attitudes is intended, I will add the relevant acronym, namely TDA, PDA or PDAC.

First, it is clear that given any proposition and any agent, one can construct an answer to both of the questions stated in the explanations of (TDA) and (PDA) above. Although these answers may be expressed at different levels of detail, it is possible to speak both of an agent's theoretical and of her practical doxastic attitude towards a particular proposition. These descriptions are clearly not the same thing: the judgment of a proposition's truth (TDA) can be a very balanced report, which is quite independent of the circumstances one finds oneself in at that moment. On the other hand, whether one lets one's reasoning depend on that proposition in a particular context (PDAC), is a yes-or-no decision which may well turn out differently in different types of circumstances or for different types of possible actions. As such, a very subtle policy (PDA) can be generated.

Second, the demarcation between contexts in which the agent relies on a proposition and those in which she does not (PDA) is determined at least by the positive consequences the agent foresees in case she is right and the negative ones she is willing to accept in case she is mistaken. These consequences, which are considered only from the agent's perspective (in other words, irrespective of the actual consequences), can vary a great deal and are often hard to compare. In accordance with Bayesian decision theory, the weighted sum of the relevant consequences can be called the *expected utility* for the agent of relying on a certain proposition in a certain context. But as it is not needed for our purposes that agents actually make such calculations, it suffices to assume that agents can compare the consequences they foresee qualitatively.

Third, the attitudes are defined descriptively without reference to rational behavior or to any normative theory. For rational agents, theoretical and practical doxastic attitudes are of course related: propositions of which one is fairly confident that they are true will be relied on in a wide variety of circumstances, while propositions that one suspects of being false will be relied on only in contexts in which the penalty of being mistaken is rather low.

In fact, Bayesian decision theory provides a method for calculating the most rational practical doxastic attitude in a certain context (PDAC) given an agent's degrees of belief towards the relevant propositions (a quantitative description of her TDA) and (quantified) expected utilities of relying on those propositions in that context. However, agents are clearly not always able to perform these quantifications and calculations effectively. This explains why in everyday circumstances, even if an agent intends to be rational, her theoretical and practical attitudes will sometimes appear to be at odds. Also, even rational agents differ in their perceptions of the utilities: two agents having the same degree of confidence in a proposition might rely on it differently in similar circumstances. This explains why the various folk notions describing doxastic attitudes allow for independent descriptions of an agent's theoretical and practical doxastic attitude towards a certain proposition (see Section 2.4).

Fourth, the theoretical doxastic attitude resembles what classical epistemologists typically have in mind when talking about doxastic attitudes (as it reports the agent's perception of the truth of a proposition). To them, the practical doxastic attitude may seem an awkward addition. Yet it is a genuine doxastic attitude. For recall the three requirements stipulated in the definition of doxastic attitudes. First, the theoretical doxastic attitude clearly has a mind-to-world direction of fit: an agent adopts a policy to trust *p* depending on how she perceives the world and what might happen in it, and therefore her policy reflects her perception of the world.[4] Secondly, there is no world-to-mind direction of fit with respect to *p*: in purely specifying the circumstances in which she trusts *p*, an agent does not express any desire that the world should confirm to the content of *p*. Thirdly, the attitude is defined solely in term of the agent's internal perception of the circumstances, the consequences she herself foresees and her assessment of the trustworthiness of the proposition, all of which are criteria internal to her.

Fifth, it is common to define the philosophical notion of 'degrees of belief' technically in terms of dispositions to bet, which would reduce the theoretical doxastic attitude (TDA) to a mere variation of the practical doxastic attitude (PDA). Such an operationalist view, which has proven to be an excellent starting point for rational decision theory, is, however, not a problem for the framework I am proposing here. My goal is to distinguish two qualitatively distinct human modes of assessing a proposition, resulting in two doxastic attitudes, which can be independently described in a qualitative way, a distinction that is reflected in the various doxastic

[4] To clarify this point, consider the following example: an agent *S* decides to accept the proposition *p*, having no specified theoretical attitude towards it, for a certain *research context* (a context in which the consequences of being mistaken are negligible; see Section 3.3 for a more precise definition). Suppose that during this research, *S* gathers evidence that *p* is very unlikely. Apart from specifying *S*'s theoretical doxastic attitude towards *p*, this evidence will also lead *S* to adapt her practical doxastic attitude: *S* will now accept *p* in hardly any context (where before she was willing to trust it for research contexts). In other words, the agent aims to match her policy (her attitude) to the perceived external world.

folk notions (see Section 2.4). I accept that, for the theoretical attitude, it may be possible that humans can only make qualitative comparisons, and that, if the attitude needs to be operationalized quantitatively (for use in a normative theory of decision making), this can probably be done only by equating theoretical attitudes (TDA) with practical doxastic attitudes for certain artificial and purified contexts (PDAC) such as “no strings attached” bets.[5] Yet though it can be argued that the quantitative operationalization of the notion ‘degrees of belief’ is, in a technical sense, an (artificial) practical doxastic attitude, the notion can still be used qualitatively as a primitive doxastic concept to describe the theoretical doxastic attitude, as this operationalization is not required for describing various folk notions of doxastic attitudes.

In summary, then, and taking the agent’s evidence to be fixed at a certain moment, the theoretical doxastic attitude (TDA) is a context-insensitive doxastic attitude that allows for a range of degrees of confidence in the truth of p, while the practical doxastic attitude (PDA) is a context-sensitive attitude that reduces to a yes-or-no decision in each context (PDAC) depending on the expected utility of the two options in that context. For rational agents, these two attitudes towards a certain proposition are related, but the nature of this relation depends on how each particular agent balances her theoretical appraisal with expected utility.

2.4 Three Categories of Doxastic Concepts

The many known doxastic concepts, such as ‘doubting’, ‘accepting’, ‘assuming’, ‘having some confidence’, ‘suspending judgment’, ‘hypothesizing’, ‘being certain of’, ‘suspecting’ and ‘believing’ (in its specific and intuitive folk psychological meaning) may all be regarded as (partial) descriptions of the nature of either one or both of the two doxastic attitudes I have distinguished.

Of these doxastic concepts, some, such as ‘having a particular degree of confidence in (the truth of) p’, ‘giving p some credit’ or ‘being (un)certain of p’ give a clear description of the nature of the theoretical doxastic attitude of the agent towards the proposition. They specify up to a certain level of detail how the agent judges the truth of p, but give hardly any in-

[5] In real life, winning or losing a bet has not only monetary consequences, but also psychological and social ones in the form of joy, sadness, self confidence boosts or dips, gain and/or loss of social prestige, etc. Therefore, it is hard to call the bet contexts used to define ‘degrees of belief’ actual real-life contexts.

formation about when the agent intends to let her reasoning depend on p. For instance, suppose that an agent acknowledges that her chances of recovering from a disease are fifty-fifty (TDA). In other words, her degree of confidence in the truth of either possibility, of recovering or not, is equally large. This information tells us nothing about her practical doxastic attitudes (PDA). An optimistic person might base all her practical reasoning and actions on the premise that she will recover, while a pessimist might do the opposite. As concepts of this type describe only the theoretical doxastic attitude of an agent towards p, they can be called, in short, *theoretical doxastic concepts*. Of these, 'having *a particular degree* of confidence in the truth of p' can be regarded as the basic or primitive notion, because it allows for a description of any theoretical doxastic attitude by specifying 'a particular degree' qualitatively. For example, 'being certain' means having full confidence, while 'giving some credit' means that one takes 'a particular degree' to mean a substantial amount, but generally less than the amount of confidence in the other option.

Other doxastic concepts, such as 'accepting that p is the case', 'suspending judgment as to whether p is the case', 'taking p to be a relevant possibility' are examples of *practical doxastic concepts*. They indicate the type of circumstances or contexts in which the agent will let her reasoning depend on p (or not) (PDA), while giving hardly any further information about exactly how much confidence the agent has in the truth of p (TDA). For instance, in most circumstances, people accept that in general their partner will not lie to them (PDA), but if asked how certain they are about this (TDA), some would answer that they have some doubts whether this is really the case, while others would be fully confident. Similarly, if an agent suspends judgment as to whether p is the case, and thus does not rely on p in any context (PDA), one does not know whether, theoretically, p or $\neg p$ seems more plausible to her (TDA). Of the practical doxastic concepts, 'accepting' can be considered the primitive notion, because it allows for a description of any practical doxastic attitude (PDA) by specifying in which contexts the agent accepts the proposition (PDAC).

Finally, some doxastic concepts, such as 'believing that p', 'doubting whether p', 'being ignorant about p' have both a theoretical and a practical meaning, or, in other words, describe to some degree the nature of both the agent's theoretical and practical doxastic attitude towards p. For instance, when an agent believes p (in its intuitive folk meaning), we certainly know that she has a high degree of confidence in the truth of p (TDA), but we also know (because people state their beliefs when prompted to give reasons for their actions) that she will be willing to base her practical reasoning on p as a premise in a large range of circumstances (PDA). The ambiguity of this notion arises from the fact that one can emphasize one part or the other, the theoretical or the practical, as we will see in the following sections.

The remainder of this chapter will examine how to understand this dual nature, both theoretical and practical, of the folk notion of 'belief'. This will be done by applying the conceptual framework presented thus far in order to reassess two important debates in epistemology: namely concerning the difference between 'belief' and 'acceptance' and the difference between '(plain) beliefs' and 'degrees of belief'.

The main goal of this analysis will be to show that 'belief' cannot be retained as a specific fine-grained primitive doxastic concept (apart from its technical coarse-grained meaning). If one tries to capture the intuitive sense of the folk notion of belief, one obtains a complex and, hence, secondary notion, reducible to a suitable expression of 'degrees of confidence' and 'acceptances'. I will argue that these two concepts are far better suited than 'belief' to be considered as primitive doxastic concepts, because each of them specifies only one of the two doxastic attitudes. Still, precisely because of this dual nature of the folk notion of belief, the notion of 'belief' can be retained in its coarse-grained philosophical sense, as denoting any doxastic attitude (either practically or theoretically) that assents to its content, as long as one takes care to specify the attitude more precisely in detailed philosophical analysis.

2.5 Belief and Acceptance

The notion 'acceptance' was introduced by Van Fraassen (1980, p. 4) to describe the attitude of scientists towards their most empirically adequate theories. According to him, acceptance of a theory does not necessarily entail that one believes it (1980, p. 9, 46), yet at the same time encompasses more than belief, because the attitude of acceptance has the pragmatic dimension of commitment to a theory, which is a question not of truth but of usefulness (1980, p. 88).

Given the importance of the notion of acceptance in general and its difference from belief, it soon became a research topic for epistemology. The most influential epistemological account to date has been given by Cohen, who defines *acceptance* of p as having or adopting

> [...] a policy of deeming, positing or postulating that p – i.e. of including that proposition or rule among one's premises for deciding what to do or think in a particular context, whether or not one feels it to be true that p. (Cohen, 1992, p. 4)

Cohen further states that acceptance, unlike belief, is more or less under an agent's voluntary control (1992, p. 20) and acknowledges implicitly that acceptance is a context-dependent notion (1992, p. 14). These two characteristics are also stressed by other authors such as Bratman (1992, pp. 5, 9) and Engel (1998, pp. 145-148). Engel further holds that, while truth is the criterion for evaluating beliefs, utility is the criterion for acceptances. A final explanation of the distinction between these two notions is given by Lehrer (2000, p. 209), who approaches the topic from a somewhat different point of view. According to him, belief is a first-order doxastic state, while acceptance is a second-order "metamental" state based on a reflective evaluation of one's first-order beliefs. Yet I am tempted here to follow Engel (2012), who notes that Lehrer's account neglects the important pragmatic aspect of acceptance as well as the idea of trust, which is inherent in the notion. Therefore, I do regard acceptance as a first-order attitude having propositions as its content, not beliefs. Yet this does not prevent one from regarding beliefs, in the spirit of Lehrer's view, as constitutive in the formation of one's acceptances. If the acceptance towards a proposition is consciously formed (by e.g. applying a kind of decision theory), this decision will clearly have taken into account beliefs about this proposition and related ones, such as the foreseen consequences of particular actions.

Using the framework introduced in this chapter, it seems at first sight possible to describe the distinction between these two notions as the difference between a theoretical doxastic concept ('belief') and a practical one ('acceptance'). Of the four contrasting features between beliefs and acceptances that are pointed out in the literature, the context-sensitivity of acceptances (and practical doxastic attitudes in general) and utility as their evaluation criterion have already been discussed in previous sections. The other two contrasting features relate to the fact that an agent's practical doxastic attitude (PDAC) can be the result of a decision. Given that such a decision takes into account the agent's theoretical doxastic attitude (TDA), among other things such as an assessment of the circumstances, it can be understood why acceptances are more under an agent's voluntary control and influenced by her theoretical doxastic attitudes.

Notwithstanding the *prima facie* plausibility of this first analysis of the distinction between ‘belief’ and ‘acceptance’, Frankish argues convincingly that distinguishing these attitudes as such – in our framework, considering belief as a theoretical doxastic concept and acceptance as a practical one – is problematic, because it “suggests that acceptance is not a form of belief at all, but a wholly different attitude” (2004, p. 86). He agrees that there are acceptances that are not beliefs, but maintains that “it would be perverse to claim that none of them are” (2004, p. 87). In other words, people do believe some (if not most) of the states present in their conscious practical reasoning. Frankish’s concern is a genuine one. It may be pointed out in response that regarding beliefs and acceptances as distinct attitudes does not imply that an agent could not hold both of them towards a single proposition. But the fact that beliefs can serve as premises even if no form of decision theory or other form of conscious consideration is applied suggests that the adoption of a new belief must in itself directly imply the acceptance of this newly believed proposition for certain circumstances. In other words, acceptance for certain circumstances is part of the meaning of the attitude of believing, such that the folk notion of ‘belief’ cannot be a purely theoretical doxastic concept.

Frankish explains this problem by classifying plain beliefs as a subspecies of acceptances, i.e. those that are “epistemically motivated and unrestricted as to context” (2009, p. 86). His explanation, however, seems at least a little awkward, because context-dependency is an inherent feature of Cohen’s definition of acceptance, which Frankish himself embraces. Frankish’s idea of unrestrictedness as to context implies that a belief can serve as a premise for practical reasoning in any context. But if a believed proposition may be considered a true premise in any context, this is the same, it seems, as adopting a policy of trusting the belief in any context: for there is no longer any demarcation between contexts in which one can trust the belief and those in which one cannot. This view is hugely problematic. Kaplan (1996, p. 104), who calls it the *act view*, argues that it is fallacious – a fact of which Frankish (2009, p. 82) is aware – because agents are not certain of their beliefs. Hence, they will, for example, not bet on the truth of their beliefs if the stakes are too high, even if they are fully convinced. The act view would instruct them to always trust their belief and accept any bet.

The initial explanation of the distinction between ‘belief’ and ‘acceptance’, outlined above, can be modified as follows in order to cope with Frankish’s concern. ‘Acceptance’ is, as noted, a purely practical doxastic concept, but the folk notion of ‘belief’ actually has both a theoretical and a practical meaning. On the one hand, it means that an agent has *at least* a rather high degree of confidence in the truth of the proposition. Exactly how high need not be numerically expressible, but a decent amount that is clearly larger that the amount of confidence in the opposite proposition is always minimally implied. On the other hand, it also means that the agent is willing to base her practical reasoning on this proposition in *at least* all contexts where the negative consequences in case she is mistaken seem acceptable to her.[6] This includes contexts where she cannot or does not assess these consequences, but where she has no reason to think that much depends on whether she trusts this proposition or not.

Keeping this in mind, one can identify the well-known examples in the literature on 'acceptance', in which an agent does not act or reason on her beliefs, as contexts where these negative consequences are unacceptable for the agent. Consider the following example, often cited and originally developed by Cohen (1992, p. 25): an attorney accepts that her client is innocent in the context of a particular trial, even though her own belief is that he is guilty. She does not accept her own belief in the context of the trial because the negative consequences of acting on that belief are unacceptable in this context, not only for her personal career but also, and more importantly, for the social institution of the judicial system. In contexts where the negative consequences of accepting her own belief are not so prominent, for example when she talks about the case with her husband/wife, the attorney might express and reason upon her own belief.

In conclusion, the folk notion of 'belief' describes the nature of both the theoretical and the practical attitude of an agent towards a certain proposition, and should therefore be handled with care. This double meaning – on the one hand, having a sufficiently high degree of confidence in the proposition's truth (TDA), and on the other, being willing to rely on it in at least all contexts where the consequences of being mistaken seem to be acceptable (PDA) – also explains the diverging views one finds in the debate about 'belief' and 'acceptance', because it is possible to lay the emphasis more on the theoretical or on the practical aspect of belief. When Van Fraassen, Cohen and others try to identify the differences between acceptances and beliefs, they appeal to intuitions about the theoretical meaning of 'belief',

[6] Of course, holding a belief might entail that one accepts it in many more contexts. For instance, if I come to believe that there are no cars coming down the road by having a look in both directions, I will accept this proposition in the present context in which I have to decide whether I will cross the road, even though the consequences of being mistaken – being hit by a car – are not at all acceptable to me.

a meaning which ‘acceptance’ lacks. But when Frankish rightly points out that some acceptances actually are beliefs, he appeals to existing intuitions about the practical meaning of ‘belief’.

2.6 Belief, Degrees of Belief and the Bayesian Challenge

The conceptual framework of this chapter and the double meaning of the folk notion of belief can also help explain the distinction between the concepts *‘(plain) belief’* and *‘partial belief’*, the attitude of having a particular *degree of confidence* or *degree of belief* in a proposition, as well as the requirement put on any explication of this distinction by the Bayesian Challenge. This is the name given by Kaplan (1996, pp. 89-101) to a problem that has been formulated in various ways by different authors; see for example Jeffrey (1970, pp. 158-161) for an early formulation and Frankish (2009, p. 76) for a fairly recent one. Let us consider Frankish’s formulation here. As he writes:

> Bayesian decision theory teaches us that the rational way to make decisions is to assign degrees of probability and desirability to the various possible outcomes of candidate actions and then choose the one that offers the best trade-off of desirability and likely success. [...] How can flat-out belief and desire have the psychological importance they seem to have, given their apparent irrelevance to rational action? (Frankish, 2009, p. 76)

It is my own view, and the view of the authors who have formulated the Bayesian Challenge, that any account of the relation between plain belief and degrees of belief must also give a satisfying answer to this challenge. Generally speaking, three strategies to specify the relation between ‘plain belief’ and ‘partial belief’ are discernible in the literature.

A first strategy, and the one that has been most extensively explored, is what Foley has called the Lockean Thesis:

> To say that we believe a proposition is just to say that we are sufficiently confident of its truth for our attitude to be one of belief. (Foley, 1992, p. 111)

Yet this strategy, which, in our framework, identifies 'believing that *p*' as a theoretical doxastic concept and defines it in terms of a threshold for the degree of belief in *p*, faces two severe threats.

First, this strategy has to cope with the famous lottery and preface paradoxes (Kyburg, 1961, p. 197; Makinson, 1965), which show that the Lockean thesis can yield inconsistent beliefs when combined with the aggregation principle for beliefs (which states that the conjunction of two beliefs is also a belief). These paradoxes are typically met by softening or qualifying the aggregation principle,[7] but this is generally done by introducing some context-sensitivity, which is hard to bring into accordance with the idea that degrees of belief (to which beliefs can, according to the Lockean thesis, be reduced) are, like all theoretical doxastic concepts, context-independently defined.[8]

Second, this strategy also fails to meet the Bayesian Challenge, given that this challenge to explain the psychological importance of plain beliefs appeals particularly to intuitions of 'belief' as a practical doxastic concept. Theoretically, there may be a very minimal difference between an acquired belief and a proposition that falls just short of the threshold for belief, as degrees of belief are considered to be on a continuous scale. The Bayesian view perfectly explains how even a small difference in this regard can lead to widely divergent decisions based on this belief. It cannot explain, however, why agents, once they have acquired a belief, tend to take it into account in the most diverse situations, even situations to which the acquired belief is only marginally significant. This behavior can only be understood if we assume that an agent does not run a full Bayesian analysis for any decision but simply adopts a policy to start relying on a belief in a large set of contexts once she has acquired it.

A second common strategy to specify the relation between '(plain) belief' and 'partial belief' is to regard 'plain belief' as a kind of behavioral disposition arising from an agent's partial beliefs, e.g. a disposition to assert the belief as a proposition (Kaplan, 1996, p. 109) or to accept it (Frankish, 2009, p. 86). These strategies identify belief solely as a practical doxastic concept. However, while this identification may meet the Bayesian Chal-

[7] See Douven (2002, 2008) for some alternative approaches.

[8] It might be argued that if degrees of belief are defined in terms of betting behavior, they are in fact context-dependent. But the artificial context of a "no strings attached" bet, which is created to operationalize the idea of degrees of belief and has no real occurence, should not be confused with the context in which a real agent is situated and in which she needs to take a decision. Her degree of belief in a proposition is independent of this actual context.

lenge, it fails to accord with our common (theoretical) intuitions about the context-insensitivity of beliefs. As long as there are no changes in the evidence an agent perceives, she will likely suppose that her beliefs hold in any context she may find herself in, while a characterization of 'belief' as a practical doxastic concept requires – to avoid the pitfall of the aforementioned act view – that one limits the set of circumstances in which the belief holds.

Finally, some authors, such as Bratman (1992) and Jeffrey (1970), seem implicitly to deny the existence of plain beliefs and reduce them in every case either to a degree of belief or to an acceptance in certain contexts.

To implement this third strategy explicitly seems to me the best proposal. The intuitive folk notion of belief entails both, theoretically, that the agent has a sufficiently high context-insensitive degree of confidence in the truth of the proposition and, practically, that the agent has adopted a policy of relying on this proposition in at least all circumstances where the consequences of being mistaken seem acceptable to her.

This duality in the meaning of the notion can give rise to ambiguity, hence making it unfit for the philosophical analysis of doxastic concepts. Consider again some of the examples described above, which pop up in the literature. Take the attorney who believes that her client is actually guilty: does this mean that, although the attorney is quite confident about her client's guilt (TDA), she practically bases all her reasoning on his innocence (PDA)? Or does it mean that, except for her public appearances in court, she reasons on the basis of his guilt to determine her strategy (PDA)? Or consider another example, of a woman who believes that her husband/wife is not cheating on her. Does this just mean that she takes this to be the case without questioning it (PDA), although she has to admit that she cannot be fully certain (TDA)? Or does it mean that she is also wholeheartedly confident about it (TDA)? Clearly, the notion of 'belief' is not precise enough to describe the particular attitudes in these examples.

In light of these considerations, it seems clear that we can gain precision in our analyses of doxastic concepts by replacing any fine-grained specific concept of 'belief' with the more precise and primitive concepts of 'degrees of confidence' and 'acceptance', for using the latter concepts makes it possible to clarify whether the theoretical, the practical or both attitudes are meant. Still, when agents report on their doxastic attitudes towards p, the attitudes of 'having a high degree of confidence in p' and 'being willing to rely on p if the negative consequences seem acceptable' are often present together. Therefore, I see no problem in retaining the folk notion of 'belief' as a somewhat ambiguous but sufficiently clear shorthand to denote both attitudes in daily life. Also, precisely because of its rather broad meaning, 'belief' in a coarse-grained sense can be retained as a technical concept referring to any doxastic concept that expresses an attitude of assent to its content. In detailed philosophical analysis, however, much precision can be gained by eliminating altogether the idea that there exists a specific and unambiguous fine-grained doxastic notion of 'belief'.

2.7 Conclusion

In this chapter, I have shown that, in the literature on 'doxastic attitudes', the notion of 'belief' is used both in a coarse-grained sense to indicate any doxastic attitude that indicates assent towards a proposition, and in a more specific, fine-grained sense to be contrasted with other doxastic concepts such as 'acceptance' or 'having a specific degree of belief'. I have argued that, while the coarse-grained meaning of 'belief' is technically sound and useful for philosophical analysis, the fine-grained meaning, which draws on the intuitive folk notion of belief, is utterly ambiguous.

In order to dispel this ambiguity, I have presented a new framework for describing fine-grained doxastic attitudes which is not reliant on a specific and intuitively clear fine-grained concept of 'belief'. In this framework, I distinguish between an agent's theoretical doxastic attitude (her credence in p) and her practical doxastic attitude (her policy on trusting p to be used as a premise for her practical reasoning). Given this distinction, all well-known doxastic concepts can be placed into one of three categories: theoretical doxastic concepts (of which 'having a certain degree of confidence' is the primitive notion), practical doxastic concepts (of which 'acceptance' is the primitive notion) and doxastic concepts that describe both attitudes, such as the folk notion of 'belief'.

After introducing this framework, I have argued for a reductionist stance concerning the idea of an unambiguous and specific fine-grained notion of 'belief' and showed that much precision can be gained in philosophical analysis by using a suitable combination of 'degrees of belief' and 'acceptances' whenever the folk notion of 'belief' is intended.

The applications of this new framework need not, and should not, be restricted to the analysis of 'belief'. An interesting question for further research is whether this framework can provide us with insights into the specific nature of other important doxastic concepts, such as ‘suspending judgment’ and various forms of ignorance. In the following chapter, I will show how this framework can be fruitfully applied to explicate the doxastic attitude of ‘entertaining a hypothesis’. Furthermore, it needs to be investigated whether this reductionist stance on a specific fine-grained notion of belief might also give us more precision in other epistemological debates that rely heavily on the notion of belief, such as debates about rationality, justification and the theory of knowledge.

3 The Attitude of Entertaining a Hypothesis

I can live with doubt and uncertainty and not knowing. I think it's much more interesting to live not knowing than to have answers which might be wrong.

— Richard Feynman, BBC Horizon, 1981

This chapter is based on an unpublished paper. I am indebted to Jan Willem Wieland, Tim De Mey, Christoph Kelp, Bert Leuridan, Atocha Aliseda and Maarten Van Dyck for their helpful comments on earlier drafts, terminology and presentations.

In this paper, it is first argued that the classical epistemological triad 'belief - disbelief - withholding (judgment)' should be supplemented with the attitude of entertaining a hypothesis or hypothesizing, after which this last doxastic attitude is characterized by means of the conceptual framework developed in Chapter 2.

3.1 Uncertainties and Doxastic Attitudes

Psychological uncertainties, i.e. propositions of which an agent is not certain whether they are true or false, figure prominently in human reasoning.[1] They arise not only when information is obtained from dubious or unreliable sources. Human reasoning consists also in a large part of defeasible inferences, such as induction or abduction, and each of these add in general some uncertainty to the propositions present in one's reasoning.

Yet it is commonly accepted that uncertainty does not prevent us from rationally relying on such uncertain propositions as premises for our further

[1] Two different kinds of *(un)certainty* can be distinguished: a belief is *psychologically certain* if the agent is fully convinced of its truth (yet might be mistaken), whereas it is *epistemically certain* if it has the highest possible epistemic status, generally including a warrant for its truth (Reed, 2011). This chapter deals only with psychological certainties and uncertainties.

reasoning or for determining our course of action. For example, somebody who just finds out that her wallet is no longer in her pocket, will generally infer that it might have been stolen. She will then trust this conclusion sufficiently to base some of her actions on it, such as calling the issuer of her credit cards to block them, something she would never do if she was certain that her wallet was not stolen. But she is also not certain that it is actually stolen; it might not even be the most probable explanation: she could have lost it in her office where a colleague found it, or maybe her memory is failing her and she simply forgot it at home. Still, in this situation, it is rationally justified to perform particular actions on the uncertain premise that it has been stolen. In general, keeping particular uncertain propositions in mind and basing part of our actions on them is beneficial for us, even we have no firm belief in these propositions and even if they are not the most likely statements or explanations.

In epistemology, one's *doxastic attitude* towards a proposition,[2] can be described in a coarse-grained manner by identifying it as one of the following *triad* of doxastic concepts: *belief* (if the agent's attitude is one of assent, or, in other words, she takes the proposition to be true), *disbelief* (if she takes it to be false, which is generally considered to be the same as believing its opposite) or *withholding judgment* (if she takes neither the proposition nor its opposite to be true).[3] Although this triad of technical concepts is the basis of many qualitative formal approaches to modeling human beliefs and human reasoning (e.g. doxastic logics), these three concepts are limited and somehow insufficient to describe an agent's attitude towards the uncertainties in her reasoning.

Describing certainties in human reasoning presents no problem: if we are certain of a proposition, we will regard it as true. The certainties in an agent's reasoning can thus be considered as a subset of her beliefs. Uncertainties in human reasoning, on the other hand, are not so easy to describe with sufficient accuracy in terms of this triad of doxastic concepts. Certainly, some uncertainties can be counted as beliefs. For instance, an agent might believe that her children are at school on a regular school day, or

[2] This is roughly the nature of an agent's opinion about that proposition after being confronted with it. For a more precise definition of *doxastic attitudes*, see Section 2.2. For the distinction between coarse-grained and fine-grained descriptions of a doxastic attitude, see Section 2.1.

[3] For references to the literature on this *Triad* view, see Section 2.1. In describing doxastic attitudes in a coarse-grained manner, it is, as explained in Chapter 2, not necessary to make a distinction between an agent's theoretical and practical doxastic attitudes.

that her colleagues will, as they usually do, go for a drink on Friday. Yet, if pressed, the agent in these cases has to admit that she is not certain about these statements as they are the result of inductive reasoning. There are also uncertainties that can be accurately described as disbeliefs. For example, an agent might not believe that her 16-year old daughter never kissed a boy (even if she has no indication at all), or that her colleague never drinks alcohol (even if she has never seen her doing so).

But not all uncertainties in an agent's reasoning qualify for one of these two categories. The main reason is that uncertainties are often mutually exclusive: there might be several contradicting explanations for an observation, or there might be several scenarios possible. In such situations, agents often reason further and act on the basis of several of these options without making any commitment to the truth of any of these. They are, therefore, not expressing genuine beliefs or disbeliefs. Consider the following example:

> This morning, I found out (p_1) that the mailman has not come by. My initial thought was (p_2) that the book I ordered had not yet arrived and there was no other mail. But when I saw my neighbor a bit later, she told me (p_3) that the mailman didn't look too well yesterday. So, at that point I realized that it was also possible (p_4) that the mailman is sick today.

At first, it is certainly rational for the agent to believe p_2. But this attitude towards p_2 is dynamic: after getting the information p_3, she has to re-assess and adjust her attitude towards p_2, as p_4 has emerged as a possible option as well. Although the explanations p_2 and p_4 are not strictly mutually exclusive, the agent will generally not believe both of them at the same time (as either one of them alone suffices to explain p_1). Neither will she, if she has no further information, disbelieve both of them (as they both are reasonable explanations for p_1). She could believe one of them in favor of the other, but this is definitely not, epistemically, her only option. It would even be counterproductive for her. She would better neither believe nor disbelieve any of them, yet take both of them as relevant premises to reason and act upon. For instance, if the agent needs the book she ordered urgently, it is most rational for her to contact both the book retailer to confirm whether the book has been sent, and the post office to inform whether the mailman is sick and whether, in case he is sick, she can pick up her mail at the post office.

The main reason why people do not and should not form a belief in one of several mutually exclusive options, is that they generally lack resources, interest or time to really form a (rationally justified) belief in one of the possibilities. In our example, the agent actually does not care about the exact explanation of why the book has not yet arrived, she just wants to make sure that she has done everything she could to receive the book as quickly as possible. To make this point even more clear, consider the following example. Suppose you find a box of chocolates in the coffee room at the office. There may be many explanations: it could be someone's birthday, it is perhaps a leftover from some kind of celebration last night, maybe somebody bought them for his wife, but forgot them, etc. But if there is no one around to ask and you have no particular interest, these options will remain open and uncertain, while you (grab a chocolate and) go back to focus on your own work. The awareness of the options, although they are all uncertain, is still valuable: it reassures you that the box of chocolates is perfectly explainable and does not require any further action on your part. Now compare this situation with one in which you find a suspicious looking package that clearly does not belong there. Now it is clear that you should take some action and, for instance, report it. This action is then based on the awareness of some possible explanations for the package's presence. Yet you do not necessarily have the epistemic duty to find the actual explanation and form a belief in it. You are perfectly allowed to just perform an action (report it) based on your uncertain possible conclusions, but forgo any investigation into the package's actual origin. While, in this case, lack of resources such as time or interest can be a valuable excuse for not forming a belief, it is not an excuse for taking no action (as action based on uncertainties can be justified).

The kind of uncertainties described in the previous examples can and should clearly not be (epistemologically) described as beliefs or disbeliefs. But is the third attitude from the triad, withholding (judgment), appropriate to describe such uncertainties? At first sight, this seems to work: as such uncertainties figure in the agent's reasoning, she has been confronted with them, after which she formed neither a belief nor a disbelief in the proposition. In fact, this is how the attitude to such propositions would be modeled in many formal approaches.[4]

Yet, this analysis misses an important part of the picture. A withheld

[4] For instance, in a standard doxastic logic with a belief-operator B_a, the doxastic attitude of an agent a towards the uncertainty p would be modeled as $\neg B_a p \wedge \neg B_a \neg p$.

judgment is a perfectly symmetrical attitude towards p and its opposite $\neg p$. The attitudes towards the uncertainties in the previous examples and their opposites are not symmetrical. While an agent probably will acknowledge that their opposites are possibly true, these opposites will have no further relevance for the agent. On the other hand, the adopted doxastic attitudes towards the uncertainties in these examples have genuine epistemic and pragmatic value for the agent: they can (temporarily) meet a need for explanation (epistemic), they can figure as a ground for action (pragmatic) or they can be a motive to further pursue the truth value of the proposition (epistemic and pragmatic).

Let me explain the value of these adopted doxastic attitudes by means of the examples. That the box of chocolates on the table might be for someone's birthday will in general sufficiently fulfil the agent's need for explanation that might have arisen upon seeing that box (even if she knows that there are more possibilities); the opposite statement, i.e. it is nobody's birthday, will, although also possible, have no epistemic value for the agent. This statement would only have some value for her if she came to believe it, i.e. take it to be true (in which case she would be again in need of some explanation for the box of chocolates). Realizing the possibility of a package bomb is a sufficient and rational ground to report the package; realizing the possibility of the opposite statement, i.e. that there is no package bomb, is no ground for any action on the agent's part at all. Considering the possibility that the mailman is sick is a good basis to start pursuing it and to contact the post office; the possibility that he is not sick is irrelevant and not a ground for any further action.

In conclusion, if we want to describe the doxastic attitude towards every type of uncertainty in human reasoning, the classical epistemological triad 'belief - disbelief - withholding' does not suffice, and needs to be supplemented with a fourth attitude, which I will call *entertaining a hypothesis* or, in short, *hypothesizing*. This is the doxastic attitude an agent adopts towards a proposition which she does not take to be true or false as such, but of which the possible truth makes it relevant enough to take it into further consideration or pursue it, to let it figure as a ground for action, or to let it sufficiently fulfil a need for explanation. In other words, acquiring a hypothesis and keeping it in mind has both epistemic and pragmatic value, which cannot accurately be described by the concepts 'withholding' and 'belief'.[5]

[5] The term 'hypothesis' is sometimes used to indicate a mere logical possibility. It should

'Hypothesis' is of course not a new concept, and it is well-recognized that hypotheses are part of human reasoning. Therefore, it might be argued that, although a proper characterization of the attitude towards them is in order – I will attempt to offer one in Section 3.3 – it is not really necessary to supplement the triad 'belief - disbelief - withholding' with this attitude. For many purposes, one may argue, a hypothesis can be described with sufficient accuracy as a withholding of judgment, especially if its more specific properties, such as its plausibility or relevance for the agent, do not matter. If these properties do matter, this can be properly expressed by a related belief of forms such as "it is probable that *p*" or "it might be that *p*".

Above, I have already argued why it is not a good idea to characterize a hypothesis as a withholding of judgment: the doxastic attitude an agent holds towards a hypothesis is different from the attitude she holds towards the opposite of that hypothesis, something the symmetrical attitude of withholding judgment cannot capture. In the next section, I will show that the attitude of entertaining a hypothesis can also not be analyzed as or reduced to a related belief in a way that uniformly applies to all hypotheses.

3.2 Hypotheses and Beliefs

In this section, I will investigate the relation between the doxastic concepts 'entertaining a hypothesis' and 'holding a belief' and see whether the former can be specified in terms of the latter or whether they are independent. Some technical preliminaries are, however, in order.

The concept of belief that I use for this analysis is 'belief' in its technical coarse-grained meaning, as used in the classical epistemological triad 'belief - disbelief - withholding'. In this meaning, 'believing that *p*' can denote any attitude of assent towards *p*, or, in other words, any attitude that takes *p* to be true.[6]

The concept of hypothesizing that I use for this analysis, is the folk notion of entertaining a relevant hypothesis. The use of a non-technical folk notion should not worry us here, because at this point my aim is only to show the need for a notion of hypothesizing to supplement the triad

be clear that my analysis does not extend to this meaning (see also Section 3.2). My analysis is aimed at capturing the folk notion of entertaining a hypothesis relevant to the agent.

[6] It is necessary to use such a broad (technical) definition of belief, because the more specific folk notion of belief can lead to ambiguity (see Chapter 2).

'belief - disbelief - withholding', and to argue that such a notion is not reducible to this triad. In Section 3.3, I will then try to characterize the folk notion of entertaining a hypothesis in the framework of theoretical and practical doxastic attitudes, developed in Chapter 2.

Finally, it is sometimes hard to avoid intuitions about the relation between 'hypothesis' and 'belief' that originate in the way the attitude of entertaining a hypothesis is referred to in natural language, especially written language. For instance, a common way for an agent to express that she entertains the hypothesis p is to state that "it might be that p". Yet this very statement also indicates that the agent believes the expression "it might be that p". As any affirmative statement can be considered as one that the author takes to be true, any expression of an attitude that takes the form of an affirmative statement (which occurs, certainly in written language, for all kinds of attitudes such as desires, fears, hopes, etc.) can be considered as an expression of belief in that statement. Yet, the fact that our (written) language cannot express every attitude in its own particular way should not be seen as an *a priori* reason to assume that all kinds of propositional attitudes are reducible to beliefs in related propositions. At most, it shows that such an attitude implies a belief in a related statement. To illustrate this point: it is not inconceivable, in spoken language, to use a somewhat higher tone to express hypotheses, in order to display one's uncertainty, while using a firm, affirmative tone to express beliefs. In this way, different attitudes towards a single proposition are clear from the different modes of expression.

I will now survey the various strategies that can be used to analyze the attitude of entertaining a hypothesis in terms of related attitudes of belief, and show that every such attempt misses part of the essence of hypotheses. Of course, most of these strategies will capture the attitude towards some hypotheses, but none of them captures the attitude towards any hypothesis. As such, it is not possible to reduce the attitude of hypothesizing in a uniform way to related attitudes of belief.[7] I can of course not guarantee that my list of reduction strategies is exhaustive, yet I think that I have covered most of the plausible options. To reduce intuitions that originate from our use of written language, I will uniformly denote the attitude towards a hypothesis with the expressions "entertaining the hypothesis p" and "hypothesizing that p" (by which I mean the act of keeping the hypothesis p in

[7] It might be argued that this speaks against the possibility of a uniform characterization of the folk notion of a relevant hypothesis. Yet I will present such a characterization (which is not reducible to related beliefs) in Section 3.3.

mind once it has been formed) as counterparts of the expressions “holding the belief *p*” and “believing that *p*”.

Entertaining a hypothesis is possibly believing that proposition A first strategy originates from the observation that hypotheses are often propositions that the agent does not believe yet, but of which she thinks she might come to believe them later on. Hence, it could be said that hypotheses are possible (future) beliefs, i.e. propositions that will become actual beliefs if certain actualization conditions are fulfilled. Yet to assert that this same relation applies to the attitudes of hypothesizing and believing would be a category mistake. The concepts ‘possible’ and ‘actual’ refer only to the belief status of that proposition, not to all attitude statuses: an agent entertaining a hypothesis has not only a possible attitude towards that proposition, but also an actual and real attitude towards that proposition, though not one of belief.

Entertaining a hypothesis is a derivative attitude from believing a disjunction A second strategy consists in denying that the attitudes towards hypotheses are isolated attitudes and arguing that they are always related to the other possible options the agent conceives in the context at hand. Hence, one could argue that conceiving of multiple options in a context implies the thought that one of these options is true; in other words, believing their disjunction. ‘Entertaining a hypothesis’ is then merely a derivative concept that describes the attitude towards a single disjunct of a believed disjunction.

The main argument for this strategy is the common view that the main criterion for believing a particular hypothesis is the elimination of the other hypotheses or options. If this elimination is conceived as the sequential application of the logical rule *disjunctive syllogism*, this view implies that initially the disjunction of all options should be believed.

Certainly, in many practical contexts, this view is convincing and a good normative procedure to update one’s doxastic attitudes. For example, if I do not have my wallet in my pocket, I do believe either that someone took it, or it fell out of my pocket at some point, or I did not bring it in the first place. When I retrace my steps and find my wallet neither along the way nor at home, I will normally come to believe that someone took it.

However, there is no *a priori* reason why the agent always should have formed a belief in the disjunction of the various options apparent to her.

Unless the agent has some justification for why the disjunction of the options she has in mind is exhaustive, she generally will not and certainly does not have to believe the disjunction of the ideas she has in mind, precisely because she is aware of the fact that there are further options she cannot think of at that moment. In such a case, the best attitude towards the disjunction of the various options in a certain context is to consider that disjunction also simply as a hypothesis. Take the example of the mailman. There might be numerous other reasons why the mailman did not come by this morning: there might be a strike at the post office; his daily route might have been adjusted, which causes him to come by later; the book retailer might have written the wrong address on the package, etc. So there is no reason why the agent should have formed a belief in the disjunction of the two options she has in mind.

A second reason why entertaining a hypothesis should not be considered as a derivative attitude towards a disjunct of a believed disjunction is that agents evaluate the relevance of hypotheses individually according to the context at hand. Reducing them to a single believed disjunction treats all possible options as equal. This would make it hard to explain why, with just a slight adjustment in the situation, an agent retains some hypotheses while considering others as suddenly irrelevant. For instance, in the example of the mailman, the hypothesis that the mailman is sick is only relevant because the agent saw her neighbor who told her that the mailman did not look too well the other day. She would not have considered this hypothesis as relevant if she did not meet her neighbor. Yet the hypothesis that the package has not arrived yet would be relevant in either case.

Entertaining a hypothesis is believing a related proposition A third strategy to analyze the attitude of entertaining a hypothesis in terms of related attitudes of belief is to argue that entertaining the hypothesis p is nothing more than believing a related proposition. I will survey some of the most plausible candidates. Particularly in this section we must take care not to be misled by intuitions coming from our use of language, because almost all of these candidates are in certain contexts suitable ways for an agent to express that she entertains a particular hypothesis.

Entertaining the hypothesis p is believing that "it is possible that p" There are actually many ways to spell out the notion of possibility, but we can assume that what an agent expresses when she states that p is possible is that p is compatible with her set of background beliefs. We should take care, however, not to exclude the possibility of hypotheses that conflict with some beliefs the agent holds. For instance, in facing an anomaly an agent already realizes that some of her beliefs are incompatible with her observation and that she will most likely have to revise some of them. In order to do so, she must be capable to entertain hypotheses that conflict with some of her beliefs. Therefore, believing that "*p* is possible" could be best interpreted as believing that "*p* is compatible with the major part of an agent's set of (background) beliefs".

This reduction neglects the aspect that hypotheses must somehow be relevant for an agent. Even amongst the statements that are fully compatible with all of an agent's beliefs, there are many highly improbable and irrelevant options. If I cannot find my wallet, it would be compatible with (the major part of) my beliefs that some form of extraterrestrial intelligence pulled a prank on me, or that I am actually just a brain in a vat and something went wrong with the wiring. Clearly, such options should not be entertained as rational hypotheses in the context of an agent who lost her wallet.[8] But also more mundane possibilities, such as that it is possible that there are exactly 923 Roman Catholic churches in Rome or that the author of this thesis might have got precisely three new grey hairs this very morning should not be rationally entertained as hypotheses, because hypotheses need to be somehow relevant for the agent, i.e. they need to contribute in some way to purposes the agent values, such as leading to justified beliefs, suggesting ways out of an impasse, suggesting actions that could possibly prevent some harm or bring some benefit, or indicating that something puzzling is explainable even if no real explanation could be provided. Mere possibilities have no merit in any of these roles, and should, therefore, not be called hypotheses. Just as for the doxastic attitude of belief, there is a normative component in judging whether a proposition should be entertained as a hypothesis in a particular context or in general.

Entertaining the hypothesis *p* is believing that "it is sufficiently probable that *p*" As highly unlikely possibilities were the main problem for the previous suggestion, we could, by characterizing "entertaining the hypothesis p" as believing that "p is sufficiently probable", make sure that the hypothesis has at least some merit relevant to the main reason why we attribute value to hypotheses, i.e. they can conduct us to justified (true) beliefs.

[8] Unless, perhaps, she is an academic philosopher.

To assess this suggestion, we should first make clear which type of probability is meant. There are at least four different interpretations of probability in the literature: three *objective* interpretations (frequency, propensity and chance) and one *subjective* interpretation in terms of degrees of belief (Williamson, 2005, pp. 7-13).

Of the objective interpretations, the *frequency* interpretation (in short, that probability is the limit of the frequency of positive occurences) is the most straightforward and undisputed. Yet it is applicable only to repeatable contexts, and not all statements that one might entertain as hypothesis are related to such a repeatable context. Even if the context is repeatable and has occurred several times before, a problem can arise, as various relevant hypotheses would be assigned a probability of zero according to this interpretation. Consider, for instance, the following case. An agent arrives home at the end of the day, and finds the back door open. At first, she only entertains the hypothesis (and maybe even believes) that her husband/wife left it open when (s)he left. But, suddenly, she realizes that there is also another option: a burglar might have broken into her house. The first hypothesis she entertains might be connected to a degree of probability. If her husband/wife, absent-minded as (s)he is, happens to leave the back door open on a regular basis, she can calculate or estimate this probability. But, given that (in our example) no one has previously broken into her house, the probability of the second hypothesis can, according to the frequency interpretation, only be judged to be zero.

To account for the probability of hypotheses in non-repeatable contexts (or contexts, like the burglar case, that are judged to have too few occurrences to allow for a correct estimation of the frequency limit), one might turn to one of the other two objective interpretations of probability: *propensity* or *chance*. Each of these, in their own way, attempts to generalize the basic idea of the frequency interpretation to single and non-repeatable cases by relating these cases to a larger class of cases. The technical details of how this related class is constructed does not matter for our purposes and are subject to various problems, but let us grant that this can be done. For instance, in our burglar case, the agent could refer to crime statistics of the neighbourhood or city she lives in to assign a non-zero probability to the hypothesis that a burglar might have broken into her house. Yet this operation could only be called a rationalization that the agent constructs afterwards. At the moment she arrives at her house and hypothesized the possibility, she might have been entirely unaware of the rate of crime in her neighbourhood.

Finally, it might be argued that the attitude of hypothesizing should not be linked with objective probability, but with *subjective* probability. In other words, one could claim that an agent entertains the hypothesis p if and only if she has a sufficient *degree of confidence* (or *degree of belief*) in the truth of p to entertain it as a hypothesis.

This thesis is fully analogous to the Lockean thesis for the attitude of belief (see Section 2.6). Although it is actually not an analysis of the concept 'hypothesis' in terms of the concept 'belief', but in terms of 'degrees of belief', it seems appropriate to discuss this suggestion here.[9]

Analyzing hypotheses in terms of a sufficient degree of belief implies (supposing we have a reliable way to quantify degrees of belief) that a threshold should be specified. While this task is already not easy for the Lockean Thesis for belief (for which any value in the half-open interval $]0.5, 1]$ can be argued for), it is impossible to uniformly impose a threshold for entertaining a hypothesis because such a threshold would depend on the context. For instance, if I plan to go for a walk and estimate my degree of confidence in the fact that it will rain during that walk at about 0.01, I will not consider this as a relevant hypothesis (and, hence, not bother to take an umbrella); yet, my degree of confidence in the fact that my house will burn down next year may perhaps be quantified as 0.0001. Still, I do consider this as a relevant hypothesis (otherwise I would not have insured my house). Therefore, the doxastic attitude of entertaining a hypothesis cannot be purely analyzed in terms of a sufficient degree of confidence in its truth, since whether we entertain a proposition as hypothesis depends on the circumstances.

Entertaining the hypothesis p is believing that "it is plausible that p" or that "it might well be that p" Finally, since all previous attempts seem to maroon on the fact that they do not somehow take into account the relevance of the hypothesis for the agent, we could maybe analyze the attitude of entertaining a hypothesis in terms of a belief in a statement that

[9] As I argued in Section 2.6, I have my reservations about the viability of the Lockean thesis for beliefs (considered either as a coarse-grained notion, or as the folk notion). But, even if the Lockean thesis were applicable to both the attitudes of belief and hypothesizing, the analysis of these two doxastic attitudes in terms of a suitable degree of belief would not be a sufficient reason to argue against the need to supplement the triad 'belief - disbelief - withholding' with a notion of hypothesis. In fact, as I will discuss in Section 3.3, I do consider (the folk notion of) 'hypothesis' to be reducible, not to 'degrees of belief' alone, but to a combination of 'degrees of belief' and 'acceptances' (like the folk notion of belief, see Chapter 2).

expresses this relevance, such as "*p* is plausible", "it is to be kept in mind that *p* is possible" or "it might well be that *p*".

But such statements are actually expressions of the fact that the agent has already developed an attitude towards *p* of entertaining it as a hypothesis. In other words, beliefs of this kind are *second-order doxastic attitudes*, i.e. doxastic attitudes the content of which is an expression or report of a doxastic attitude itself. In essence, believing that "*p* is plausible" is equivalent with believing that "I entertain the hypothesis *p*".

Such second-order beliefs are formed when the agent expresses or reports on her doxastic attitudes. After having been confronted with a proposition *p* and having formed a doxastic attitude towards *p*, an agent can by means of introspection report on it and state, for instance, "I consider *p* as a relevant hypothesis". Yet in expressing a (doxastic) attitude report in an affirmative way, the agent also expresses that she believes that affirmative statement, i.e. that doxastic attitude report. Hence, expressing attitudes in an affirmative way (which is possible for many attitudes) implies at the same time the formation of a second-order belief towards that expression, yet such beliefs are not the same as the first-order attitudes themselves.

My argument for that this is the case for beliefs in statements of the form "*p* is plausible" (or a similar form) takes the form of a *reductio* argument.

In order to analyze the attitude of entertaining the hypothesis *p* (which is, if *p* contains no doxastic attitude reports itself, a first-order attitude) in terms of a belief in a statement of the form "it is plausible that *p*" (or a similar form), two conditions should be met: (1) a belief of that form should be held for every proposition that is entertained as a hypothesis, and vice versa (otherwise the suggestion was not a viable candidate to reduce the attitude of entertaining a hypothesis to in the first place); and (2) a belief in a statement of that form has to be a first-order doxastic attitude (otherwise we would reduce a first-order attitude towards *p* to a second-order attitude).

Let us take the conjunction of these two conditions as the *reductio hypothesis*, more precisely (for any proposition *p* that contains no doxastic attitude reports):

(H_1) An agent entertains a hypothesis *p* if and only if she also holds the uniquely related belief that "it is plausible that *p*".

(H_2) Believing that "it is plausible that *p*" is a first-order belief.

On the one hand, as any entertained first-order hypothesis, according to H_1, uniquely relates to a held belief that is, according to H_2, a first-order doxastic attitude, the size of the set of an agent's entertained first-order hypotheses has to be smaller than or equal to the size of the set of her held first-order beliefs. More formally, for every agent a, taking $\mathcal{P}$ to be the set of all propositions (in the considered language) that contain no doxastic attitude reports and $B_a p$ and $H_a q$ operators indicating that the agent a holds the belief p and entertains the hypothesis q respectively:

$$|\{p \mid p \in \mathcal{P} \wedge H_a p\}| \leqslant |\{p \mid p \in \mathcal{P} \wedge B_a p\}| \tag{3.1}$$

On the other hand, holding the first-order belief p will, if the agent is minimally rational, also imply a belief in the statement "it is plausible that p". This is, by virtue of the reductio hypothesis H_2, a first-order belief of the agent, and, moreover, according to H_1, a belief that is uniquely related to an entertained first-order hypothesis, i.e. p. Therefore, any first-order belief in a statement implies that that proposition is also entertained as a hypothesis. Hence, the set of an agent's held first-order beliefs is a subset of her set of entertained first-order hypotheses, or more formally:

$$\{p \mid p \in \mathcal{P} \wedge B_a p\} \subseteq \{p \mid p \in \mathcal{P} \wedge H_a p\} \tag{3.2}$$

From (3.1) and (3.2) we can conclude that:

$$\{p \mid p \in \mathcal{P} \wedge B_a p\} = \{p \mid p \in \mathcal{P} \wedge H_a p\}$$

In other words, the set of an agent's held first-order beliefs is the same as the set of her entertained first-order hypotheses – a contradiction with the fact that hypotheses need not be believed.

Therefore, the reductio hypothesis $H_1 \wedge H_2$ does not hold, and, hence, either $\neg H_1$, i.e. believing that "it is plausible that p" is not the same as entertaining the hypothesis p (in which case the attitude of hypothesizing should clearly not be reduced to a belief in such an expression) or $\neg H_2$, i.e. believing that "it is plausible that p" is a higher-order doxastic attitude the content of which is a doxastic attitude report itself. In our case, the only attitude the expression "it is plausible that p" could be a report of is that the agent entertains the hypothesis p. Hence, the attitude of entertaining the hypothesis p can clearly not be reduced to such a belief.

Conclusion Through my overview of various strategies to reduce the doxastic attitude of entertaining a hypothesis to related attitudes of belief, I have shown that entertaining a hypothesis is a genuine doxastic attitude, which is *sui generis*. This implies that, as the attitude of hypothesizing is not reducible to the other attitudes of the triad 'belief - disbelief - withholding', and as the triad appeared to be inadequate to describe the doxastic attitude towards the uncertainties in an agent's reasoning (which we identified as hypotheses), we can conclude that the triad should be supplemented with a notion of hypothesis to make it adequate to describe the agent's doxastic attitude towards the various statements present in her reasoning.

3.3 The Doxastic Attitude of Entertaining a Hypothesis

So far, I have used a folk notion of the attitude towards a (relevant) hypothesis. In this section, I will specify this notion more precisely by drawing on the framework I constructed in Chapter 2 to describe doxastic attitudes in a more fine-grained manner.

As the various examples I used throughout this chapter already suggest, I take the folk notion of a (relevant) hypothesis to be a doxastic concept that specifies both the theoretical and practical doxastic attitude of an agent towards a proposition, just like the folk notion of belief (see Chapter 2). In the case of entertaining a hypothesis, the practical meaning is clearly the dominant one of these two: to entertain a hypothesis is in the first place to adopt a policy to rely on that statement in specific contexts or circumstances. Yet to some extent the theoretical doxastic attitude is also specified: to be able to entertain a proposition as hypothesis, the agent must have a degree of confidence in it that is strictly greater than having zero confidence. In other words, the agent must at least acknowledge that it is possibly true. It is, however, not possible to raise this threshold for confidence, because (as we have seen in the previous discussion of whether the attitude towards hypotheses could be analyzed in terms of a sufficient degree of belief) hypotheses that have an immense impact on an agent's life are often entertained even with extremely low degrees of belief (see, for instance, the hypothesis that a suspicious package is a package bomb). The only thing that is really required regarding the theoretical doxastic attitude is that the agent has a non-zero degree of confidence in the possible truth of the hypotheses she entertains.

Having some (non-zero) degree of confidence in the truth of a proposition is clearly not enough to entertain it as a hypothesis, as one could also adopt an attitude of withholding judgment or disbelief. For instance, towards the proposition that "there are 923 Roman Catholic churches in

Rome”, I would adopt an attitude of withholding judgment. It might be the case that it is so, but it could also be a few dozen more or less than 923, and, unless I have some special interest in the accuracy of this number, it does not matter for me at all. Hence, my attitude towards this proposition will be symmetrical with my attitude towards any other proposition stating a reasonable number; in other words, it would be an attitude of withholding. For an example of disbelief, suppose I have a degree of confidence of 0.01 in the proposition (p) “it will start to rain in the next hour” and I have decided to go for a walk now. Then my attitude towards p would normally be one of disbelief. The possible occurrence of p is clearly relevant for me, but my low degree of confidence causes me to form a belief in its opposite $\neg p$, and, basing my reasoning on this belief, I decide not to take an umbrella with me. Compare this with the case in which my degree of confidence in p is 0.25. In this case, I typically would neither form a belief nor disbelief in p. Yet, as the possible truth of p is clearly relevant for me, my attitude towards p is now not one of withholding judgment, but of entertaining it as a hypothesis; and this hypothesis can cause me to decide to take an umbrella during my walk.

Hence, the doxastic concept ‘entertaining a hypothesis’ is a description not only of the agent’s theoretical doxastic attitude, but also of her practical doxastic attitude. In order to specify this, let me start by defining a few types of contexts.

I define a *research context* as a context that satisfies the following three criteria:

(a) the context is clearly constrained in space, time and resources.
(b) in this context, the agent’s main purpose is to improve her doxastic attitudes; in other words, the agent’s intentionality is *mind-to-world*.
(c) the actions available to an agent are limited to those that have, from the agent’s perspective, negligible negative consequences in case these actions are based on wrong assumptions, except for the possible loss of the invested resources.

Scientific research is obviously an example of such a context, yet many everyday situations can also be accurately described as a research context, such as (in our mailman example) calling the post office to inform whether the mailman is sick and whether the package has already arrived.

I define a *context requiring action in face of uncertainty about p* as a context that satisfies the following criteria:

(a) in the context, the agent has a clear option to perform a certain action.
(b) from the perspective of the agent, performing this action has, in case *p* turns out to be true, a clear benefit or prevents a clear harm.
(c) the agent perceives the negative consequences of this action in case *p* turns out to be false as clearly not weighing up to the benefit of her actions in case *p* turns out to be true, scaled by her degree of confidence in *p*

A clear example of such a context is, for instance, the discovery of a suspicious package, in which case the trouble of reporting it does not offset the harm that could be caused in case it is a bomb, even if the agent considers this possibility to be very unlikely. But also more mundane situations fit this definition. For instance, if I have a 0.25 degree of confidence in the fact that it will start to rain in the next hour, the trouble of having to carry an umbrella does not match the trouble of being soaked given the substantial degree of confidence I have.

By means of these two types of contexts, I can now specify the practical doxastic meaning of entertaining the hypothesis *p* as adopting a policy to rely on the proposition *p* for research contexts and contexts requiring action in face of uncertainty about *p*.

In summary, the notion of 'hypothesis' can be defined as a doxastic concept that specifies both the theoretical and practical doxastic attitude of an agent towards a proposition. *Entertaining the hypothesis p* means both that the agent has a non-zero degree of confidence in the truth of *p*, and that the agent has a policy to rely on *p* or trust *p* in research contexts and in contexts requiring action in face of uncertainty about *p*.

3.4 Hypotheses and Science, Rationality and Skepticism

The view sketched in this chapter, i.e. considering the attitude towards a hypothesis as a necessary supplement to the classical epistemological triad 'belief - disbelief - withholding' and my characterization of this attitude in terms of the agent's theoretical and practical doxastic attitude towards that proposition opens up various ways for future research concerning particular epistemological issues.

Connection between Epistemology and the Philosophy of Science The main goal of including this epistemological part in my dissertation has been to understand the attitude towards a hypothesis not merely as a matter of scientific methodology, but also more generally as a common human doxastic attitude.

It can be easily observed that my definition of scientific hypotheses (introduced in Chapter 1), i.e. statements about the empirical world that have an unknown or underdetermined truth status and that are advanced as tentative answers to particular research questions,[10] fits the characterization of the attitude towards a hypothesis presented in this chapter. As scientific hypotheses have an unknown truth status, scientists do not take them to be true as such and have, therefore, not adopted an attitude of belief towards them.[11] But, as they consider them as tentative answers to the research questions they pursue, scientists certainly have a non-zero degree of confidence in the possible truth of scientific hypotheses. For the same reason, these scientific hypotheses are also clearly relevant and valuable to them. Therefore, scientists are inclined to trust scientific hypotheses for research contexts that are aimed at improving their doxastic attitudes towards these hypotheses.

The fact that scientific hypotheses are a specific type of the common doxastic attitude of entertaining a hypothesis justifies why important aspects of hypotheses formation and their role in science can also be studied in subfields of philosophy that aim to capture human reasoning more broadly, such as epistemology or logic.

The Dynamics of Human Reasoning As human thinking is in constant, vigorous flux, the doxastic attitude of an agent towards a certain statement may change quickly depending on new information or further reasoning: hypotheses can become beliefs if the available alternatives turn out to be impossible; beliefs can become hypotheses if one is confronted with a new idea; a proposition towards which one has so far withheld judgment, might prove relevant after all and become a hypothesis; a hypothesis might prove

[10] For an elaboration of this position, see Section 9.2.

[11] The intended notion of belief in this context is its coarse-grained technical meaning of “taking a proposition to be true”. It does happen that scientists, certainly in later stages of research, develop an attitude towards their hypotheses that could be described as belief in its folk meaning. Yet, it has been argued (e.g. Van Fraassen, 1980) that belief (in its folk meaning) is an inappropiate attitude in scientific research, as the goal of scientists should be to construct well-accepted theories (an attitude of acceptance by many scholars for the contexts of their intended application).

to be so unlikely that it becomes, on further consideration, a disbelief; or hypotheses and beliefs may lead to a quest for confirmation which in turn might result in knowledge.

To model these dynamics, one can conceive of 'hypothesizing' as a doxastic attitude towards a proposition that is weaker than 'belief' (in the same way that 'belief' is weaker than 'knowledge'), but which still expresses a genuine and positive doxastic attitude towards a proposition and which is clearly more than having no opinion or withholding judgment about it. This leaves open the possibility of constructing a normative theory about which requirements a rational hypothesis should further fulfill in order to be considered as a justified belief.

Conceiving the relations between hypotheses, beliefs and knowledge in this way makes it possible to construct doxastic-hypothetic logics or even epistemic-doxastic-hypothetic logics, in which human reasoning that incorporates both hypotheses and beliefs (and maybe even knowledge) could be modeled. Traditionally, the attitudes of belief and knowledge are modeled by means of modal box operators: $\Box_B$ or B for belief and $\Box_K$ or K for knowledge (Hintikka, 1962). Such a box operator could certainly not be used to model the attitude of entertaining a hypothesis, as it can be rational to entertain mutually exclusive hypotheses. Yet, what might work is a diamond operator $\Diamond_H$, for which the axiom $\Diamond_H A \supset \neg\Box_B\neg A$ holds. While the construction of such a logic is already a worthy challenge in itself, it also makes it possible to create logics that can model in a qualitative way transitions from one doxastic attitude to another. For instance, such a logic could be capable of the (non-monotonic) abduction of particular hypotheses given certain sets of beliefs. While the logic $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$ of chapter 4 is clearly not intended to be such a doxastic-hypothetic logic (it has only one modal operator), it might give an idea of how this can be worked out, as in this logic hypotheses are modeled with the aid of a diamond operator.

A skeptical attitude and living with uncertainties Acknowledging the existence of a doxastic attitude towards a proposition that does not rationally imply a disbelief in the other options, yet recognizes the value of the proposition's possible truth (such that it can be a basis for further action or reasoning), also allows us to specify better the existence of a skeptical attitude.

A *skeptical attitude* can be considered as an attitude to form beliefs only if they are rational, or, in other words, if they are sufficiently warranted.

Of course, the nature of this warrant and the level of its threshold are a matter of debate, ranging from a minimally critical attitude to a full-blown Pyrrhonism, which denies the rationality of any belief whatsoever.

A key problem for skeptical attitudes that require high warrants is often that they lack sufficient ground for actions, as it is generally assumed that (conscious) actions require certain evaluative stances or beliefs (Wieland, 2012). If an agent only very hesitatingly adopts new beliefs or even no beliefs at all, it appears that she will have problems functioning and acting normally in society. An often cited example, apparently originating from Sextus Empiricus, is that skeptics cannot help someone so long as they cannot bring themselves to form a belief that this person is indeed in distress.

For skeptical attitudes with high warrants but that accept the possibility of sufficiently warranted beliefs (hence, not full-blown Pyrrhonism), the doxastic attitude of entertaining hypotheses might explain how normal action is still possible, because, as we saw, hypotheses can be a sufficient ground for action in certain contexts. The classic example of the person in distress poses no problem in this case, because as soon as the agent realizes that someone might be in distress, she can consider the context as a context requiring action in face of uncertainty about whether that person is actually in distress.

However, use of the attitude of hypothesizing cannot solve this problem for the actual Pyrrhonist, because, as Wieland (2012) shows, a Pyrrhonist is not only unable to form the belief that someone is in distress, but also unable to form the belief that people in distress should be helped. As a Pyrrhonist has suspended all his beliefs, he has also suspended his beliefs concerning desirable outcomes or what he should do. Therefore, in my conceptual framework, a Pyrrhonist cannot judge whether a context is a context requiring action in the face of uncertainty.

The doxastic attitude of entertaining a hypothesis, as I defined it, clearly depends on the existence of beliefs. More precisely, it requires certain beliefs about which ends are valuable and relevant for the agent. This should not, however, necessarily contradict the idea that rational beliefs are hypotheses that are sufficiently warranted. As I explained above, the doxastic attitude towards a proposition is dynamic, which means that rational consideration can lead one not only to consider a hypothesis as a belief, but also the other way around, to entertain a previously held belief again as a mere hypothesis. Given that no human agent, in becoming rational, starts from a blank slate, previously held beliefs about the value of certain ends, which may or may not be rational, can initially act as the background against which new hypotheses can be entertained. Yet, as rational thinking develops, hypotheses, beliefs and values could become more and more rationally justified and aligned. Of course, it remains for future research how such a process could exactly develop over time.

In conclusion, I want to state that the most important advantage of carving out an epistemological niche for the attitude of entertaining a hypothesis is that doing so allows people to rationally endorse and settle with their uncertainties, without making them incapable of normal and moral behaviour and without obliging them to adopt unwarranted beliefs.

Part II

Logical Patterns

motivation

In this part, I will zoom in to the perspective of individual reasoning steps. Focusing on the micro structure of hypothesis formation processes shows us that the reasoning steps at the core of these processes are often similar. Therefore, we can describe these steps in terms of generalized patterns, by abstracting away from the specific content of the reasoning and retaining only the formal structure of the inferences – a structure, which can then be studied by means of suitable logics. As a result, although it focuses on the micro structure of human hypothesis formation, this is the most formal part of the dissertation.

My main goal is to formally explicate some of the more common patterns of hypothesis formation in science by applying existing logical tools. At the same time, I will make use of my own contributions to reflect critically on the prospects and shortcomings of this project.

The core assumptions at the heart of this part of the dissertation are (1) that such patterns can be found in the micro structure of the human reasoning processes of hypothesis formation; (2) that such patterns can be studied formally; and (3) that this can be done by using logics as a tool. Let us turn to the motivation for adopting each of these assumptions.

Patterns of Hypothesis Formation In the historical introduction to the literature on discovery and abduction (Section 1.2), we saw that the quest to characterize rational discovery in science under a single schema was abandoned around 1980. The main reasons were that such attempts (e.g. Hanson's proposal to call abduction "the logic of discovery") often did not provide much detailed guidance for actual discovery processes, and that even these general attempts always captured only a part of the discovery process (e.g. *Inference to the Best Explanation* describes only the selection of hypotheses, not their formation).

Around the same time, research from different fields such as philosophy of science based on historical cases, artificial intelligence and cognitive science resulted in a new consensus that there is a plenitude of patterns, heuristics and methods of discovery, which are open to normative guidance, yet this guidance might be content-, subject-, or context-dependent.

My first assumption, i.e. that these patterns in hypothesis formation exist, could be defended by simply referring to this part of the literature. Yet I want to use this opportunity to add some qualification to my claim. Trying to formally explicate particular patterns of hypothesis formation would have little impact were there an infinity of rather seldom and *ad hoc* patterns. Hence, I assume not merely that such patterns exist, but also that the majority of instances of hypothesis formation can be described by a rather limited number of such patterns.[1]

To substantiate this claim, I start by referring to the literature on abduction (which I take, as explained in Section 1.4, to be a subcase of hypothesis formation in general), in which various authors have tried to provide classifications of patterns of abduction (Thagard, 1988; Schurz, 2008a; Hoffmann, 2010). Although these attempts differ slightly, some general patterns clearly stand out.

Before I give my personal classification of these major patterns found in abductive reasoning, it is important to note that abductive inferences form explanatory hypotheses for observed facts using the agent's background beliefs (or knowledge). Therefore, these patterns have the structure of the inference of a hypothesis (HYP) from some observed facts (OBS) and some of the agent's background beliefs (or knowledge) (BBK).

In line with the Fregean tradition, I consider *factual statements* as statements of a *concept* with regard to one or more *objects* (or a logical combination of such statements). For instance, the statement "There was a civil war in France in 1789" can be analyzed as the concept "civil war" with regard to "France in 1789". A *fact* is a true factual statement. As such, concepts can

[1] As I do not argue for this claim on *a priori* grounds, I do not assume that an exhaustive set of such patterns can be given. In fact, I see no arguments against the *a priori* possibility of constructing new patterns of hypothesis formation, whether by humans or by artificially intelligent agents.

also be considered as the *class* of all objects (or tuples of objects) for which the concept with regard to that object (or tuple of objects) is a fact. An *observed fact* is a factual statement describing an agent's observation that she considers to be true. This can be broadly conceived to include also, for instance, a graph or a table of measurements in an article. Together, the observed facts form the *trigger* for the agent.

In my semi-formal description of these patterns, I express that p should be considered as a hypothesis by using a formulation of the form "It might be that p"; beliefs and observed facts can be expressed simply by stating their content. Concepts are denoted by uppercase letters, objects by lowercase letters. Addition of a subscript denotes a finite list of objects or concepts (including, unless stated otherwise, the possibility of a single object or concept).

1. ***Abduction of a Singular Fact***

   (OBS) F with regard to x_i
   (BBK) E with regard to some objects explains F with regard to those objects

   ---

   (HYP) It might be that E with regard to x_i

   Some examples of this pattern, which has also been called "simple abduction" (Thagard, 1988), "factual abduction" (Schurz, 2008a) and "selective fact abduction" (Hoffmann, 2010), are:

   - the inference that the hominid who has been dubbed Lucy might have been bipedal, from observing the particular structure of her pelvis and knee bones and knowledge about how the structure of pelvis and knee bones relates to the locomotion of animals.
   - the inference that two particles might have opposite electric charges, from observing their attraction and knowledge of the Coulomb force.

2. ***Abduction of a Generalization***

   (OBS) F with regard to all observed objects of class D
   (BBK) E with regard to some objects explains F with regard to those objects

   ---

   (HYP) It might be that E with regard to all objects of class D

Some examples of this pattern, which has also been called “rule abduction” (Thagard, 1988), “law abduction” (Schurz, 2008a) and “selective law abduction” (Hoffmann, 2010), are:

- the inference that all hominids of the last three million years might have been bipedal, from observing the similar structure of the pelvis and knee bones of all observed hominid skeletons dated to be younger than three million years and knowledge about how the structure of pelvis and knee bones relates to the locomotion of animals.
- the inference that all emitted radiation from a particular chemical element might be electrically neutral, from observing in all experiments conducted so far that radiation emitted by this element continues in a straight path in an external magnetic field perpendicular to the stream of radiation and knowledge of the Lorentz force and Newton’s second law.

3. ***Existential Abduction***, or the abduction of the existence of unknown objects from a particular class

(Obs) F with regard to x_i
(Bbk) the existence of objects y_i of class E would explain F with regard to x_i

---

(Hyp) It might be that there exist objects y_i of class E

Some examples of this pattern, which was already called “existential abduction” by Thagard (1988), and has also been called “first-order existential abduction” (Schurz, 2008a) and “selective type abduction” (Hoffmann, 2010), are:

- the inference that a hominid of the genus *Australopithecus* might have lived in this area, from observing a set of vulcanized foot imprints and the belief that these foot imprints are of an *Australopithecus*.
- the inference that there might be other charged particles in the chamber, from observing deflections in the path of a charged particle in a chamber without external electric or magnetic fields and knowledge of the Coulomb and Lorentz forces and Newton’s second law.

4. ***Conceptual Abduction***, or the abduction of a new concept

   (OBS) F_i with regard to multiple x_j individually
   (BBK) No known concept explains why F_i for all x_j

   (HYP) It might be that there is a similarity between the x_j, which can be labeled with a new concept E, that explains why F_i with regard to all the various x_j individually

   Some examples of this pattern, which largely coincides with the various types of "second order abduction" Schurz (2008a) suggests,[2] and several of the types of "creative abduction" conceived by Hoffmann (2010), are:

   - the inference that there might be a new species of hominids, from observing various hominid fossils that are similar in many ways and believing that these fossils cannot be classified in the current taxonomy of hominids.
   - the inference that there might exist a new type of interaction, from observing similar interactive behavior between certain types of particles in similar experiments and believing that this behavior cannot be explained by the already known interactions, properties of the involved particles and properties of the experimental setup.

Using the terminology of Magnani (2001) and following the distinction of Schurz (2008a), the first two patterns, abduction of a singular fact and abduction of a generalization, can be considered as instances of *selective abduction*, as the agent selects an appropriate hypothesis in her background knowledge, while the latter two, existential abduction and conceptual abduction, can be called *creative abduction*, as the agent creates a new hypothetical concept or object.[3]

[2] It was Schurz who pointed out that this pattern is rational and useful for science only if the observation concerns several objects each individually having the same or similar properties, so that some form of conceptual unification is obtained. Otherwise, for each fact it could be suggested that there exists an *ad hoc* power that explains (only) this single fact.

[3] Hoffmann (2010) would dispute this distinction, as he sees the third pattern (existential abduction) in the first place as the selection of an already known type (e.g. the genus *Australopithecus*), and not so much as the creation of a new token (someone of this genus of which his/her existence is now hypothesized).

As stated before, my list is not exhaustive. Further patterns have been identified, such as the abduction of a new perspective (Hoffmann, 2010), e.g. to suggest that a problem might have a geometrical solution instead of an algebraic one; "analogical abduction" (Thagard, 1988), e.g. explaining similar properties of water and light, by hypothesizing that light could also be wave-like; or "theoretical model abduction" (Schurz, 2008a), i.e. explaining some observation by suggesting suitable initial conditions given some governing principles or laws. Some have even considered "visual abduction", the inference from the observation itself to a statement describing this observation, as a separate pattern (Thagard and Shelley, 1997). For some of these patterns (or instances of them), it is possible to argue that they are a special case of one of the patterns above. For instance, the suggestion of the wave nature of light can also be seen as an instance of conceptual abduction, in which the (mathematical) concept 'wave behavior' is contructed to explain the similar properties of water and light; yet it is true that the analogical nature of this inference makes it a special subpattern with interesting properties in itself.[4]

Perhaps more important to note is that these patterns are not mutually exclusive given a particular instance of abductive reasoning. For instance, the inference that leads to the explanation of why a particular piece of iron is rusted can be described both as singular fact abduction (this piece of iron underwent a reaction with oxygen) or as existential abduction (there were oxygen atoms present with which this piece of iron reacted). But in essence it describes the same explanation for the same explanandum. Also, combinations occur. For instance, if a new particle is hypothesized as an explanation for an experimental anomaly,[5] then we have both an instance of existential abduction (there is a not yet observed particle that causes the observed phenomenon) and an instance of conceptual abduction (these hypothesized particles are of a new kind).[6] Yet in the mind of the scientist, this process of hypothesis formation might have occurred in a single reasoning step.

We should not, however, be too worried about these issues, if we remember that these patterns are categories for linguistic descriptions of actual reasoning processes. Any actual instance of hypothesis formation can

[4] This is also how Schurz (2008a) presents it; in his classification, analogical abduction is one of the types of second order existential abduction he conceives of.

[5] See, for instance, Wolfgang Pauli's suggestion in the case of the β spectrum (Chapter 7).

[6] I think this particular combination coincides with Hoffmann's (2010) pattern of "creative fact abduction".

be described in several ways by means of natural language, and some of these expressions can be formally analyzed in more than one way. Therefore, I do not think that we should focus too much on the exact classification of particular instances of hypothesis formation. Yet this does not render meaningless the project of explicating various patterns of hypothesis formation. The goal of this project is to provide normative guidance for future hypothesis formation. If particular problems or observations can be looked at from different perspectives and, therefore, expressed in various ways, it is only beneficial for an agent to have multiple patterns of hypothesis formation at her disposal.

So far, I have argued only for my first assumption, i.e. that the majority of instances of hypothesis formation can be described by means of a limited number of patterns, for the case of explanatory hypothesis formation or abduction. The following patterns of hypothesis formation are normally not mentioned in the literature on abduction, as the inferred hypotheses tend to be seen as non-explanatory. But also for non-explanatory instances of hypothesis formation – I leave it open whether these patterns or their instances are really non-explanatory – there are some very general patterns, the most common ones being suggested belief revision and inductive generalization.

5. ***Suggested Belief Revision***

(Obs) F with regard to x_i
(Bbk) G with regard to x_i
(Bbk) F with regard to x_i is apparently in contradiction with G with regard to x_i

---

(Hyp) It might be that G should be revised so that F with regard to x_i is not in contradiction with G' with regard to x_i

Some examples of this pattern, which is common in cases where the trigger is an anomaly and which is related to what is studied in the field of Belief Revision (though the inference is weaker, as it is only hypothesized that a belief should be revised), are:

- to suggest that the idea that apes not belonging to the genus *Homo* are not bipedal should be revised, upon observing that Lucy (of the genus *Australopithecus*) is bipedal and noticing that this is in contradiction with this idea.
- to suggest that the principle of energy conservation might not

hold in the case of β decay, upon observing energy curves of the emitted β particles and noticing that these are apparently in contradiction with the belief that the conservation of energy holds in all cases.[7]

6. ***Inductive Generalization***

(OBS) F with regard to all observed objects of class D

(HYP) It might be that F with regard to all objects of class D

Some examples of this well-studied inductive pattern are:

- the inference that all members of the genus *Australopithecus* are bipedal, from observing that Lucy and all other fossils found of this genus are bipedal.
- the inference that all oppositely charged particles attract each other, from observing the attraction between all observed oppositely charged pairs so far.

Again, further patterns might be discerned, and various instances of hypothesis formation can often be described according to various patterns. Also, it should not be assumed that suggested belief revision is the only pattern that occurs when scientists encounter anomalies or contradictions. It often happens that scientists confronted with anomalies suspend their judgment with regard to the contradicting parts of their belief set, while pursuing other patterns that involve parts of their belief set that are consistent with the observation.

Formal Patterns By listing these various patterns, I already implicitly illustrated the second core assumption of this part of the dissertation, i.e. that all of these patterns of hypothesis formation can be formally described by a pattern schema that abstracts from the specific context. This seems to be at odds, however, with the consensus that has been reached in the literature on discovery, which stresses the context and content dependency of the various methods of discovery. The main worry for the project of this part of the dissertation is that if patterns are so dependent on the field, discipline or context, this project would have little impact.

[7] This is in essence Niels Bohr's suggestion in the case of the β spectrum (see Chapter 7).

Yet one should note that the patterns of hypothesis formation are not the same methods, procedures and heuristics that are described in the discovery literature. The listed patterns differ in two important aspects from those in the discovery literature.

First, I consider only the formation or generation of hypotheses, which is only a part of scientific discovery. It is clear that various other parts of discovery, such as hypothesis selection or the search for relevant data, are content and discipline dependent. Second, it is true that there may exist quite specific patterns for hypothesis formation in particular fields or paradigms. For instance, in particle physics, given certain observations, it is common practice to presume the possible existence of a new type of particle. But if one considers carefully the actual structure of such methods, one sees that they are often more specific instances of one of the more abstract patterns detailed above. These field dependent patterns have their value in scientific practice, but pose no argument against the possibility of a formal description of patterns of hypothesis formation.

The use of formal logics This brings us to the third, and maybe most surprising core assumption of this part of the dissertation: that these patterns can be modeled using formal logics. To those who might be surprised, I want to stress that I do not mean that any of these patterns is a valid inference in classical logic or any other (non-trivial) deductive logic. To model defeasible reasoning steps such as hypothesis formation, one has to use non-monotonic logics: logics for which an extension of a premise set does not always yield a consequence set that is a superset of the original consequence set. Or, put more simply, logics according to which new information may lead us to revoke old conclusions.

It is important to note that my purpose in using logics is not the classical purpose of the discipline of logic. Classically, the discipline of logic studies the correct way to infer further knowledge from already known facts. The correct way should guarantee the truth of the new facts, given that the old facts are true. Accordingly, this has motivated the search for the right (deductive) logic (whether it be Classical Logic or another one). My purpose, however, is to model or explicate human reasoning patterns. As these patterns are fallible, leading to conclusions that are not necessarily true even if the premises are true, it should be possible to revoke previously derived results; hence, my use of non-monotonic logics. Also, because there are many patterns of human reasoning, I naturally conceive of a plenitude of logics in order to describe them.

Let me explain this a bit more formally. A logic can be considered as a function from the power set of the sentences of a language to itself. So, given a language L and the set $\mathcal{W}$ of its well-formed formulas:

$$\mathbf{L} : \wp(\mathcal{W}) \rightarrow \wp(\mathcal{W})$$

Hence, a logic determines for every set of sentences (or premise set) Γ which sentences can be inferred from it ($Cn_{\mathbf{L}}(\Gamma) =_{df} \mathbf{L}(\Gamma)$). Therefore, as a reasoning pattern is nothing more than the inference of some statements given some initial statements, in principle, a logic can be devised to model any reasoning pattern in science.[8] If this pattern can be formally described, description by a formal logic is in principle possible.

Deductive logics, such as Classical Logic (**CL**), have the property of monotonicity, i.e. for all premise sets Γ and Γ':

$$Cn_{\mathbf{L}}(\Gamma) \subseteq Cn_{\mathbf{L}}(\Gamma \cup \Gamma')$$

Most patterns of human reasoning, however, do not meet this criterion. For instance, if an agent infers a hypothesis, she is well aware that it might need to be revoked on closer consideration of the available background knowledge or in light of new information.

Although non-monotonic reasoning has typically received less attention in the field of logic than monotonic reasoning, various frameworks for defeasible reasoning and non-monotonic logics are available. For this dissertation, I use the adaptive logics framework, which was created by Diderik Batens (Ghent University) over the past three decades.[9] This framework for devising non-monotonic logics has some advantages that suit the project of this part of the dissertation well.

First, the focus in the adaptive logics program is, in contrast with other approaches to non-monotonic reasoning, on proof theory. For these logics, a dynamic proof style has been defined in order to mimic to a certain extent actual human reasoning patterns. More in particular, these dynamic proofs display the two forms of revoking previously derived results that can also

[8] In reality, scientific and human reasoning include not only sentences or propositions, but also direct observations, sketches and various other symbolic representations. Yet for the purpose of modeling particular reasoning patterns, we can generally represent those sources by suitable propositions.

[9] For an extensive overview and thorough formal introduction, see Straßer (2013) or the online available manuscript of Batens (n.d.).

be found in human reasoning: revoking old conclusions on closer consideration of the available evidence (internal dynamics) and revoking them in light of new information (external dynamics).[10]

Second, over the years, a solid meta-theory has been built for this framework, which guarantees that if an adaptive logic is created according to certain standards (the so-called "standard format"), many important metatheoretical properties are generically proven. This creates an opportunity for projects such as mine to focus almost exclusively on the application of these formal methods without having to worry too much about proving their meta-theoretical characteristics.

Finally, as the framework is presented as a unified framework for nonmonotonic logics, it has been applied in many different contexts. Over the years, adaptive logics have been devised for paraconsistent reasoning, induction, argumentation, deontic reasoning, abduction, etc. This gave me plenty of inspiration and the foundations on which to build my own project.[11]

To this I want to add that from a logician's point of view, my logical applications might look somewhat unfamiliar as both of the logics I define (in Chapter 4 and Chapter 6) have a rather restricted modal language schema; the first logic even has a restriction on possible premises. This should be understood from the perspective that my purpose is to provide an explication of reasoning patterns, not to explore all possible derivations in a certain logic. My modal extensions of the language of first-order logic are the simplest possible ones to model the studied hypothesis formation pattern in a sensible way (given that the propositions occurring in the pattern can be described in the language of first-order logic). This method has the further advantage that, for instance, the first logic I define (in Chapter 4) has a lower complexity than standard adaptive logics (because, as it will be explained, it uses the simple strategy instead of the more common reliability or minimal abnormality strategies), which makes it fit for applications

[10] One should not be misled, however, by this idea of dynamic proofs in thinking that the consequence set of adaptive logics for a certain premise set depends on the proof. Adaptive logics are proper proof-invariant logics that assign for each premise set Γ exactly one consequence set $Cn_{\mathbf{L}}(\Gamma)$.

[11] Most applications of the adaptive logics framework have been studied at the Centre for Logic and Philosophy of Science (Ghent University). At the Centre's preprints list (`http://logica.ugent.be/centrum/writings/pubs.php`), references can be found to many papers in various contexts. The reference works mentioned earlier, Straßer (2013) and Batens (n.d.), also give a good overview of the various applications.

in the context of artifical intelligence (see Chapter 5).

strengths and weaknesses of the method

Strengths Explicating patterns of hypothesis formation by means of formal logics has a clear advantage: by reducing patterns to their formal and structural essence, an insight into the pattern's precise conditions and applications is gained that is hard to achieve purely by studying different cases. This is illustrated in Chapter 6, where I devise a logic for the abduction of generalizations, a pattern for which so far no formal characterization has been given.

In this way, the formal explication of patterns can provide the basis of normative guidance in scientific methodology, yet of a sophisticated type. The project of Logical Empiricism, which envisioned the full logical and normative explication of scientific methodology, has clearly faced its limits during the historical turn in the philosophy of science. The present approach of explicating various patterns, is strongly benchmarked on actual historical cases and open to the emergence of new patterns in the history of science. Yet by explication and rational consideration, it can still aspire to provide some normative guidance for scientific practice.

Another great advantage of the formal explication of human reasoning patterns is that it allows for the possibility to provide artificially intelligent agents (which in general lack the human capacity for context awareness unless it is explicitly provided) with formal patterns to simulate human reasoning. In the case of hypothesis formation, this possibility has presently already found applications in the AI subfields of abduction (diagnosis), planning and machine learning (as is also illustrated in Chapter 5).

Weaknesses The method of explicating patterns of hypothesis formation by means of formal adaptive logics also has certain drawbacks, however.

First, formal logics are expressed in terms of a formal language, in which not all elements of human reasoning processes can be represented. This leads inevitably to certain losses. A very obvious example is that in general only propositions can be represented in logics. That means that all observations, figures or other symbolic representations must be reduced to descriptions of them.

A more important example for my project is the implication relation. The adaptive logics framework I use is, certainly for ampliative logics such as those for abduction or induction, largely built around the use of a classical material implication (mostly to keep things sufficiently simple).[12] As a result of this, I have to represent all relations between a hypothesis and the observations that led to their formation (their triggers) by material implications. It is clear that this is a strong reduction of the actual richness of such relations. Hypotheses do not have to imply their triggers: they can also just be correlated with them or be probabilistically likely; or the relation can be much more specific, as in the case of an explanatory or causal relation. Problems of this nature lead me to argue in Chapter 6 for the formal representation of an "explanatory framework" in the language of any logic for abduction. Finally, even if an implication is the suitable description of the relation, it is well-known that the material implication has its limits in describing actual human reasoning.

Related to this is the problem that it is hard to model an agent's intentionality in formal logics. Hypotheses are not always proposed to relate directly to their triggers; agents might have other purposes in mind in suggesting hypotheses.

Second, if one sets out to model actual historical human reasoning processes by means of dynamic logical proofs (as the adaptive framework allows us to do), one quickly finds that it is no easy task to boil down those actual processes to the micro structure of their individual reasoning steps. As human agents often combine individual steps and seldom take note of each individual step, this type of models always contains an aspect of simulation.

Human reasoning also does not proceed linearly step by step as proofs do: it contains circular motions, off-topic deviations and irrational connections that cannot be captured by formal logics. Therefore, models of such reasoning processes are always to a great extent idealized.

Natural languages are also immensely more complex than any formal language can aspire to be. Therefore, models of human reasoning are unavoidably simplifications. Furthermore, as formal logics state everything explicitly, any modeler of human reasoning has to simplify deliberately the actual cases, only to achieve a certain degree of comprehensibility.

Altogether, it is clear that formal models of human reasoning processes

[12] This is an issue relevant beyond the field of adaptive logics. Paul (2000, p. 36) has claimed that most approaches to abduction use a material implication that is implicitly interpreted as some kind of explanatory or causal relation.

are, in fact, only models: they contain abstractions, simulations, simplifications and idealizations. And although these techniques are the key characteristics of models, such as those used in science (see Chapter 9), it is not always easy to evade the criticism that formal logics can only handle toy examples.

Third, certain patterns of creative hypothesis formation, i.e. those that introduce the hypothetical existence of new concepts, cannot be modeled by first-order logics. They require the use of second-order logics, and this is a possibility of which, at present, the adaptive logics framework is not capable.

Fourth, as we are purely concerned with hypothesis formation and not with hypothesis selection, formal methods will generate sets of possible hypotheses that grow exponentially in relation to the growth of the agent's background knowledge. It is clear that this also poses a limit to the application of these methods to real world problems.

Finally, one might question the normativity of this project (and more generally of the adaptive logics program). By aiming to describe actual human reasoning processes, this branch of logics appears to put a descriptive ideal first, which contrasts sharply with the strongly normative ideals in the field of logic in general. The standard answer to this question is that adaptive logics attempt to provide both: on the one hand, they aim to describe actual reasoning patterns; on the other, once these patterns are identified, they aim to prescribe how these patters should be rationally applied. Yet this does not answer how the trade-off between these two goals of description and normativity should be conceived: is it better to have a large set of logics that is able to describe virtually any pattern actually found in human reasoning, or should we keep this set trimmed and qualify most actual human reasoning as failing to accord with the highest normative standards? Therefore, it remains a legitimate criticism that the goals of description and prescription cannot be so easily joined: how their trade-off should be dealt with needs further theoretical underpinning.

overview of my contributions

Various approaches such as inductive logics, abductive logics and belief revision have addressed the formal explication of different patterns of hypothesis formation. My contributions all concern the explication of patterns of abduction. So far, most research in abductive reasoning has focused on singular fact abduction. I have made an effort to extend this analysis to logics for the abduction of generalizations. In addition, I have tried to better tailor the existing work on singular fact abduction to abductive reasoning as it is actually found in scientific practice.

In Chapter 4, I devise the logic $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$ for singular fact abduction that is particularly well suited for scientific reasoning, as it presents a natural way to handle the problem of multiple explanatory hypotheses in science. To illustrate this, a small case study concerning the origin of the moon is included.

In Chapter 5, a set-based approach to the proof theory of adaptive logics is presented. Translation of the syntax of the logic $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$ to this framework allows for a consideration of the problem of abduction as it is conceived in the field of artificial intelligence. The main issue in modeling abduction in AI (which is always conceived as singular fact abduction), is how to handle fast growing knowledge bases, a problem the present logic excels at given its simple strategy and rather low complexity.

In Chapter 6, I present the logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$ for the abduction of generalizations, a pattern of hypothesis formation that has, so far, not received any formal explication. Also, in this chapter the notion of "explanatory framework" is introduced, and it is argued that this notion is a valuable asset for any logic that aims to model abduction.

4 Singular Fact Abduction in Science

The test of a first-rate intelligence is the ability to hold two opposed ideas in the mind at the same time, and still retain the ability to function.
— F. Scott Fitzgerald, *The Crack-up*, 1936

This chapter is based on the paper "Modelling Abduction in Science by means of a Modal Adaptive Logic", published in Foundations of Science *(Gauderis, 2013a). I am indebted to Mathieu Beirlaen, Hans Lycke, Joke Meheus, Bert Leuridan, Peter Verdée, Erik Weber, Dagmar Provijn, Atocha Aliseda and two anonymous referees for their helpful comments on earlier drafts.*

In this paper, a new logic for singular fact abduction in a scientific context is presented. Unlike other logics for singular fact abduction, it deals with the problem of multiple explanatory hypotheses in a natural way. The modeling capability of this logic is illustrated by including a small case study on the origin of the moon.

The content of the original article is retained, except for the addition of section 4.6, which is the result of a recent discussion with Peter Verdée, and I need to thank him for pressing me on this issue. For general consistency with the remainder of this dissertation, small stylistic corrections have been made (including a change in spelling to American English).

4.1 Introduction

The aim of this chapter is to present a new adaptive logic, called $\mathbf{MLA}^{s}_{s}$, that enables us to model abductive reasoning processes. The goal of these processes is to derive possible explanatory hypotheses (*explanantia*) for puzzling phenomena (*explananda*). For that purpose, this logic contains, in addition to deductive inference steps, defeasible reasoning steps based on an argumentation schema known as *Affirming the Consequent* (combined with Universal Instantiation):

$$(\forall\alpha)(A(\alpha)\supset B(\alpha)), B(\beta)/A(\beta)$$

By using this schema I restrict my field of application in two ways. First, I consider abduction only in a *strict* sense, which means that the conditional linking explananda and explanantia must be given. In other words, the modeling of any sort of *creative abduction* – in which conditionals are created – is beyond the scope of this chapter.[1] Second, I opt for a predicate logic. This is so because I use a material implication to model the relation between *explanans* and *explanandum*. As it is well known that $B \vdash_{\mathbf{CL}} A \supset B$, a propositional logic would allow us to derive anything as a hypothesis. In the predicative case, the use of the universal quantifier can avoid this.[2] Moreover, it raises no major problem for modeling real life situations, as the case study illustrates.

Adaptive Logics This logic is constructed by means of the techniques of the adaptive logics program.[3] The reasons why an adaptive logic is fit for this job are threefold.

First, it allows for a direct implementation of defeasible reasoning steps (*in casu* applications of *Affirming the Consequent*). This makes it possible to construct logical proofs that nicely integrate defeasible (in this case ampliative) and deductive inferences. This corresponds to natural reasoning processes.

Second, the formal apparatus of an adaptive logic instructs exactly which formulas would falsify a (defeasible) reasoning step. As these formulas are assumed to be false (so long as one cannot derive them), they are called *abnormalities* in the adaptive logic literature. So, if one or a combination of these abnormalities is derived in a proof, it instructs in a formal way which defeasible steps cannot be maintained. This possibility of defeating previous reasoning steps mirrors nicely the dynamics found in actual human reasoning.

Third, for all adaptive logics in standard format, such as the presented logic $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$, there are generic proofs for most of the important metatheo-

[1] For a more elaborate discussion of creative abduction, see Schurz (2008a, pp. 212-231).

[2] For example, compare $\vdash_{\mathbf{CL}} B(\beta)\supset(A(\beta)\supset B(\beta))$ with $\nvdash_{\mathbf{CL}} B(\beta)\supset(\forall\alpha)(A(\alpha)\supset B(\alpha))$.

[3] The general characteristics of adaptive logics will be explained in the next section. A systematic and thorough overview can be found in Batens (2007).

retical properties (including soundness and completeness).[4]

The Problem of Multiple Explanatory Hypotheses This is not the first attempt to explicate abductive reasoning by means of an adaptive logic and this result draws on earlier suggestions. However, these earlier attempts did not completely deal with the problem of multiple explanatory hypotheses.

To explain this problem, consider the following example. Suppose we have to explain the puzzling fact Pa while our background knowledge contains both $(\forall x)(Qx \supset Px)$ and $(\forall x)(Rx \supset Px)$. There are two ways in which one can proceed. First, we can construct a logic in which we can derive only the disjunction $(Qa \vee Ra)$ and not the individual hypotheses Qa and Ra. This first way, called *practical abduction*[5] and adequately modeled by the logics $\mathbf{LA^r}$ and $\mathbf{LA^r_s}$,[6] is suitable for modeling situations in which one has to *act* on the basis of the conclusions before having the chance to find out which hypothesis actually is the case. A good example is how people react to unexpected behavior. If someone suddenly starts to shout, people will typically react in a hesitant way, taking into account that either they themselves are somehow at fault or that the shouting person is just frustrated or crazy and acting inappropriately.

Second, someone with a theoretical perspective (for instance, a scientist or a detective) is interested in finding out which of the various hypotheses is the actual explanation. Therefore it is important that she can *abduce* the individual hypotheses Qa and Ra in order to examine them further one by one. Although there exist adaptive logics that model this *theoretical* kind of abduction[7], these logics have a quite complex proof theory. This is because, on the one hand, one has to be able to derive Qa and Ra separately, but on the other, one has to prevent the derivation of their conjunction $(Qa \wedge Ra)$, because it seems counterintuitive to take the conjunction of two possible hypotheses as an explanation. Moreover, if the two hypotheses are actually incompatible, it would lead to explosion in a classical context.

Capturing Hypotheses as Logical Possibilities There is actually a more elegant and natural way out of this problem by adding modalities to the

[4] An overview of these can be found in Batens (2007).

[5] According to the definition suggested in Meheus and Batens (2006, pp. 224–225) and used in Lycke (2009).

[6] See Meheus and Batens (2006); Meheus (2007, 2011).

[7] See, for instance, Lycke (2009) and another solution of Lycke (personal communication).

language and deriving the hypotheses $\Diamond Qa$ and $\Diamond Ra$. As $(\Diamond Qa \wedge \Diamond Ra)$ does not imply $\Diamond(Qa \wedge Ra)$ in any standard modal logic, the conjunction problem is automatically solved. This new approach, which I will adopt in what follows, also nicely coincides with the common idea that hypotheses are possibilities. These features make the logic $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$[8] very suitable for the modeling of actual theoretical abductive reasoning processes as the case study will illustrate.

Structure of the chapter In the next section, I will first introduce the main characteristics of an adaptive logic in standard format for readers not familiar with the adaptive logics program. The approach will be general and not limited to abductive contexts. In section 4.3, I provide the groundwork for the logic $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$ by stipulating the deductive framework, i.e. the language schema and the non-defeasible reasoning steps of the logic. The fourth section will introduce in an informal way the defeasible part of the logic $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$ with examples that illustrate how this logic fulfills the different desiderata for modeling abductive reasoning contexts. This informal approach is chosen to give more insight into the functioning of the logic. A formal presentation of the logic $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$ is given in Section 4.5, followed by some philosophical considerations about its consequence set in the next section. Finally, in Section 4.7, I will use this logic to model a more elaborate example taken from the recent history of astronomy.

4.2 General Characterization of Adaptive Logics

Definition An *adaptive logic in standard format* is defined by a triple:

(i) A *lower limit logic* (henceforth **LLL**): a reflexive, transitive, monotonic and compact logic that has a characteristic semantics.[9]

[8] $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$ stands for Modal Logic for Abduction. The subscript **s** is added to indicate that this logic captures the singular fact variant. Like all names of adaptive logics in standard format, the superscript indicates the strategy used (see Section 4.2) – the simple strategy in this case.

[9] Strictly speaking, the standard format for adaptive logics requires that a lower limit logic contains, in addition to the **LLL**-operators, also the operators of **CL** (Classical Logic). However, these operators have merely a technical role (in the generic meta-theory for adaptive logics) and are not used in the applications presented here. Therefore, given the introductory nature of this section, I will not go into further detail. In the logics presented in this dissertation, the condition is implicitly assumed to be satisfied.

(ii) A *set of abnormalities* Ω: a set of **LLL**-*contingent* formulas (formulas that are not theorems of **LLL**) characterized by a logical form, or a union of such sets.

(iii) An adaptive *strategy*.

The lower limit logic **LLL** specifies the stable part of the adaptive logic. Its rules are unconditionally valid in the adaptive logic, and anything that follows from the premises by **LLL** will never be revoked. Apart from that, it is also possible in an adaptive logic to derive defeasible consequences. These are obtained by assuming that the elements of the set of abnormalities are "as much as possible" false. The adaptive strategy is needed to specify "as much as possible". This will become clearer further on.

Dynamic Proof Theory As stated before, a key advantage of adaptive logics is their *dynamic proof theory* which models human reasoning. This dynamics is possible because a *line* in an adaptive proof has – along with a line number, a formula and a justification – a fourth element, i.e. the *condition*. A condition is a finite subset of the set of abnormalities and specifies which abnormalities need to be assumed to be false for the formula on that line to be derivable.

The inference rules in an adaptive logic reduce to three generic rules. Where Γ is the set of premises, Θ a finite subset of the set of abnormalities Ω and $Dab(\Theta)$ the (classical) disjunction of the abnormalities in Θ, and where

$$A \qquad \Delta$$

abbreviates that A occurs in the proof on the condition Δ, the inference rules are given by the generic rules:

PREM If $A \in \Gamma$:

$$\begin{array}{ll} \vdots & \vdots \\ \hline A & \emptyset \end{array}$$

RU If $A_1, \ldots, A_n \vdash_{\mathbf{LLL}} B$:

$$\begin{array}{ll} A_1 & \Delta_1 \\ \vdots & \vdots \\ A_n & \Delta_n \\ \hline B & \Delta_1 \cup \ldots \cup \Delta_n \end{array}$$

RC If $A_1, ..., A_n \vdash_{\mathbf{LLL}} B \vee Dab(\Theta)$

$$\begin{array}{ll} A_1 & \Delta_1 \\ \vdots & \vdots \\ A_n & \Delta_n \\ \hline B & \Delta_1 \cup \ldots \cup \Delta_n \cup \Theta \end{array}$$

The premise rule PREM states that a premise may be introduced at any line of a proof on the empty condition. The unconditional inference rule RU states that, if $A_1, \ldots, A_n \vdash_{\mathbf{LLL}} B$ and $A_1, \ldots, A_n$ occur in the proof on the conditions $\Delta_1, \ldots, \Delta_n$, we may add B on the condition $\Delta_1 \cup \ldots \cup \Delta_n$. The strength of an adaptive logic comes from the third rule, the conditional inference rule RC, which works analogously to RU, but introduces new conditions. So, it allows one to take defeasible steps based on the assumption that the abnormalities are false.[10] Several examples of how these rules are employed in actual proofs can be found in section 4.4.

The only thing we still need is a criterion that defines when we consider a line of the proof to be defeated. At first sight, it seems straightforward to mark[11] lines of which one of the elements of the condition is *unconditionally*[12] derived from the premises. But this strategy, called the *simple strategy*, usually has a serious flaw. If it is possible to derive unconditionally a disjunction of abnormalities $Dab(\Delta)$ that is *minimal*, i.e. if there is no $\Delta' \subset \Delta$ such that $Dab(\Delta')$ can be unconditionally derived, the simple strategy would ignore this information. This is problematic, however, because at least one of the disjuncts of the ignored disjunction has to be true. Therefore, more advanced strategies have been developed. The best-known of these are *reliability* and *minimal abnormality*. We can use the simple strategy only in cases where

$$\Gamma \vdash_{\mathbf{LLL}} Dab(\Delta) \text{ only if there is an } A \in \Delta \text{ such that } \Gamma \vdash_{\mathbf{LLL}} A$$

with $Dab(\Delta)$ any disjunction of abnormalities out of Ω. For adaptive logics in standard format, the first letter of the name of the strategy (simple strategy, reliability or minimal abnormality) is added in superscript to the name of the logic.

[10] The rule also makes clear that any adaptive proof can be transformed into a Fitch-style proof in the **LLL** by writing down for each line the disjunction of the formula and all of the abnormalities in the condition.

[11] Defeated lines in a proof are marked instead of deleted, because, in general, it is possible that they may later become unmarked in an extension of the proof.

[12] *Unconditionally* derived is to be understood as derived on the empty condition.

4.3 The Deductive Framework

Formal Language Schema Let $\mathcal{L}$ be the standard predicative language of classical logic **CL** with logical symbols $\neg, \supset, \wedge, \vee, \equiv, \forall$ and $\exists$. I will further use $\mathcal{C}$, $\mathcal{V}$, $\mathcal{F}$ and $\mathcal{W}$ to refer respectively to the sets of individual constants, individual variables, all (well-formed) formulas of $\mathcal{L}$ and the closed (well-formed) formulas of $\mathcal{L}$.

$\mathcal{L}_M$, the language of the logic, is $\mathcal{L}$ extended with the modal operator $\Box$. $\mathcal{W}_M$, the set of closed formulas of $\mathcal{L}_M$ is the smallest set that satisfies the following conditions:

1. if $A \in \mathcal{W}$, then $A, \Box A \in \mathcal{W}_M$
2. if $A \in \mathcal{W}_M$, then $\neg A \in \mathcal{W}_M$
3. if $A, B \in \mathcal{W}_M$, then $A \wedge B, A \vee B, A \supset B, A \equiv B \in \mathcal{W}_M$

It is important to notice that there are no occurrences of modal operators within the scope of another modal operator or a quantifier.

I further define the set $\mathcal{W}_\Gamma$, the subset of $\mathcal{W}_M$ the elements of which can act as premises in the logic, as:

$$\mathcal{W}_\Gamma = \{\Box A \mid A \in \mathcal{W}\}$$

It is easily seen that $\mathcal{W}_\Gamma \subset \mathcal{W}_M$.

Lower Limit Logic The **LLL** will be the predicative version of **D**, restricted to the language schema $\mathcal{W}_M$. **D** is characterized by a full axiomatization of predicate **CL** together with two axioms, an inference rule and a definition:

K	$\Box(A \supset B) \supset (\Box A \supset \Box B)$
D	$\Box A \supset \neg\Box\neg A$
NEC	if $\vdash A$, then $\vdash \Box A$
$\Diamond_{df}$	$\Diamond A =_{df} \neg\Box\neg A$

This logic is one of the weakest normal modal logics that exist and is obtained by adding the **D**-axiom to the axiomatization of the better-known minimal normal modal logic **K**.

The semantics for this logic can be expressed by a standard possible world Kripke semantics where the accessibility relation R between possible worlds is *serial*, i.e. for every world w in the model, there is at least one world w' in the model such that Rww'.

Intended Interpretation As indicated in the introduction, explanatory hypotheses – the results of abductive inferences – will be represented by formulas of the form $\Diamond A$ $(A \in \mathcal{W})$. I will use formulas of the form $\Box B$ to represent explananda, other observational data and relevant background knowledge. Otherwise, this information would not be able to revoke derived hypotheses.[13] The reason why I choose **D** instead of **K** is that I assume that the explananda and background information are together consistent. This assumption is modeled by the **D**-axiom.[14]

4.4 Informal Presentation of the Logic $\mathbf{MLA}^{\mathrm{s}}_{\mathrm{s}}$

Abductive Contexts and the Set of Abnormalities In specifying the set of abnormalities and the strategy, we have to check whether they allow us to model abductive reasoning according to our expectations.

Apart from the fact that by means of this logic we should be able to derive hypotheses according to the schema of *Affirming the Consequent*, we have to make sure that we cannot derive – as a side effect – random hypotheses which are not related to the explanandum. In addition, it is quite straightforward to demand that a logic for hypothesis formation can handle contradictory hypotheses. Finally, as I pointed out in the introduction, it is a nice feature of adaptive logics that they enable us to integrate defeasible and deductive steps. Therefore, we may require that the logic can handle further predictions (based on earlier derived hypotheses) and evidence for or against them in a natural way.

Since the final form of the abnormalities is quite complex – although the idea behind it is straightforward – I will first consider two more basic proposals that are constitutive for the final form and show why they are insufficient. Obviously, only closed well-formed formulas can be an element of any set of abnormalities. This will not be explicitly stated each time.

[13] For instance, $\neg A$ and $\Diamond A$ are not contradictory, whereas $\Box\neg A$ and $\Diamond A$ are.

[14] For instance, the premise set $\{\Box\neg Pa, \Box(\forall x)Px\}$ is a set modeling an inconsistent set of background knowledge and observations. However, in the logic **K**, this set would not be considered inconsistent, because we cannot derive anything from this set by *Ex Falso Quodlibet*. To be able to do this, we need the **D**-axiom.

First proposal Ω_1 This first proposal is a modal version of the set of abnormalities of the logic $\mathbf{LA}^{\mathbf{r}}_{\mathbf{s}}$.[15] In this and the further definitions, the meta variables A and B represent (well-formed) formulas, α a variable and β a constant of the language $\mathcal{L}$.

$$\begin{aligned} \Omega_1 \quad = \quad & \{\Box((\forall\alpha)(A(\alpha) \supset B(\alpha)) \wedge (B(\beta) \wedge \neg A(\beta))) \mid \\ & \text{No predicate that occurs in } B \text{ occurs in } A\} \end{aligned}$$

This means that a derived hypothesis will be defeated if one shows explicitly that the hypothesis cannot be the case. The second line in the definition is to prevent self-explanatory hypotheses.

Simple Strategy For this logic we can use the *simple strategy*, which means, as stated before, that we have to mark lines for which one of the elements of the condition is unconditionally derived. We can easily see that the condition for use of the simple strategy, i.e.

$$\Gamma \vdash_{\mathbf{LLL}} Dab(\Delta) \text{ only if there is an } A \in \Delta \text{ such that } \Gamma \vdash_{\mathbf{LLL}} A,$$

is fulfilled here. Since all premises have the form $\Box A$, the only option to derive a disjunction of abnormalities would be to apply addition, i.e. to derive $(\Box A \vee \Box B)$ from $\Box A$ (or $\Box B$), because it is well-known that $\Box(A \vee B) \nvdash \Box A \vee \Box B$ in any standard modal logic.[16]

Contradictory hypotheses The following example shows that this logic allows us to derive hypotheses according to the schema *Affirming the Consequent* and is able to handle contradictory hypotheses without causing explosion.

1	$\Box(\forall x)(Px \supset Qx)$	-;PREM	$\emptyset$
2	$\Box(\forall x)(\neg Px \supset Rx)$	-;PREM	$\emptyset$
3	$\Box Qa$	-;PREM	$\emptyset$
4	$\Box Ra$	-;PREM	$\emptyset$
5	$\Diamond Pa$	1,3;RC	$\{\Box((\forall x)(Px \supset Qx) \wedge (Qa \wedge \neg Pa))\}$
6	$\Diamond \neg Pa$	2,4;RC	$\{\Box((\forall x)(\neg Px \supset Rx) \wedge (Ra \wedge \neg\neg Pa))\}$
7	$\Diamond Pa \wedge \Diamond \neg Pa$	5,6;RU	$\{\Box((\forall x)(Px \supset Qx) \wedge (Qa \wedge \neg Pa)),$ $\Box((\forall x)(\neg Px \supset Rx) \wedge (Ra \wedge \neg\neg Pa))\}$

[15] As proposed in Meheus (2011).

[16] It is also possible to derive a disjunction from the premises by means of the **K**-axiom. For instance, $\Box(A \supset B) \vdash \neg\Box A \vee \Box B$, but the first disjunct will always be equivalent to a possibility ($\Diamond\neg A$) and can, hence, not be an abnormality.

$\Diamond Pa$ and $\Diamond\neg Pa$ are both derivable hypotheses because the conditions on lines 5-7 are not unconditionally derivable from the premise set. It is also interesting to note that, because of the properties of the lower limit **D**, it is not possible to derive from these premises that $\Diamond(Pa \wedge \neg Pa)$. The conjunction of two hypotheses is never considered as a hypothesis itself, unless there is further background information that links the two hypotheses in some way.

Predictions and Evidence Suppose I extend the premise set with an additional implication.[17] Then, we can continue the example to see whether the logic can handle further predictions and (counter)evidence for these predictions in a natural way:

8	$\Box(\forall x)(Px \supset Sx)$	-;PREM	$\emptyset$
9	$\Diamond Sa$	5,8;RU	$\{\Box((\forall x)(Px \supset Qx) \wedge (Qa \wedge \neg Pa))\}$

With the extra implication we can derive the prediction $\Diamond Sa$. As long as we have no further information about this prediction (for instance, by observation), it remains a hypothesis derived on the same condition as $\Diamond Pa$. If we would test this prediction, we would have two possibilities. On the one hand, if the prediction turns out to be false, the premise $\Box\neg Sa$ could be added to the premise set. In this case, we can subsequently derive $\Box\neg Pa$, which would falsify the hypothesis $\Diamond Pa$. This is indicated in the proof by marking the now defeated lines with a $\checkmark^i$-sign, where i indicates the line at which the abnormality is derived.

⋮	⋮	⋮	⋮	
5	$\Diamond Pa$	1,3;RC	$\{\Box((\forall x)(Px \supset Qx) \wedge (Qa \wedge \neg Pa))\}$	$\checkmark^{12}$
⋮	⋮	⋮	⋮	
10	$\Box\neg Sa$	PREM	$\emptyset$	
11	$\Box\neg Pa$	8,10;RU	$\emptyset$	
12	$\Box((\forall x)(Px \supset Qx) \wedge (Qa \wedge \neg Pa))$	1,3,11;RU	$\emptyset$	

On the other hand, if the prediction turns out to be true, the premise $\Box Sa$ could be added, but this extension of the premise set would not allow us to derive $\Box Pa$. Since true predictions only *corroborate* the hypothesis

[17] Strictly speaking, this is not what is actually done. What I actually do is start a new proof with another premise set (the extended set). But it is easily seen that I can start this new proof with exactly the same lines as the old proof. This way, it looks as if I extended the old proof. This qualification needs to be considered each time I speak about "adding premises".

and do not *prove* it, while false predictions directly *falsify* the hypothesis, one can say that this logic handles predictions in a *Popperian* way.[18]

Contradictions One way a logic of abduction could generate random hypotheses as a side effect is by allowing for the abduction of contradictions. How this is possible and how the logic prevents this is illustrated in the following example.

1	$\Box Qa$	-;PREM	$\emptyset$	
2	$\Box(\forall x)((Px \wedge \neg Px) \supset Qx)$	-;RU	$\emptyset$	
3	$\Diamond(Pa \wedge \neg Pa)$	1,2;RC	$\{\Box((\forall x)((Px \wedge \neg Px) \supset Qx) \wedge (Qa \wedge \neg(Pa \wedge \neg Pa)))\}$	$\checkmark^4$
4	$\Box((\forall x)((Px \wedge \neg Px) \supset Qx) \wedge (Qa \wedge (\neg Pa \vee Pa)))$	1;RU	$\emptyset$	

Tautologies Still, there are other ways to derive random hypotheses that are not prevented by the first proposal for the set of abnormalities Ω_1. For instance, Ω_1 does not prevent that random hypotheses can be derived from a tautology, as illustrated by the following example. As it is impossible to unconditionally derive the abnormality in the condition of line 3 from the premises, the formula of line 3, the random hypothesis $\Diamond Pa$, remains derived in every possible extension of the proof.

1	$\Box(Qa \vee \neg Qa)$	-;RU	$\emptyset$
2	$\Box(\forall x)(Px \supset (Qx \vee \neg Qx))$	-;RU	$\emptyset$
3	$\Diamond Pa$	1,2;RC	$\{\Box((\forall x)(Px \supset (Qx \vee \neg Qx)) \wedge ((Qa \vee \neg Qa) \wedge \neg Pa))\}$

Therefore, let me adjust the set of abnormalities to obtain the second proposal Ω_2.

Second proposal Ω_2 No hypothesis can be abduced from a tautology if the abnormalities have the following form:

$$\begin{array}{rcl} \Omega_2 & = & \{\Box((\forall\alpha)(A(\alpha) \supset B(\alpha)) \wedge (B(\beta) \wedge \neg A(\beta))) \\ & & \vee\Box(\forall\alpha)B(\alpha) \mid \\ & & \text{No predicate that occurs in } B \text{ occurs in } A\} \end{array}$$

[18] It needs to be remembered that I devised a logic for modeling abduction and handling explanatory hypotheses, not a formal methodology of science. This logic has nothing to say about the confirmation of theories.

It is clear that we can keep using the simple strategy with this new set of abnormalities. It is also easily seen that all of the advantages and examples described above still hold. Each time we can derive an abnormality of Ω_1, we can derive the corresponding abnormality of Ω_2 by a simple application of *addition*. Finally, the problem raised by the tautologies, as illustrated in the previous example, is solved in an elegant way, because the form of the abnormalities makes sure that the abnormality will always be a theorem in case the explanandum is a theorem. So, nothing can be abduced from tautologies.

Most parsimonious *explanantia* Still, there is another way to derive random hypotheses that cannot be prevented by Ω_2. Consider, for instance, the following proof.

1	$\Box Ra$	-;PREM	$\emptyset$
2	$\Box(\forall x)(Px \supset Rx)$	-;PREM	$\emptyset$
3	$\Box(\forall x)((Px \wedge Qx) \supset Rx)$	2;RU	$\emptyset$
4	$\Diamond(Pa \wedge Qa)$	1,3;RC	$\{\Box((\forall x)((Px \wedge Qx) \supset Rx) \wedge (Ra \wedge \neg(Pa \wedge Qa))) \vee \Box(\forall x)Rx\}$
5	$\Diamond Qa$	4;RU	$\{\Box((\forall x)((Px \wedge Qx) \supset Rx) \wedge (Ra \wedge \neg(Pa \wedge Qa))) \vee \Box(\forall x)Rx\}$

The reason why we can derive the random hypothesis $\Diamond Qa$ is the absence of a mechanism to ensure that the abduced hypothesis is the most parsimonious one and not the result of *strengthening the antecedent* of an implication. Before defining the final and actual set of abnormalities that also prevents this way of generating random hypotheses, I have to introduce a new notation to keep things as perspicuous as possible.

Notation Suppose $A_{PCN}(\alpha)$ is the *prenex conjunctive normal* form of $A(\alpha)$. This is the equivalent form of $A(\alpha)$ where all quantifiers are first moved to the front of the expression and where, consequently, the remaining (quantifier-free) expression is written in conjunctive normal form, i.e. as a conjunction of disjunctions of literals.

$$\begin{aligned} &A_{PCN}(\alpha) = (Q_1\gamma_1)\ldots(Q_m\gamma_m)(A_1(\alpha) \wedge \ldots \wedge A_n(\alpha)) \\ &\text{and} \vdash A_{PCN}(\alpha) \equiv A(\alpha) \end{aligned}$$

with $m \geqslant 0, n \geqslant 1, Q_i \in \{\forall, \exists\}$ for $i \leqslant m$, $\gamma_i \in \mathcal{V}$ for $i \leqslant m$, $\alpha \in \mathcal{V}$ and $A_i(\alpha)$ disjunctions of literals in $\mathcal{F}$ for $i \leqslant n$.

Then, I can introduce the new notation $A_i^{-1}(\alpha)$ $(1 \leqslant i \leqslant n)$ so that I have a way to take out one of the conjuncts of a formula in PCN form. In cases where the conjunction consists of only one conjunct (and, obviously, no more parsimonious explanation is possible), the substitution with a random tautology will make sure that the condition for parsimony, added in the next set of abnormalities, is satisfied trivially.

$$\begin{array}{lll} \text{if } n > 1 & : & A_i^{-1}(\alpha) =_{df} (Q_1\gamma_1)\ldots(Q_m\gamma_m)(A_1(\alpha) \wedge \ldots \wedge A_{i-1}(\alpha) \wedge \\ & & \qquad\qquad A_{i+1}(\alpha) \wedge \ldots \wedge A_n(\alpha)) \\ & & \text{with } A_j \ (1 \leqslant j \leqslant n) \text{ the } j^{\text{th}} \text{ conjunct of } A_{PCN}(\alpha) \\ \text{if } n = 1 & : & A_1^{-1}(\alpha) =_{df} \top \\ & & \text{with } \top \text{ any tautology of } \mathbf{CL} \end{array}$$

Final proposal Ω With this notation I can write the logical form of the set of abnormalities Ω of the logic $\mathbf{MLA^s_s}$.

$$\begin{array}{lll} \Omega & = & \{\Box((\forall\alpha)(A(\alpha) \supset B(\alpha)) \wedge (B(\beta) \wedge \neg A(\beta))) \\ & & \vee\Box(\forall\alpha)B(\alpha) \vee \bigvee_{i=1}^{n} \Box(\forall\alpha)(A_i^{-1}(\alpha) \supset B(\alpha)) \mid \\ & & \text{No predicate that occurs in } B \text{ occurs in } A\} \end{array}$$

This form might look complex, but its functioning is quite straightforward. I have actually constructed the disjunction of the three reasons why we should refrain from considering $A(\beta)$ as a good explanatory hypothesis for the phenomenon $B(\beta)$, even if we have $(\forall\alpha)(A(\alpha) \supset B(\alpha))$. The disjunction will make sure that the hypothesis $A(\beta)$ is rejected as soon as one of the following is the case: (i) when $\neg A(\beta)$ is derived, (ii) when $B(\beta)$ is a tautology (and obviously, does not need an explanatory hypothesis) or (iii) when $A(\beta)$ has a redundant part and is therefore not an adequate explanatory hypothesis.

From now on, I will unambiguously shorten this logical form of the abnormalities as

$$!A(\beta) \triangleright B(\beta)$$

which could be read as "$A(\beta)$ is not a valid hypothesis for $B(\beta)$". For the same reasons as stated in the description of Ω_2, we can keep using the simple strategy and all of the advantages and examples described above will still hold.

Example For instance, let's have a look at how the new set of abnormalities solves the previous problem. To make things more clear, the condition will be written out fully for the first time. As such, it is clear that the third disjunct is actually a premise, and that, hence, the abnormality is unconditionally derivable.

1	$\Box Ra$	-;PREM	$\emptyset$	
2	$\Box(\forall x)(Px \supset Rx)$	-;PREM	$\emptyset$	
3	$\Box(\forall x)((Px \wedge Qx) \supset Rx)$	2;RU	$\emptyset$	
4	$\Diamond(Pa \wedge Qa)$	1,3;RC	$\{\Box((\forall x)((Px \wedge Qx) \supset Rx) \wedge (Ra \wedge \neg(Pa \wedge Qa))) \vee \Box(\forall x)Rx \vee \Box(\forall x)(Px \supset Rx) \vee \Box(\forall x)(Qx \supset Rx)\}$	$\checkmark^5$
5	$!(Pa \wedge Qa) \triangleright Ra$	2; RU	$\emptyset$	

4.5 Formal Presentation of the Logic $\mathbf{MLA}^{\mathrm{s}}_{\mathrm{s}}$

I can now present the logic $\mathbf{MLA}^{\mathrm{s}}_{\mathrm{s}}$ in a formally precise way.[19] Like any adaptive logic in standard format, the logic $\mathbf{MLA}^{\mathrm{s}}_{\mathrm{s}}$ is characterized by the triple of a lower limit logic, a set of abnormalities and an adaptive strategy. In this case, the lower limit logic is **D**, the strategy is the simple strategy and the set of abnormalities Ω is, relying on the previously introduced abbreviation, defined by

$$\Omega \quad = \quad \{!A(\beta) \triangleright B(\beta) \mid \text{No predicate that occurs in } B \text{ occurs in } A\}$$

Proof Theory The proof theory is characterized by the three generic inference rules introduced in section 2 and the following definitions.

Within adaptive logics, proofs are considered to be chains of subsequent stages. A *stage of a proof* is a sequence of lines obtained by application of the three generic rules. As such, every proof starts off with the first stage which is an empty sequence. Each time a line is added to the proof by applying one of the inference rules, the proof comes to its next stage, which is the sequence of lines written so far extended with the new line.

Definition 4.1 (Marking for the simple strategy). *Line i with condition Δ is* marked for the simple strategy *at stage s of a proof, if stage s contains a line of which an $A \in \Delta$ is the formula and $\emptyset$ the condition.*

[19] This section is limited to what I need to present this specific logic. For a more general formal presentation of adaptive logics in standard format, see Batens (2007).

Definition 4.2. *A formula A is* derived from Γ at stage s *of a proof if and only if A is the formula of a line that is unmarked at stage s.*

Definition 4.3. *A formula A is* finally derived from Γ at stage s *of a proof if and only if A is derived at line i, line i is not marked at stage s and line i remains unmarked in every extension of the proof.*[20]

Definition 4.4 (Final Derivability). *For $\Gamma \subset \mathcal{W}_\Gamma$: $\Gamma \vdash_{\mathbf{MLA}^s_s} A$ ($A \in Cn_{\mathbf{MLA}^s_s}(\Gamma)$) if and only if A is finally derived in a $\mathbf{MLA}^s_s$-proof from Γ.*

Semantics The semantics of an adaptive logic is obtained by a selection on the models of the lower limit logic. With the simple strategy, for instance, this selection includes only those models that verify the abnormalities that are derivable (by means of the lower limit logic).

Definition 4.5. *A $\mathbf{D}$-model M of the premise set Γ is* simply all right *if and only if $\{A \in \Omega \mid M \vDash A\} = \{A \in \Omega \mid \Gamma \vdash_{\mathbf{D}} A\}$.*

Definition 4.6 (Semantic Consequence). *For $\Gamma \subset \mathcal{W}_\Gamma$: $\Gamma \vDash_{\mathbf{MLA}^s_s} A$ (A* is a semantic consequence of Γ*) if and only if A is verified by all simply all right models of Γ.*

The fact that $\mathbf{MLA}^s_s$ is in standard format warrants that the following theorem holds:[21]

Theorem 4.7 (Soundness and Completeness). *$\Gamma \vdash_{\mathbf{MLA}^s_s} A$ if and only if $\Gamma \vDash_{\mathbf{MLA}^s_s} A$.*

4.6 Modeling Human Reasoning and Consequence Sets

As argued in the general introduction to the logical part of this dissertation, my main goal in employing adaptive logics is to model patterns of hypothesis formation in a formal way. Therefore, my main focus in constructing these adaptive logics is on their proof theory, because, as the examples so far have shown, this allows me to set up proofs that model an actual hypothesis formation process in a step by step fashion.

[20] This definition is slightly different from the more general definition that is used for the other strategies because, using the simple strategy, it is not possible that a marked line becomes unmarked at a later stage of a proof.

[21] An overview of all meta-theoretic properties of adaptive logics in standard format (and their proofs) can be found in Batens (2007).

Yet, as already noted several times, adaptive logics in standard format are also decent logics in the sense that they map any premise set to a unique set of consequences that are *finally derivable* from that premise set, independent of the proofs that are used to obtain these consequences. But if we look in more detail at the final consequence set of the logic $\mathbf{MLA}^s_s$, we see that this set contains far more consequences than one might expect for a logic for abduction. Apart from the explanatory hypotheses and deductive consequences of the premises, this set also contains a large number of possibilities (formulas of the form $\Diamond A$) that are deductive consequences of conditionally inferred hypotheses. In the suggested interpretation of the syntax of the logic, all these formulas should be considered as hypotheses. For instance, if we look at the proof on page 92, in which we illustrated how the logic handles predictions, we see that, if no counter evidence were found, both the original hypothesis $\Diamond Pa$ and its deductive consequence $\Diamond Sa$ would be finally derivable.

This should not, however, be conceived as a problem. The rational attitude to adopt towards these formulas is to entertain them as hypotheses. If we entertain p as a hypothesis and take $p \supset q$ to be true, than none of the other doxastic attitudes of the quadruple 'belief - disbelief - witholding - hypothesizing' (see Chapter 3) would be a suitable attitude to adopt towards q: belief and disbelief are obviously too strong, withholding judgment is clearly too weak. Hence, we should also entertain q as a hypothesis. In other words, the logic $\mathbf{MLA}^s_s$ can certainly be thought of as a logic for hypothesis formation.

For those specifically interested in the actual explanatory hypotheses inferred by means of an abductive reasoning process, it might be suggested that it is always possible to look at specific proofs that model these abductive reasoning steps to observe which hypotheses are initially inferred. Yet one should be warned. As will become clear in the case study of Part III, the explanatory hypothesis that scientists advance in a certain case is seldom the first idea they inferred by an abductive inference. Far more often, the proposed explanatory hypothesis is a consequence of such an initial idea that expresses their suggestion in connection with other parts of their background knowledge to make the explanatory link clear.

A further reason why the inclusion of deductive consequences of initial hypotheses in the consequence set should not be considered a problem is that such consequences have (as we have seen in the example of a prediction) the property that, if they are refuted, the initial hypothesis is also refuted. In a sense, by performing an abductive step, one infers a set of related propositions that all have the same condition. Therefore, if one of them is revoked, all the related hypotheses, which have the same condition in common, will also be revoked. Flach and Kakas (2000a) express this idea by using the notion of an *abductive extension*, i.e. a coherent extension of a classical consequence set with one or more abductive hypotheses and their consequences, such that as many observations of the original premise set as possible are explained.

Although the main purpose of the case study in the next section is to illustrate the modeling capacity of the logic $\mathbf{MLA}^{s}_{s}$, the idea of an abductive extension is also already displayed. In Chapter 5, I will discuss this notion further and show how one can keep track of different abductive extensions.

4.7 Case Study: The Origin of the Moon

In the first decades after NASA was founded in 1958, lunar exploration was one of its most prestigious goals. These efforts have led to the Apollo program that included six lunar landing missions between 1969 and 1972.

> There was widespread expectation that the Apollo exploration of the moon would settle the question of its origin; this had been cited frequently as one of the scientific goals of the Apollo program. (Wood, 1986, p. 18)

As history has taught us, this goal was not achieved. Seen in retrospect, one of the most important reasons for this lack of success was

> [. . .] the concentration on three classical theories of lunar origin: (1) *Capture* – capture of a planetesimal, formed elsewhere in the solar system, into Earth's orbit; (2) *Fission* – spontaneous ejection of upper mantle material into a circumterrestrial swarm due to rotational instability, probably during core formation; (3) *Coaccretion* – formation of the moon by accretion in a circumterrestrial nebula. (Hartmann, 1986, p. 579)

The main reasons[22] why these hypotheses were considered untenable can be summarized as follows. First, capture (H_1) of a planetesimal – according to the laws of celestial mechanics – can occur only if the original

[22] As listed, for instance, in the review article of Wood (1986).

trajectory of this planetesimal is within very limited constraints which include the constraint that this proto-moon should have originated at about the same (radial) distance from the sun and at about the same time as the earth. But if the moon and earth originated at the same time at roughly the same spot in the circumsolar nebula, the moon should have more or less the same chemical composition as the earth. This is not the case, because the moon contains hardly any iron, one of the heavier elements in the solar system that is abundant in the core of the earth. Second, fission (H_2) can neither explain the depletion of volatile elements on the moon's surface (in comparison with the earth's surface) nor account for the abnormally high angular momentum of the Earth-Moon system (in comparison with other planetary systems in our solar system). Finally, coaccretion (H_3) – which was until then the most supported hypothesis – can account neither for iron depletion nor for the high angular momentum.

Coming to this point, several scientists in the mid-seventies were trying to figure out a new hypothesis. Soon, a fourth hypothesis was proposed independently by Hartmann and Davis (1975) and Cameron and Ward (1976). Our attention here will be devoted to the thought process displayed in the latter paper.

Cameron and Ward started by focusing on the angular momentum of the Earth-Moon system.

> A key constraint on the origin of the Earth-Moon system is the abnormally large value of the specific angular momentum of the system, compared to that of the other planets in the solar system. (Cameron and Ward, 1976, p. 120)

Reasoning in terms of the elementary dynamics of physical bodies – in which a collision with another body can lead to an increase in angular momentum – they abduced the following hypothesis.

> This spin was presumably imparted by a collision with a major secondary body in the late stages of accumulation of the earth, with the secondary body adding its mass to the remainder of the proto-earth. (Cameron and Ward, 1976, p. 120)

After determining the characteristics of such a second body – a body roughly the size of Mars, approaching at 11 km/s and hitting the earth off center – to account for the specific angular momentum, they could deductively reason further what would be the consequences of such a giant impact. In short, a lot of volatile elements would vaporize upon shock-unloading and a disk of debris would be caught in the gravitational field of the earth. After a while, the heavier elements (including iron) would sink into the still very fluid young earth, while the lighter elements that remain in an elliptical trajectory around the earth would, over a certain amount of time, form the moon by accretion. Thus, deductively deriving further consequences of this hypothesis, they concluded that "the Moon should thus be deficient in metallic iron and volatile elements..." (p. 121) and, hence, that this hypothesis could at first sight account for all the available data, much of which had previously been problematic.

Before I start to model this case study, it is important to note that we are interested in the process of abduction or the heuristic process of forming explanatory hypotheses, not in confirmation theory or (justificational) inference to the best explanation.[23] What is to be modeled, then, is the reasoning process of scientists looking for a new explanatory hypothesis for the origin of the moon.[24] This is a different reasoning process than confirmation processes, in which one tries to decide whether there is sufficient evidence to support a certain conclusion. This explains the more qualitative nature of arguments in abductive reasoning as opposed to the more quantitative nature of these arguments in justification. I will further use the following notations:

m	"the moon"
Ex	"x exists in its actual state"
Ax	"x is part of a two-body system with unusually high angular momentum"
Fx	"x has an iron (Fe) core"
Vx	"x has a surface containing volatile elements"
Ix	"x is part of a two-body system that is the result of a collision between two proto-bodies"

[23] In discerning abduction and IBE I follow the reasoning initiated by Hintikka (1998) and elaborated by Schurz (2008a,b) that the distinction is to be found in their function and context. Abduction is a strategical or heuristic process, while IBE is a justificational process (see also Sections 1.2 and 1.4 on my interpretation of the notion 'abduction').

[24] New hypotheses can be found by means of both creative and selective abductive processes. As stated in the introduction to this chapter, the logic $\mathbf{MLA}^{s}_{s}$ does not model creative abductions, which would imply that the conditional used by Cameron and Ward would have been created. Instead, the new hypothesis found by Cameron and Ward is obtained by selecting an existing conditional in their background knowledge ("Collisions have an impact on the angular momentum in systems of physical bodies") and using it to abduce a new hypothesis for the origin of the moon.

We can thus model the relevant background knowledge of Cameron and Ward as follows. The domain of objects over which the variables can range is the set of all natural satellites of our solar system.

1	$\Box Em$	-;PREM	$\emptyset$
2	$\Box\neg Fm$	-;PREM	$\emptyset$
3	$\Box\neg Vm$	-;PREM	$\emptyset$
4	$\Box Am$	-;PREM	$\emptyset$
5	$\Box(\forall x)(Ix \supset Ax)$	-;PREM	$\emptyset$

From these premises, they could derive their new hypothesis.

6	$\Diamond Im$	4,5;RC	$\{!Im \blacktriangleright Am\}$

Note that if they would have tried to come up with one of the three older hypotheses by considering one of the three implications $\Box(\forall x)(H_i x \supset Ex)$ (with $1 \leqslant i \leqslant 3$) as an extra premise[25] and abducing the corresponding hypothesis $\Diamond H_i m$, these hypotheses would have been defeated given the premises on lines 2 – 4 and our further background knowledge about these hypotheses $\{\Box(\forall x)(H_1 x \supset Fx), \Box(\forall x)(H_2 x \supset (Vx \wedge \neg Ax)), \Box(\forall x)(H_3 x \supset (Fx \wedge \neg Ax))\}$. But, as we can see in the following extension of the proof, the new hypothesis $\Diamond Im$ actually predicts all the known (and previously problematic) data about the moon.

7	$\Box(\forall x)(Ix \supset Ex)$	-;PREM	$\emptyset$
8	$\Box(\forall x)(Ix \supset \neg Fx)$	-;PREM	$\emptyset$
9	$\Box(\forall x)(Ix \supset \neg Vx)$	-;PREM	$\emptyset$
10	$\Diamond Em$	6,7;RU	$\{!Im \blacktriangleright Am\}$
11	$\Diamond\neg Fm$	6,8;RU	$\{!Im \blacktriangleright Am\}$
12	$\Diamond\neg Vm$	6,9;RU	$\{!Im \blacktriangleright Am\}$

Since the new hypothesis is at first sight corroborated by the known data, Cameron and Ward (and other scientists in the field) could now go on and try to justify or prove that this new hypothesis is the actual explanation for the origin of the moon. This also nicely illustrates that it is not possible to sharply distinguish between the context of discovery and the context of justification.[26] Already in the initial phase of hypothesis formation, a

[25] Although these three hypotheses are not able to explain the origin of the moon, some of them are leading hypotheses for other natural satellites in our solar system. See, for instance, Canup and Ward (2002).

[26] As discussed and argued in Aliseda (2006) and elsewhere.

justificational aspect is present (which I have labeled here "corroboration with the known data").[27]

That this model of their thought process can be assumed to be more or less accurate follows from Cameron and Ward's own reflection upon their thought process, as stated in their conclusions.

> We wish to emphasize that this picture follows as a logical consequence of the process needed to provide the angular momentum of the Earth-Moon system. (p. 121)

This conclusion is correct, but omits their abductive move: only if we take the increased angular momentum to be the result of a collision, all the other characteristics follow as deductive consequences.[28] Their essential consideration was that collisions are a well-known cause of changes in the parameters of dynamic systems.

Together with the independently proposed article by Hartmann and Davis (1975),[29] Cameron and Ward's paper has led to a new successful hypothesis about the origin of the moon, which has since come to be called the "giant impact hypothesis" (Hartmann, 1986). Increased interest in this problem led to the 1984 conference on the origin of the moon in Kona. At the conference it became clear that a "major shift of confidence had occurred among lunar scientists" towards the giant impact hypothesis (Wood, 1986, p. 47). At present, this hypothesis is still the most widely favored among lunar scientists (Belbruno and Gott, 2005, p. 1), although one is still looking for more conclusive evidence by modeling this impact by means of computer simulations.

4.8 Conclusion

In this chapter, I have presented the logic $\mathbf{MLA}^s_s$ that enables us to model the singular fact abductive reasoning processes of scientists. Scientists are in general interested in the actual explanation of the puzzling phenomena

[27] For a thorough discussion on justification in scientific discovery, see Nickles (1980).

[28] As Cameron and Ward are physicists and not trained logicians, I assume that they use the notion of logical consequence in its common layman sense, i.e. as consequences that follow directly and deductively from previous reasoning.

[29] This paper mostly explains that such collisions in the initial stadia of our solar system were not as uncommon as was thought.

they investigate. This means that in the case of multiple explanatory hypotheses, scientists will further investigate the different hypotheses one by one. The logic $\mathbf{MLA}^s_s$ provides this possibility by allowing one to derive – in a defeasible way – the different hypotheses. The logic $\mathbf{MLA}^s_s$ is a decent formal logic in every possible way. Since it is formulated in the standard format of adaptive logics, this logic has a proof theory and a semantics that is sound and complete with respect to it.

While this logic is apt to model actual abductive processes in science – as the case study points out – several extensions can still enrich it. An interesting addition would be that the logic could also handle explananda that contradict the existing background knowledge (anomalies). Another extension that comes to mind is the ability to handle a structured or layered background knowledge. Finally, there is still a lot of work to be done on the heuristics behind abductive reasoning. Can a pattern be discerned in how scientists find relevant conditionals to perform their abductive reasoning steps?

5 Singular Fact Abduction in AI

> *It [the Analytical Engine] might act upon other things besides numbers, were objects found whose mutual fundamental relations could be expressed by those of the abstract science of operations, and which should be also susceptible of adaptations to the action of the operating notation and mechanism of the engine.*
>
> — Ada Byron, Lady Lovelace, 1842

> *"Funny that penguin being there, isn't it? What's it doing there?"*
>
> *"Standing."*
>
> — Monty Python's Flying Circus, 1970

This chapter is based on the article "An Adaptive Logic-based Approach to Abduction in AI" (Gauderis, 2011), published in the proceedings of the 9th International Workshop on Nonmonotonic Reasoning, Action and Change (NRAC 2011), associated with the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011). I am indebted to Peter Verdée, Bert Leuridan and four anonymous referees for their helpful comments on earlier drafts.

In this paper, a set-based formulation of the syntax of adaptive logics in standard format (using the simple strategy) is presented. The translation of the logic $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$ *into this framework will allow us to address the abduction problem as it is conceived in the field of artificial intelligence. The major difference with the problem in philosophical logic is that in AI one is not so much concerned with syntactical proof theory, but rather with how the method handles subsequent stages of theory extension in fast-growing knowledge bases. Therefore, the stress is on how (final) consequence sets of different extensions relate, a problem that is conceived hard with respect to non-monotonic reasoning (e.g. see the review article Paul, 2000). Given the rather low* *complexity of the simple strategy, the results of addressing this problem by means of the logic* $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$ *are promising.*

The scope and content of the original article are to a large extent retained. Yet several paragraphs are adapted or rewritten for general consistency with the remainder of this dissertation and the adaptive logics framework in general. Also, the spelling is changed to American English.

5.1 A Set-based Formulation of Adaptive Logics Syntax

The adaptive logics program is established to offer insight in the direct application of defeasible reasoning steps.[1] This is done by focusing on which formulas would falsify a defeasible reasoning step. In this chapter, I start by reformulating the syntax of this framework in terms of sets. This will allow me to address the problem of abduction as it is perceived and defined in the field of artificial intelligence by means of the existing adaptive logic $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$, which I presented in Chapter 4. The main difference between abduction as it is perceived in philosophical logic and in artificial intelligence is that in AI the stress in not so much on proof theory given a single premise set, but rather on how derived consequence sets can be reused for extensions of the theory.

The presented reformulation of adaptive logics syntax, which reinterprets various elements of its proof theory, is tailored to a specific subclass of adaptive logics, i.e. those in standard format using the simple strategy, but can easily be extended to include other adaptive logics in standard format using other strategies.

Consider a logical theory $\mathcal{T}$ defined in a language $\mathcal{L}$.[2] In my reformulation of adaptive logics syntax, a *formula* is a pair (A, Δ) with A standing for an ordinary well-formed formula in $\mathcal{L}$ and Δ, the *condition* of the formula, standing for a set of ordinary well-formed formulas in $\mathcal{L}$ that are assumed to be false. To express this assumption, elements of the condition Δ can be called *abnormalities*. For each adaptive logic in standard format **AL**, these abnormalities are characterized by a logical form. Also, each such adaptive logic **AL** has a so-called lower limit logic **LLL**, a monotonic logic the consequences of which are always considered to be valid in the adaptive logic. For abductive purposes, this logic is generally Classical Logic **CL** or

[1] For a philosophical defense of the use of adaptive logics, see Batens (2004).

[2] In logic-based approaches to problems in artificial intelligence, such a theory represents the background knowledge of an agent.

a suitable modal extension of **CL**.

The set of *plausibly derivable formulas* $\mathcal{P}$ from a logical theory $\mathcal{T}$ according to an adaptive logic **AL** is defined by the following three rules:

1. *Premise Rule:* if $A \in \mathcal{T}$, then $(A, \emptyset) \in \mathcal{P}$

2. *Unconditional Inference Rule:*
   if $A_1, \ldots, A_n \vdash_{\mathbf{LLL}} B$
   and $\{(A_1, \Delta_1), \ldots, (A_n, \Delta_n)\} \subseteq \mathcal{P}$ for some $\Delta_1, \ldots, \Delta_n$,
   then $(B, \Delta_1 \cup \ldots \cup \Delta_n) \in \mathcal{P}$

3. *Conditional Inference Rule:*
   if $A_1, \ldots, A_n \vdash_{\mathbf{LLL}} B \vee Dab(\Theta)$
   and $\{(A_1, \Delta_1), \ldots, (A_n, \Delta_n)\} \subseteq \mathcal{P}$ for some $\Delta_1, \ldots, \Delta_n$,
   then $(B, \Delta_1 \cup \ldots \cup \Delta_n \cup \Theta) \in \mathcal{P}$

where $Dab(\Theta)$ stands for *disjunction of abnormalities*, i.e. the classical disjunction of all elements in the finite set of abnormalities Θ.

The third rule, which adds new conditions, makes clear how defeasibly derived formulas can be modeled. The idea is that if we can deductively derive the disjunction of a defeasible result B and the formulas the truth of which would make us to withdraw B, we can defeasibly derive B on the assumption that none of these formulas is true.

Apart from the set of plausible formulas $\mathcal{P}$ we need a mechanism that selects which defeasible results should be withdrawn. This is done by defining a marking strategy. In the adaptive logics literature, several strategies have been developed, but for our purposes it is sufficient to consider only the *simple strategy*.[3] Following this strategy, the set of the *derivable formulas* or *consequences* $\mathcal{D} \subseteq \mathcal{P}$ from $\mathcal{T}$ according to **AL** consists of :

1. *Deductive Results:* if $(A, \emptyset) \in \mathcal{P}$, then $(A, \emptyset) \in \mathcal{D}$

2. *Unfalsified Defeasible Results:*
   if $(A, \Theta) \in \mathcal{P}$ (with $\Theta \neq \emptyset$)
   and if for every $\omega \in \Theta : (\omega, \emptyset) \notin \mathcal{P}$,
   then $(A, \Theta) \in \mathcal{D}$

[3] This simple strategy, although computationally simple, should be used with care, because it has no mechanism to block, for instance, formulas with one abnormality in their condition for which that abnormality can be derived as part of a disjunction but not individually. Therefore, when using the simple strategy, it should be prevented in the language or allowed structure of $\mathcal{T}$ that disjunctions of abnormalities can be derived (without at least one of the disjuncts being individually derivable).

So, apart from the deductive results – which are always derivable – an adaptive logic considers all defeasible results as derivable so long as none of the elements of their condition are deductively derivable.

From the definitions of the sets $\mathcal{P}$ and $\mathcal{D}$, we can understand how adaptive logics model the non-monotonic character of defeasible reasoning. If our theory $\mathcal{T}$ is extended to a new theory $\mathcal{T}'$ ($\mathcal{T} \subset \mathcal{T}'$), we can construct the corresponding sets $\mathcal{P}'$ and $\mathcal{D}'$. On the one hand, the set of plausibly derivable formulas will be monotonic ($\mathcal{P} \subset \mathcal{P}'$), since there is no mechanism to withdraw elements from this set and it can only grow larger. On the other hand, we know that the set of derivable formulas is non-monotonic ($\mathcal{D} \not\subset \mathcal{D}'$). It is possible that a condition of a defeasible result in $\mathcal{D}$ is suddenly – in light of the new information in $\mathcal{T}'$ – deductively derivable. So, this result will not be part of $\mathcal{D}'$. Obviously, no deductive result will ever be revoked.

These properties make this kind of logics very apt to model fast growing knowledge bases.[4] If we need a previously defeasibly derived result at a certain point, we cannot be sure whether it is still valid, because there might have been several knowledge base updates in the meantime. But, since the set of plausible formulas is monotonic, we know that this formula will still be in $\mathcal{P}$. So, instead of recalculating the whole non-monotonic set $\mathcal{D}$ after each knowledge base extension (which is the traditional approach), it is sufficient to update the monotonic set $\mathcal{P}$. If we then want to use a defeasible result at a certain stage of knowledge base expansion, we just have to check its condition. It is easily seen that a lot of repetitive recomputation is avoided by this approach, certainly in situations in which we need only a small percentage of the defeasible results at any given stage of knowledge base expansion.

Finally, it has been proven that if the adaptive logic is in standard format (i.e. the abnormalities have a fixed logical form and the lower limit logic **LLL** is sound, complete, reflexive, transitive, monotonic and compact), then the adaptive logic will have many interesting meta-theoretic properties such as soundness, completeness, proof invariance and the fixed-point property.[5]

[4] In this way, this kind of logic can offer a solution to what Paul (2000) noted as one of the main problems both of set cover-based and of some logic-based approaches to abduction.

[5] For an overview of the generic proofs of these properties, see Batens (2007).

5.2 Other Conditional Approaches in AI

So far as I can see, two other approaches to non-monotonic reasoning in AI have used the idea of directly adding conditions or restrictions to formulas. On the one hand, there is a line of research, called “Cumulative Default Reasoning”, going back to a paper of Brewka (1991) with the same title. On the other hand, in the area of argumentation theory, some work on defeasible logic programs (see e.g. García and Simari, 2004) is also based on formulas together with consistency conditions that need to be satisfied to make these formulas acceptable.

The main difference with these research programs is that the abnormalities in adaptive logics are characterized by a fixed logical form. This means that, for instance, the logical form of the logic for abduction in this chapter is the fixed form of abnormalities for any theory or premise set to which we want to apply abductive reasoning. In other words, as soon as any logical theory is given, all abnormalities for that theory and, hence, all plausible and finally derivable abductive results are already determined. In the two other approaches, the conditions of defeasible steps must be specified in the premise set, which gives a complicating element of choice.

5.3 The Problem of Multiple Explanatory Hypotheses

Abduction in AI is the search for explanations for particular observations given a logical theory. Whether a sentence ϕ is considered as an explanation for an observation ω given a theory $\mathcal{T}$ depends on some formal conditions, of which, in general, the following three are considered crucial: (1) ϕ together with $\mathcal{T}$ implies ω; (2) ϕ is logically consistent with $\mathcal{T}$; and (3) ϕ is the most ‘parsimonious’ explanation for w.

The main problem for abduction is that the different defeasible results – the abduced hypotheses – can be mutually exclusive. For instance, if Tweety is a non-flying bird, he may be a penguin or an ostrich. But the formulas $(penguin(Tweety), \Theta_1)$ and $(ostrich(Tweety), \Theta_2)$ are incompatible.[6]

An elegant solution to this problem is found by translating this problem into a modal framework. When we introduce a possibility operator $\Diamond$ to indicate hypotheses and the corresponding necessity operator ($\Box =_{df} \neg\Diamond\neg$)

[6] At this point, we abstract away from the exact conditions; the details of these will be explained in Section 5.4.

to represent background knowledge, we evade this problem. The Tweety-example translates, for instance, as follows (for variables ranging over the domain of all birds):

Background Knowledge:

$$(\Box\forall x(penguin(x) \supset \neg flies(x)), \emptyset)$$
$$(\Box\forall x(ostrich(x) \supset \neg flies(x)), \emptyset)$$
$$(\Box\neg flies(Tweety), \emptyset)$$

Plausible defeasible results:

$$(\Diamond penguin(Tweety), \Theta_1)$$
$$(\Diamond ostrich(Tweety), \Theta_2)$$

So, with this addition the set $\mathcal{D}$ is again consistent. However, in order to have easily accessible reference sets, it is not really necessary to maintain the modal operators explicitly, because we can quite easily make a translation to a hierarchical set-approach by borrowing some ideas of the Kripke semantics for modal logics. It is important, however, to remember that, although we borrow some ideas of Kripke semantics, we are constructing a syntactical representation for abductive extensions of theories and not a semantics for the underlying logic. The notion world (set) will denote a (syntactical) set of formulas, not a semantic concept.

We define the actual world w as the set of all formulas of the knowledge base and all of its deductive consequences. The elements of the set w are the only formulas that have a $\Box$-operator in our modal logic. Subsequently, for every abduced hypothesis we define a new world set that contains it. This world is hierarchically directly beneath the world from which the formula is abduced. This new set contains, further, the formulas of all the world sets hierarchically above, and will be closed under deduction. To make this hierarchy clear, we will use the names $w_1, w_2, \ldots$ for the worlds containing hypotheses directly abduced from the knowledge base, $w_{1.1}$, $w_{1.2}, \ldots, w_{2.1}, \ldots$ for hypotheses abduced from a first-level world, etc. As can be verified, the actual world w is a subset of every other defined world set. In general, for every $w_i : w_i \subset w_{i.j}$ for every $w_{i.j}$ defined hierarchically under w_i. Therefore, it suffices to keep track in the representation of a formula of the highest world set that contains it.[7]

[7] This approach allows for a particular ordinary formula to appear in two world sets that

Using these ideas, our Tweety example can be respresented as follows:

$$\begin{array}{rl}
(\forall x(penguin(x) \supset \neg flies(x)), \emptyset) & w \\
(\forall x(ostrich(x) \supset \neg flies(x)), \emptyset) & w \\
(\neg flies(Tweety), \emptyset) & w \\
(penguin(Tweety), \Theta_1) & w_1 \\
(ostrich(Tweety), \Theta_2) & w_2
\end{array}$$

Since the *hierarchical system of sets w_i* contains all the information of the set $\mathcal{P}$ (the plausibly derivable results) of a modal logic for abduction, the definition of the set $\mathcal{D}$ can be applied to this system of sets too. It is clear that only the deductive consequences of the premises – the only formulas with an empty condition – will be in set w. Further, since all formulas in a world set that are not comprised in the world set hierarchically directly above have the same condition, i.e. the union of the abnormality of the hypothesis for which the world is created and the condition of the world set hierarchically directly above, the definition of $\mathcal{D}$ does not only select at the level of individual formulas, but also at the level of the world sets.[8]

In other words, the definition of $\mathcal{D}$ selects a hierarchical subsystem of the initial hierarchical system of world sets. The different sets in this subsystem are equivalent with what Flach and Kakas (2000a) called *abductive extensions* of some theory (the deductive closure of which is equivalent with the actual world set w). In this way, the logic can handle mutually contradictory hypotheses, without the risk that any set of formulas will turn out to be inconsistent. Different explanations lead to different abductive extensions: one in which Tweety is an ostrich and one in which Tweety is a penguin.

are not hierarchically connected, e.g. w_1 and $w_{2.1}$. This happens when the formula can be derived via two different defeasible ways resulting in different conditions. We will see below that all formulas of a world set that are not comprised in the world set just above it have the same condition. Therefore, as the condition is an inherent part of a formula, they are, strictly speaking, two different formulas. As long as one of them is selected according to the definition of the set $\mathcal{D}$, the ordinary formula can be considered defeasibly derivable.

[8] It is true that each world set also contains all formulas of the world sets hierarchically above. But since these formulas are contained in those worlds above, no information is lost if we allow that $\mathcal{D}$ can select at the level of the world sets.

5.4 A Set-based Formulation of the Logic $\mathbf{MLA}^{\mathrm{s}}_{\mathrm{s}}$

So far, in this chapter we have shown (in Section 5.1) how we can represent the syntax of adaptive logics in terms of the sets $\mathcal{P}$ and $\mathcal{D}$, and (in Section 5.3) how we can cope with contradictory hypotheses by using a hierarchical system of world sets, which is equivalent to the various abductive extensions. In this section we will now use this set representation to reformulate the syntax of the logic $\mathbf{MLA}^{\mathrm{s}}_{\mathrm{s}}$, which has been developed in Gauderis (2013a) (see Chapter 4) and is designed to handle contradictory hypotheses in abduction. The reformulation in terms of sets is performed with the aim of integrating the adaptive approach with other approaches to abduction in AI. The problem of abduction in AI is typically defined in terms of an abductive system (Paul, 2000):

Definition 5.1 (Abductive System). *An* abductive system $\mathcal{T}$ *is a triple* $(\mathcal{H}, \mathcal{O}, d)$ *consisting of the following three sets*

- *a set of* clauses $\mathcal{H}$ *of the form*

  $$\forall\alpha((A_1(\alpha) \wedge \ldots \wedge A_n(\alpha)) \supset B(\alpha))$$

  with $A_1(\alpha), \ldots, A_n(\alpha), B(\alpha)$ *literals and* α *ranging over* d.

- *a set of* observations $\mathcal{O}$ *of the form* $C(\gamma)$ *with* C *a literal and* $\gamma \in d$ *a constant .*

- *a domain* d *of* constants.

All formulas are closed formulas defined over first-order predicate logic.

Furthermore, the notation does not imply that predicates should be of rank 1. Predicates can have any rank; the only preliminaries are that in the clauses all A_i and B share a common variable, and that the observations have at least one variable that is replaced by a constant. Obviously, for predicates of higher rank, extra quantifiers for the other variables need to be added to make sure that all formulas are closed.

Definition 5.2. *The* background knowledge *or* actual world w *of an abductive system* $\mathcal{T} = (\mathcal{H}, \mathcal{O}, d)$ *is the set*

$$w = \{(p,\, \emptyset) \mid \mathcal{H} \cup \mathcal{O} \vdash p\}$$

Since it was the goal of an adaptive logic-approach to implement directly defeasible reasoning steps, we will consider instances of the Peircean schema for abduction (Peirce, 1958, CP 5.189):

> The surprising fact, C, is observed;
> But if A were true, C would be a matter of course,
> Hence, there is reason to suspect that A is true.

When we translate his schema to the elements of $\mathcal{T} = (\mathcal{H}, \mathcal{O}, d)$, we get the following schema:

$$\begin{array}{l} \forall\alpha((A_1(\alpha) \wedge \ldots \wedge A_n(\alpha)) \supset B(\alpha)) \\ \underline{B(\gamma) \qquad\qquad\qquad\qquad\qquad\qquad\qquad} \\ A_1(\gamma) \wedge \ldots \wedge A_n(\gamma) \end{array}$$

To implement this schema – better-known as the logical fallacy *Affirming the Consequent* – in an adaptive logic, we need to specify the logical form of the conditions that would falsify the application of this rule. As we can see from the way in which the conditional inference rule is introduced in the first section, the disjunction of the hypothesis and all defeating conditions needs to be derivable from the theory. To specify these conditions, we will first look at the different *desiderata* for our abductions.[9]

Obviously, it is straightforward that if the negation of the hypothesis can be derived from our background knowledge, the abduction is falsified. If we know that Tweety lives in the wild in the African savannah, we know that he cannot be a penguin. So, in light of this information, the penguin hypothesis can no longer be considered derivable: $(penguin(Tweety), \Theta_1) \notin \mathcal{D}$. But the hypothesis still remains in the monotonic set of 'initially' plausible results: $(penguin(Tweety), \Theta_1) \in \mathcal{P}$.

So, if we define $A(\alpha)$ to denote the full conjunction,

$$A(\alpha) =_{\text{df}} A_1(\alpha) \wedge \ldots \wedge A_n(\alpha)$$

the first defeating condition that could revoke the abductive step is

$$\forall\alpha(A(\alpha) \supset B(\alpha)) \wedge B(\gamma) \wedge \neg A(\gamma).$$

[9] As this chapter is based on a stand-alone article, the following paragraphs unavoidably have some overlap with Chapter 4, yet adjusted to the specific aims of the original paper.

To avoid self-explanations we will further add the condition that $A(\alpha)$ and $B(\alpha)$ share no predicates.

The reason why this condition also states the two premises of the abductive schema is because, in an adaptive logic, we can apply the conditional rule each time the disjunction of a formula and a condition is derivable. So, if we didn't state the two premises in the abnormality, we could derive anything as a hypothesis since $\vdash C(\gamma) \vee \neg C(\gamma)$ for any $C(\gamma)$. But with the current form, only hypotheses for which the two premises are true can be derived. This abnormality would already be sufficient to create an adaptive logic.

Still, we want to add some other defeating conditions. This could be done by replacing the abnormality by a disjunction of the already found condition and the other required defeating conditions. Then, each time one of the defeating conditions is derivable, the whole disjunction is derivable (by addition), and so the formula is defeated.

Often, it is stated that the abduced hypothesis must be as parsimonious as possible. One of the main reasons for this is that one has to avoid random explanations. For instance, have a look at the following example:

$$\begin{aligned}
&\mathcal{H} = \{\forall x(penguin(x) \supset \neg flies(x))\}\\
&\mathcal{O} = \{\neg flies(Tweety)\}\\
&d = \{x \mid x \text{ is a bird}\}
\end{aligned}$$

The following formulas are derivable from this system:

$$\begin{array}{rl}
(\forall x((penguin(x) \wedge is\_green(x)) \supset \neg flies(x)), \emptyset) & w\\
(penguin(Tweety) \wedge is\_green(Tweety), \Theta_1) & w_1\\
(is\_green(Tweety), \Theta_1) & w_1
\end{array}$$

The fact that $Tweety$ is green is not an explanation for the fact that $Tweety$ doesn't fly, nor is it something that follows from our background knowledge. Since we want to avoid that our abductions yield these kinds of random hypotheses, we will add a mechanism to ensure that our hypothesis is the most parsimonious one.

A final condition that we have to add is that our observation is not a tautology. Since we use a material implication, anything could be derived as an explanation for a tautology, because $\vdash C(\gamma) \supset \top$ for any $C(\gamma)$.

Now we can define the defeasible reasoning steps. Therefore we will need a new notation, the purpose of which is to lift out one element from the conjunction $A_1(\alpha) \wedge \ldots \wedge A_n(\alpha)$. This will be used to check for more parsimonious explanations.

Notation 5.3 $(A_i^{-1}(\alpha))$.

$$\begin{array}{lcl} \textit{if } n > 1 & : & A_i^{-1}(\alpha) =_{df} A_1(\alpha) \wedge \ldots \wedge A_{i-1}(\alpha) \wedge A_{i+1}(\alpha) \wedge \ldots \wedge A_n(\alpha) \\ \textit{if } n = 1 & : & A_1^{-1}(\alpha) =_{df} \top \end{array}$$

Definition 5.4. *The* set of abnormalities Ω *of the adaptive logic* $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$ *for an abductive system* T *is given by*

$$\begin{array}{rcl} \Omega & = & \{(\forall\alpha(A(\alpha) \supset B(\alpha)) \wedge B(\gamma) \wedge \neg A(\gamma)) \\ & & \vee\ \ \forall\alpha B(\alpha) \vee \displaystyle\bigvee_{i=1}^{n} \forall\alpha(A_i^{-1}(\alpha) \supset B(\alpha)) \mid \gamma \in d, \\ & & \alpha \textit{ ranging over } d, A_i \textit{ and } B \textit{ literals}, B \notin \{A_i\}\} \end{array}$$

One can verify that the generic conditional rule for adaptive logics (see Section 5.1) for which the abnormalities are characterized by the form above ($\Theta \subseteq \Omega$) is equivalent to the following inference rule, written in the style of the Peircean schema. To keep the computational complexity of testing the condition as low as possible, we can apply a simplifying procedure on Θ that consists in replacing disjunctions by their individual disjuncts and removing conjuncts that are premises. This procedure guarantees that as soon as one of the members of the simplified condition, which will be called Ξ, is unconditionally derivable, a member of the original condition Θ is also unconditionally derivable (by simple applications of addition and conjunction).

Definition 5.5. Defeasible Inference rule for Abduction

$$\begin{array}{ll} (\forall\alpha(A_1(\alpha) \wedge \ldots \wedge A_n(\alpha) \supset B(\alpha)),\ \emptyset) & w \\ (B(\gamma),\ \Xi_i) & w_i \\ \hline (A_1(\gamma) \wedge \ldots \wedge A_n(\gamma),\ \Xi_{i.j}) & w_{i.j} \end{array}$$

with $w_{i.j}$ *a new world set hierarchically directly beneath* w_i *and*

$$\begin{array}{rcl} \Xi_{i.j} & = & \Xi_i \ \cup\ \{\neg A_1(\gamma),\ \ldots,\ \neg A_n(\gamma),\ \forall\alpha B(\alpha),\ \forall\alpha(A_1^{-1}(\alpha) \supset B(\alpha)),\ \ldots, \\ & & \forall\alpha(A_n^{-1}(\alpha) \supset B(\alpha))\} \end{array}$$

So, it is possible to abduce further on hypothetical observations (and, in that way, generate further abductive extensions), but the implications need to be present in the background knowledge w. It is quite obvious, that if the abduced hypothesis is already abduced before (from, for instance, another implication), the resulting world set will contain (partly) the same formulas, but with other conditions.

5.5 An Elaborate Example: a Bird called Tweety

Motivation and comparison with other approaches In this section we will consider an intricate example of the dynamics of this framework. Our main goal will be to illustrate the key advantage of this approach, i.e. that it is no longer needed to recalculate all non-monotonic results at any stage of a growing knowledge base, but only to update the monotonic set of plausible formulas and to check the non-monotonic derivability of the specific formulas needed at that stage.

This is the main difference with other approaches to abduction such as the ones explicated, for instance, in Paul (2000), Flach and Kakas (2000a) or Kakas and Denecker (2002). Since these approaches focus on a fixed and not an expanding knowledge base, they require in cases of expansion a full re-computation to keep the set of derived non-monotonic results up to date. It remains open, however, whether the presented adaptive approach also yields better results for fixed knowledge bases.

Initial system $\mathcal{T}$ Our example will be an abductive learning situation concerning the observation of a non-flying bird, called Tweety. Initially, our abductive system $\mathcal{T} = (\mathcal{H}, \mathcal{O}, d)$ contains in addition to this observation only very limited background knowledge.

$$\mathcal{H} = \{\forall x(penguin(x) \supset \neg flies(x)), \forall x(ostrich(x) \supset \neg flies(x))\}$$
$$\mathcal{O} = \{\neg flies(Tweety)\}$$
$$d = \{x \mid x \text{ is a bird}\}$$

Thus, our background knowledge contains the following formulas:

$$(\forall x(penguin(x) \supset \neg flies(x)), \emptyset) \quad w \tag{5.1}$$
$$(\forall x(ostrich(x) \supset \neg flies(x)), \emptyset) \quad w \tag{5.2}$$
$$(\neg flies(Tweety), \emptyset) \quad w \tag{5.3}$$

The following abductive hypotheses are also part of $\mathcal{P}$:

$$(penguin(Tweety), \Xi_1) \quad w_1 \tag{5.4}$$

$$(ostrich(Tweety), \Xi_2) \quad w_2 \tag{5.5}$$

with the sets Ξ_1 and Ξ_2 defined as

$$\begin{aligned} \Xi_1 &= \{\neg penguin(Tweety), \forall x\ \neg flies(x)\} \\ \Xi_2 &= \{\neg ostrich(Tweety), \forall x\ \neg flies(x)\} \end{aligned}$$

Since both implications have only one conjunct in the antecedent, their parsimony conditions – as defined in the general logical form – trivially coincide with the second condition. Since none of the conditions is deductively derivable in w, both (5.4) and (5.5) are elements of the set of derivable formulas $\mathcal{D}$.

First Extension $\mathcal{T}'$ At this stage, we discover that Tweety can swim, something we know ostriches can't do.

$$\begin{aligned} &\mathcal{H}' = \mathcal{H} \cup \{\forall x(ostrich(x) \supset \neg swims(x))\}, \\ &\mathcal{O}' = \mathcal{O} \cup \{swims(Tweety)\} \\ &d = \{x \mid x \text{ is a bird}\} \end{aligned}$$

From this, the following formulas can be derived:

$$(\forall x(swims(x) \supset \neg ostrich(x)), \emptyset) \quad w \tag{5.6}$$

$$(\neg ostrich(Tweety), \emptyset) \quad w \tag{5.7}$$

As the background information is extended, we know that all previously derived hypotheses are still in the set of plausible hypotheses ($\mathcal{P} \subseteq \mathcal{P}'$). If we now want to check whether these hypotheses are in the new set of derivable hypotheses $\mathcal{D}'$, we need to check whether or not their conditions are derivable from this extended information. But – this has already been cited several times as the key advantage of this system – we don't need to check all hypotheses. Since we do not have any further information on the penguin case and we also do not need that idea at the moment, we can choose to just leave the hypothesis (5.4) for what it is (and save, hence, a computation). At this stage we only want to check whether this new information is a problem for the ostrich hypothesis; and indeed, it is easily seen that (5.5)$\notin \mathcal{D}'$.

Second Extension $\mathcal{T}''$ At this stage, we will further investigate the penguin hypothesis and retrieve additional background information about penguins.

$$\begin{aligned}
&\mathcal{H}'' = \mathcal{H}' \cup \{\forall x(penguin(x) \supset eats\_fish(x)),\\
&\quad \forall x((on\_south\_pole(x) \wedge in\_wild(x)) \supset penguin(x))\}\\
&\mathcal{O}'' = \mathcal{O}'\\
&d = \{x \mid x \text{ is a bird}\}
\end{aligned}$$

The following formulas can now further be retrieved:

$$(eats\_fish(Tweety), \Xi_1) \quad w_1 \tag{5.8}$$

$$(on\_south\_pole(Tweety), \Xi_{1.1}) \quad w_{1.1} \tag{5.9}$$

$$(in\_wild(Tweety), \Xi_{1.1}) \quad w_{1.1} \tag{5.10}$$

with the set $\Xi_{1.1}$ defined as

$$\begin{aligned}
\Xi_{1.1} \quad = \quad & \Xi_1 \cup \{\neg on\_south\_pole(Tweety), \neg in\_wild(Tweety),\\
& \forall x\ penguin(x), \forall x(on\_south\_pole(x) \supset penguin(x)),\\
& \forall x(in\_wild(x) \supset penguin(x))\}
\end{aligned}$$

This stage is added to illustrate the other aspects of adaptive reasoning. First, as (5.8) illustrates, there is no problem in reasoning further with previously deductively derived hypotheses. Only, to reason further, we must first check the condition of these hypotheses. This poses no problem here, as we can easily verify that (5.4)$\in \mathcal{D}''$. The deductively derived formula has the same conditions as the hypothesis on which it is built (and is contained in the same world). So, these results stand so long as the hypotheses on the assumption of which they are derived hold. This characteristic of adaptive logics is very interesting, because it allows one to derive predictions that can be tested in further investigation. In this example, we can test whether Tweety eats fish. In case this experiment fails and $\neg eats\_fish(Tweety)$ is added to the observations in the next extension of the theory, the hypothesis (and all results derived on its assumption) will be falsified.

Second, the set of conditions $\Xi_{1.1}$ for the formulas (5.9) and (5.10) contains now also conditions that check for parsimony. Let us illustrate their functioning with a final extension.

Third Extension $\mathcal{T}'''$ At this stage, we learn that even in captivity the only birds that can survive at the South Pole are penguins. In addition to that, we get to know that Tweety is held in captivity.

$$\begin{aligned}
&\mathcal{H}''' = \mathcal{H}'' \cup \{\forall x(on\_south\_pole(x) \supset penguin(x))\},\\
&\mathcal{O}''' = \mathcal{O}'' \cup \{\neg in\_wild(Tweety)\}\\
&d = \{x \mid x \text{ is a bird}\}
\end{aligned}$$

If we now check the parsimony conditions of $\Xi_{1.1}$, we see that an element of this condition can be derived from our background knowledge. Hence, none of the formulas assigned to world $w_{1.1}$ are any longer derivable on this condition. Yet one might wonder whether this parsimony condition should not let us keep (5.9) and only withdraw (5.10). That this is not a good way forward is proven by the fact that in that case (5.9) would still be falsified, because $\Xi_{1.1}$ also contains $\neg in\_wild(Tweety)$, our new observation. In fact, we do not need the world $w_{1.1}$ to maintain the South Pole hypothesis of (5.9), as it can now be derived from $\mathcal{H}'''$ in another world, which has no conditions on whether or not Tweety lives in the wild.

$$(on\_south\_pole(Tweety), \Xi_{1.2}) \quad w_{1.2} \tag{5.11}$$

with the set $\Xi_{1.2}$ defined as

$$\Xi_{1.2} \quad = \quad \Xi_1 \cup \{\neg on\_south\_pole(Tweety), \forall x\, penguin(x)\}$$

So, at the end, we find that the set $\mathcal{D}'''$ of derivable formulas consists of all formulas derivable in the world $w_{1.2}$ (which is an abductive extension of the worlds w and w_1). The formulas of w_2 and $w_{1.1}$ are not an element of the final set of derivable results $\mathcal{D}'''$.

5.6 Conclusion

In this chapter I have presented a new logic-based approach to the problem of abduction in AI, which is based on the adaptive logics program. The main advantages of this approach are :

1. Each abduced formula is presented together with the specific conditions that would defeat it. In that way, it is not necessary to check the whole system for consistency after each extension of the background knowledge. Only the formulas needed at a certain stage need to be checked. Furthermore, it allows for the conditions to contain additional requirements, such as parsimony.

2. In comparison with other approaches that add conditions to formulas, the conditions are here fixed by a logical form and hence only determined by the (classical) premise set. In this way, there is no element of choice in stating conditions (as, for instance, in default logics).

3. By integrating a hierarchical system of sets, it provides an intuitive representation of multiple hypotheses without causing conflicts between contradictory hypotheses.

4. It allows for further deductive and abductive reasoning on previously retrieved abduced hypotheses.

5. The approach is based on a proper, sound and complete fixed point logic ($\mathbf{MLA}^{s}_{s}$).

Limitations and Future Research It has been argued that these advantages make this approach apt for systems in which not all non-monotonically derivable results are needed at every stage of expansion of a knowledge base. Still, it needs to be examined whether an integration with existing systems (for a fixed knowledge base) does not yield better results. Furthermore, since the key feature of this approach is that it saves computations in expanding knowledge bases, it needs to be investigated whether there is any possibility of integration with assumption-based Truth Maintenance Systems (building on the ideas of Reiter and de Kleer, 1987).

6 Abduction of Generalizations

Though it be too obvious to escape observation, that different ideas are connected together; I do not find that any philosopher has attempted to enumerate or class all the principles of association; a subject, however, that seems worthy of curiosity.

— David Hume, *An Enquiry Concerning Human Understanding*, 1748

This chapter is based on the paper "Abduction of Generalizations", co-authored by Frederik Van De Putte and published in Theoria *(Gauderis and Van De Putte, 2012). We are indebted to Laszlo Kosolosky, Dagmar Provijn, Bert Leuridan, Peter Verdée, Joke Meheus and two anonymous referees for their helpful comments on earlier drafts.*

In this paper, a logic for the abduction of a generalization is presented, a pattern which has, so far, not been modeled in terms of a formal logic. Furthermore, the notion of explanatory framework *is introduced, which is a valuable asset for any logic that aspires to model abductive patterns.*

The content of the original article is largely retained. In order to avoid repetition in the formal presentation of adaptive logics, elements presented already in previous chapters have been removed from Section 6.4. Further, small stylistic corrections have been made for general consistency with the remainder of this dissertation.

Recently, Mathieu Beirlaen has found a problem for some adaptive logics for abduction, which affects the logic in this paper. The problem and some suggestions to solve this problem are presented in a new section (6.5).

6.1 Introduction

Abduction is generally defined as "the process of forming an explanatory hypothesis" (Peirce, 1998, p. 216). In this chapter we will focus on a specific "pattern of abduction" (to use a phrase introduced by Schurz (2008a)). Consider the following example (Schurz, 2008a, p. 212):

(P1) Pineapples taste sweet.
(P1) Everything that contains sugar, tastes sweet.

(C) Pineapples contain sugar.

Schurz called this type of inference "law abduction". The name "rule abduction" has also been used for a similar pattern (Thagard, 1988). But, as 'law' and 'rule' are heavily debated concepts in the philosophy of science and in philosophy in general, we will use the more neutral term *abduction of a generalization* (henceforth AG) for this specific pattern. More examples and a general characterization of AG will be presented in Section 6.2. It will be argued that this pattern is ubiquitous in both everyday and scientific reasoning, and is commonly recognized as a useful – though also fallible – means of extending one's knowledge.

Notwithstanding the importance of AG, little effort has been made so far to study the characteristics of this inference pattern or to explicate it by means of a formal logic. As will be explained in Section 6.2.2, most scholars in AI and formal logic have focused on singular fact abduction, whereas philosophers of science have taken a more general, but informal point of view on abduction. It is our aim to treat AG as a distinct subject matter to see how one may understand and formalize it.

Outline A first analysis of AG is provided in Section 6.2. We describe this pattern informally, showing that it is a widespread inference pattern; secondly, we explain why it has been neglected in formal logic and philosophy of science; finally, we argue for the specific importance of AG in scientific contexts.

In Section 6.3, we turn our focus to the problems that emerge when representing AG formally. We argue that a distinction in the object language is needed between what we call *mere generalizations* and the *explanatory framework* for any logic that models AG; and, moreover, that this distinction is useful in any logic for abduction. In general, as AG is a non-monotonic inference form, we also discuss how the dynamic features can be represented.

Finally, in Section 6.4, the logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$ is presented. This is a logic for the abduction of generalizations (symbolized by the $\forall$ subscript), formulated in the standard format of adaptive logics, using the *reliability* strategy (symbolized by the **r** superscript). After arguing why this framework is well-suited for the current application, we will illustrate the proof theory of $\mathbf{LA}^{\mathbf{r}}_{\forall}$, which allows us to model the dynamic interaction of AG and classical inferences.

Preliminaries Let $\mathcal{L}$ be the standard language of classical first-order predicate logic, obtained from a set of constants $\mathcal{C} = \{a, b, c, \ldots\}$, a set of variables $\mathcal{V} = \{x, y, z, \ldots\}$, a set of predicates $\mathcal{P} = \{P, Q, R, \ldots\}$, the connectives $\neg, \vee, \wedge, \supset, \equiv$ and quantifiers $\forall, \exists$. $\mathcal{W}$ is the set of formulas in $\mathcal{L}$. Depending on the context, A, B, C are used either as metavariables for members of $\mathcal{W}$, or for (conglomerates of) predicates, e.g. $(P \wedge Q) \vee \neg R$. The metavariables $\alpha, \beta, \ldots$ refer to constants and variables.

6.2 Abduction of a Generalization

6.2.1 The phenomenon

We define *abduction of a generalization* (AG) as every inference that fits the following pattern:

> It has been observed that all A are B.[1]
> Also, being C is regarded as an explanation for being B.
> Therefore, the hypothesis that all A are C is raised.

Hence, by AG we generate hypotheses that explain why all observed objects of a certain class have a specific property. In Section 6.3, we will explain how this definition can be operationalized in a first-order modal language. But first, let us point out some general characteristics of AG.

First of all, consider the classical definition of abduction by Peirce (1958, 5.189):

[1] Strictly speaking, this is shorthand for "All observed A are B, and therefore it is believed that all A are B." In essence, this pattern contains an instance of the pattern *Inductive Generalization*. This is important, because, although we will model this premise as a generalization, it cannot be forgotten that abduction always starts from observed cases.

> The surprising fact, X, is observed;
> But if Y were true, X would be a matter of course,
> Hence, there is reason to suspect that Y is true.[2]

Note that AG does not entirely fit this definition. In AG we do not seek an explanation for a certain observation, but for a *generalization* based on a series of such observations. However, if one is willing to accept this natural extension of the concept, we can make AG fit the above schema nicely. Both the surprising fact X and the hypothesis Y are generalizations, respectively "all A are B", and "all A are C". The second line of Peirce's schema follows deductively if "being C" implies "being B".[3]

This leads to another important consideration about the *Peircean* or *classical* notion of abduction: it is defined in a *deterministic* way, i.e. the truth of Y implies the truth of X. Although we do not suggest that this notion of abduction cannot be meaningfully extended to other accounts in which the motivation to adopt the abductive hypothesis is, for instance, probabilistic ($P(X|Y)$ is high) or comparative ($P(X|Y) > P(X|\neg Y)$), we restrict ourselves in this chapter to the classical case, as does most of the literature on abduction. As it is also assumed that Y explains X,[4] this restriction will have consequences for the formalization of AG in Section 6.3.1.

Second, AG is distinct from what is called *singular fact abduction*, in which both the surprising fact and the hypothesis are singular facts (see Chapter 4). In a first-order language, both the *explanandum* and explanatory hypothesis of a singular fact abduction are modeled as objects having a certain property (such as Pa). In contrast, in AG they will be modeled by generalizations (such as $\forall x(Px \supset Qx)$). Existing models for abduction usually limit themselves to singular fact abduction, as we will see in the next section.

Third, AG is not a novel reasoning pattern. It has been known at least since Aristotle, who treats something similar in his *Posterior Analytica* when he considers the "middle term" of a definition. This pattern is, in his view, the essence of a good definition: it should not only say what the *definien-*

[2] To avoid confusion with our definition of AG, the schematic letters A and C originally used by Peirce are replaced by X and Y.

[3] This may be easier to grasp when spelled out in first-order predicate logic: we have that $\forall x(Cx \supset Bx) \vdash_{\mathbf{CL}} \forall x(Ax \supset Cx) \supset \forall x(Ax \supset Bx)$.

[4] Applying the above schema as such is justified only in case of abduction, i.e. the formation of explanatory hypotheses. If Y does not explain X, flagrant examples of the logical fallacy *affirming the consequent* that have little value qua hypothesis will be obtained.

dum (A) is, it should also be an explanation (C) for its observed properties (B). As an example, he explains why horned animals (A) lack upper incisors (B) by defining horned animals as a subclass of animals that have inflected hard material from their mouth to their heads (C). According to Aristotle, this is a good definition of a class because it explains the properties of that class.[5] However, the reasoning pattern we are considering is much broader than what Aristotle had in mind. A, B and C can be any properties, and neither should A be a definiendum, nor C a definiens. Furthermore, as explained in Eco (1983), Aristotle's desire to have a strict taxonomy of definitions confronts him with the fact that different properties need different explanations. As we do not presuppose a taxonomical or hierarchical structure connecting the three predicates, we naturally consider different explananda for different properties.

Fourth, AG is frequently applied in human reasoning, often in combination with or following an instance of singular fact abduction. For instance, people do not only wonder why their heads hurt (they drank too much last night) or why there is a thunderstorm (it was very hot during the day). Not much of a reflective mind is needed to start also asking questions such as why it is that every time one drinks a bit too much, one suffers from headaches, or why thunderstorms often follow hot days. In other words, people wonder not only why certain facts are the case, but also why certain regularities occur in their environment.

6.2.2 The Lack of Formal Characterizations of AG

The lack of models for AG will be explained by pointing out how the application of the concept of abduction in a variety of fields has caused a growing divergence in definitions and interpretations. This will also clarify the relation between our current project and the literature on abduction.

Broadly speaking, two main currents in research on abduction can be discerned. On the one hand, research in AI and formal logic mostly focuses on a *syllogistic* interpretation of Peirce's work, in which abduction

[5] See Aristotle's *Posterior Analytica* (n.d.), section II.10 for his distinction between two types of definitions and sections II.12-14 for his view on the role of the middle term in a definition. A good treatment of the analogy between Aristotelian definitions and Peircean abduction can be found in Eco (1983). In our opinion, Schurz (2008a) refers to the wrong concept when he links AG (in his words: law abduction) to Aristotle. The concept "hitting upon the middle term" is only employed in the definition of quick wit (Aristotle, n.d., I.34), in which it is illustrated with an example of a singular fact abduction. In our view, a predecessor of AG can be found only in Aristotle's treatment of the role of the middle term in definitions.

is introduced as part of a triad that is clarified with the following famous beans-example of Peirce (1958, CP 2.623):

> All the beans from this bag are white. (Rule)
> These beans are from this bag. (Case)
> These beans are white. (Result)

All reasoning deriving a result from a case and a rule is called *deductive*, all reasoning deriving a rule from a case and a result *inductive*, and all reasoning deriving a case from a rule and a result *abductive*. Of these three, only deductive reasoning is analytic and infallible; abduction and induction are called by Peirce synthetic or ampliative (Peirce, 1958, CP 2.623).

Having this schema in mind, researchers in AI or formal logic generally focus on instances of singular fact abduction, which are variations on the following pattern:

$$B(\alpha), \forall\beta(A(\beta) \rightarrow B(\beta))/A(\alpha)$$

This pattern is usually combined with the condition that the hypothesis should be explanatory. Aliseda (2006) adds a further condition suggested by Peirce, i.e. that the observed fact should be surprising (in the sense that $B\alpha$ cannot be derived from the background theory alone).

One noteable exception to the exclusive focus on singular fact abduction is Thagard (1988). He obtains a pattern similar to AG, which he calls "rule abduction", by adding to his logic program PI the ability to generalize the results of singular fact abductions. Although his model does not abduce *from* generalizations, it has the same goal as an AG, i.e. to derive an explanation for why all elements of a given class share a certain property.

On the other hand, research in philosophy of science usually starts from a *methodological* interpretation of Peirce. In his later writings Peirce distinguishes abduction, induction and deduction as different steps in a methodology of science (Peirce, 1998, pp. 212-218). Abduction is the process of forming an explanatory hypothesis, from which deduction can draw predictions, which then can be tested by induction.[6] Research in this tradition (see e.g. Magnani 2009) considers abduction as a very broad concept including analogical reasoning, visual abduction, common cause reasoning,

[6] It is generally acknowledged (see e.g. Flach and Kakas 2000a, pp. 5-8) that both interpretations can be found in Peirce's work, although they are not fully compatible. They represent an evolution in his thinking, as he hinted himself when he remarked that he "was too much taken up in considering syllogistic forms" (Peirce, 1958, 2.102). See also Sections 1.2 and 1.4.

etc. Here, Peirce's definition of abduction (see Section 6.2.1) is seen as an expression in the metalanguage, in which the term 'fact' can refer to any proposition. Some (see e.g. Harman 1965; Lipton 2004; Douven 2011) have tried to capture this concept of abduction under the single schema of *inference to the best explanation* (IBE).[7] However, such attempts to reduce the broadness of the considered concept prevent the discovery of interesting features of more specific patterns of abduction. Even more, it is not exactly clear whether IBE and abduction refer to the same thing in the process of discovery. Schurz explains this as follows (2008a, p. 205):

> The majority of the recent literature on abduction has aimed at *one most general* schema of abduction (for example IBE) which matches every particular case. I do not think that good heuristic rules for generating explanatory hypotheses can be found along this route, because these rules are dependent of the specific type of abduction scenario.

In this article, Schurz subsequently presents a taxonomy of distinct patterns of abduction. Having this in mind, we think it is best to remain pluralistic on the logical form of abduction. We should maintain the rich concept of abduction as it is understood in the philosophy of science, but, in order to provide the formal rigor which is characteristic of the logic and AI community, we should focus separately on each of the different specific forms of abduction.

6.2.3 The Ubiquity of AG in Scientific Practice

At the end of Section 6.2.1, we mentioned several examples in which abduction of a generalization is triggered by a question concerning the result of a singular fact abduction. This question arises from the need for a deeper understanding of the observed relations. Even in these simple examples, the hypotheses resulting from the second abduction often have a more scientific outlook. One is not satisfied with the information that somebody has put this particular banana in the fridge as an explanation for the fact that it has a dark brown color. One wants to understand why bananas change color when they are put in cold places.

[7] These scholars consider Peirce's remark that abduction should be as economical as possible (Peirce, 1958, 7.220) to be an essential and crucial condition.

We can recognize this curious spirit in the endeavors of many scientists. For instance, Descartes was not satisfied with the folk explanation of the rainbow, i.e. that a rainbow appears because the sun breaks through shortly after a rain shower. He wanted to understand *why* rainbows appear whenever the sun shines while it rains. We will argue that AG is at least as important in scientific practice as singular fact abduction by considering two general characteristics of this practice.[8]

First, in scientific practice one attempts to formulate theories, which have both a *universal* and *falsifiable* nature.[9] One does not merely want an explanation why, for instance, this particular person suffers from this disease. One wants to understand why and how this disease is transmitted in general. Formulating theories about particularities is seldom considered good scientific practice, and such theories are then also often labeled as *ad hoc*. Theories are thus mainly formulated for a whole class of objects and, by consequence, formulated in terms of generalizations. These generalizations allow us to derive singular fact predictions by means of which theories can be tested. Therefore, in the formation process of such theories, reasoning methods resulting in generalizations, such as inductive generalization or abduction of a generalization, are essential.

Second, augmented *unification* (as characterized, for instance, by Kitcher, 1993) is generally seen as an indicator of scientific progress.[10] Each application of AG is in essence a unification step, because it explains an observational generalization, e.g. "All A are B", by characterizing its antecedent (A) as a subclass of a more general class (C) for which the observed properties (B) hold. Therefore, AG is a key method in enhancing unification in scientific practice. The most interesting examples in the history of science can be found when a new theory is proposed as a solution for some anomalies of an existing theory. In that case, the proponents of the new theory also need to show that most of the already known and well-tested observational laws, which are explained by the old theory, can be explained by the new theory. For instance, Newton could explain Huygens' pendulum law using his general laws of motion by pointing out how

[8] This claim is about the *scientific practice* and not about *scientific explanation*. In scientific explanation, a scientific theory is employed to explain a certain fact (either a singular fact or a generalization). Scientific practice is the activity of forming such scientific theories and expanding current scientific knowledge.

[9] Universality should not be taken as an absolute notion, but as an achievable level of generality relative to the methods and scope of the specific field.

[10] Both the instrumentalist and realist views concerning the nature of scientific progress seem to agree on this point (Niiniluoto, 2011).

the different parameters of the pendulum law could be translated into his general mathematical framework. In the same way, Bohr could explain by means of his atomic model why the wavelengths of the visible emission spectrum of hydrogen can be calculated by the Balmer formula.[11]

6.3 Introducing the Formal Framework

6.3.1 The Explanatory Framework

The pattern presented in the definition of AG (on p. 123) could be formally explicated as follows:[12]

$$\begin{array}{ll} \text{(P1)} & \forall\alpha(A(\alpha) \supset B(\alpha)) \\ \text{(P2)} & \forall\alpha(C(\alpha) \supset B(\alpha)) \\ \hline \text{(H)} & \forall\alpha(A(\alpha) \supset C(\alpha)) \end{array}$$

However, we must be careful here: the definition stipulates that C-hood *explains* B-hood, not just that everything that has the property C also has the property B. In other words, where (P1) and (H) can be of any kind, the set of possible candidates for (P2) is restricted.[13] We call this set the *explanatory framework*. It consists of all generalizations of the form $\forall\alpha(F(\alpha) \supset G(\alpha))$ where being F provides an explanation for being G. Whether or not a generalization belongs to the explanatory framework, may depend on the phenomenon we are trying to explain. In other words, it is contextually defined. All we assume is that it is clear for each generalization, given the abductive problem at hand, whether it is a member of the explanatory framework or not. In the latter case we call it a *mere generalization*.

With this new terminology, we are now able to characterize all the lines of the above schema: (P1) is the *explanandum*, i.e. the mere generalization that is to be explained; (P2) is a generalization that is part of the explanatory framework for the current abductive context; (H) is the *explanatory hypothesis*. An *explanation* or *explanans* for (P1) consists of an explanatory

[11] A philosophical introduction to the circumstances of these two major milestones in science can be found in Smith (2008) and Faye (2008).

[12] This first representation makes use of the fact that we restrict ourselves to the deterministic case.

[13] In our opinion, Schurz (2008a) puts too little emphasis on this point in his discussion of AG, or "law abduction" as he calls it. In his schema, (P2) is called a "background law", but as far as we see, no explicit definition or circumscription is provided.

hypothesis together with one or more elements of the explanatory framework that connect the hypothesis to the explanandum.

Now what does it actually mean that F-hood explains G-hood? Needless to say, the philosophical literature abounds in theories of explanation. However, as we have here restricted ourselves to classical abduction (see Section 6.2.1), certain preconditions apply. First, if F-hood explains G-hood, then F-hood should also imply G-hood. Second, as abduction is an inference, only argumentative accounts of explanation are relevant. Hence, the choices to explicate the notion 'explanation' in the definition of the *explanatory framework of a (classical) abductive problem* are limited to accounts of explanation that have the structure of a deductive argument such as a DN-argument (e.g. Hempel 1965), a causal argument (e.g. Hausman 1998) or an augmented unification argument (e.g. Kitcher 1993).[14]

In any of these accounts, (P2) has a specific status: it must be lawlike, refer to an underlying causal mechanism, or be a more general argumentation scheme. We use the more abstract term *explanatory framework* to express this status of (P2). This specific status turns AG into a fundamentally *asymmetric* inference. It is not possible to derive $\forall\alpha(C(\alpha) \supset A(\alpha))$ from the same premises, since A-hood does not explain B-hood. Hence, if a logic explicates AG, it should be able to represent this asymmetry between (P1) and (P2) in its object language.

Before we explain how this can be done, let us briefly give an extra reason to motivate the distinction between the explanatory framework and mere generalizations as a valuable asset for any logic that models abductive processes in general. Mere generalizations are often used in abductions that involve knowledge about methods or procedures. For an example in singular fact abduction, consider the following premises:

P_1 The Geiger counter produces audible clicks near the object a, but turns back to silence if an aluminum plate is brought between a and the Geiger counter.

P_2 If the Geiger counter produces audible clicks near an object, but turns back to silence if an aluminum plate is brought between the object and the Geiger counter, the object emits β radiation.

P_3 If an object contains ${}^{14}_{6}C$, the object emits β-radiation.

[14] It is not implied that there are no other valuable accounts of explanation. We only claim that (classical) abductive hypotheses (the only ones that are of our concern here) are part of a deductive argument that forms an explanation for the explanandum.

Without the distinction between the explanatory framework and mere generalizations, a logic for singular fact abduction treats P_2 and P_3 as having the same formal structure (a mere generalization). But a physicist interested in explaining the emitted β radiation from a is interested only in the hypothesis suggested by P_3, as the behavior of the Geiger counter provides no explanation. On the other hand, the mere generalization P_2 is needed to derive the fact that a emits β-radiation in the first place (as the only thing the physicist can observe is P_1). Hence, P_2 cannot be omitted from this abductive reasoning context. Only a logic that is able to represent explanatory frameworks can handle this case properly.

6.3.2 A Modal Approach

In Section 6.4, we will present the logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$. This system is a non-monotonic extension of the well-known modal logic **T**, and allows us to model instances of AG in the modal language $\mathcal{L}^{\Box}$. We will first define $\mathcal{L}^{\Box}$ and **T** formally, after which we will offer some comments concerning our choice of these two.

Let $\mathcal{L}^{\Box}$ denote the extension of $\mathcal{L}$ with the modal necessity operator $\Box$. The set of formulas $\mathcal{W}^{\Box}$ is the smallest set for which the following holds:

For all $A \in \mathcal{W}$: $\quad A, \Box A \in \mathcal{W}^{\Box}$
For all $A, B \in \mathcal{W}^{\Box}$: $\quad \neg A, A \vee B, A \wedge B, A \supset B, A \equiv B \in \mathcal{W}^{\Box}$

Note that by this definition, we exclude the occurrence of boxes within the scope of quantifiers, of iterations of boxes and, more generally, of nested boxes.[15] For instance, $\Box\forall x(Px \supset Qx)$ and $Pa \vee \Box\exists x(\neg Rx)$ are members of $\mathcal{W}^{\Box}$, whereas $\Box\Box\forall xPx$ and $\forall x\Box Px$ are not.

An axiomatization for the predicate version of **T** over the language $\mathcal{W}^{\Box}$ is obtained by taking the axioms of classical first-order predicate logic (henceforth **CL**) and adding the following axioms (closed under *modus ponens*):

K $\quad \Box(A \supset B) \supset (\Box A \supset \Box B)$

[15] It might be possible to do without these restrictions on the language, given a number of additional axioms such as the 4-axiom ($\Box A \equiv \Box\Box A$), the Barcan formula and/or the inverse Barcan formula. This would however severely complicate the logical apparatus, whereas the extended language would contain several expressions that have no sensible interpretation in terms of explanatory frameworks.

RN where $A \in \mathcal{W}$: if $\vdash A$ then $\vdash \Box A$
T $\Box A \supset A$

A semantics of **T** that is sound and complete with this axiomatization can be found in Batens et al. (2003, pp. 46-47), which is a typical Kripke semantics in terms of a set of worlds and an accessibility relation over them.

The language $\mathcal{L}^{\Box}$ allows us to represent the premises involved in abductive reasoning processes with the expressive power of classical first-order logic, but gives us the extra operator $\Box$, which allows us to indicate at the object level that a certain generalization is in the explanatory framework. Let $\mathcal{F}^{\circ}$ denote the set of *purely functional formulas*, i.e. formulas that contain no individual constants, quantifiers, or sentential letters. For example, $Px \wedge (Qxy \vee Rx)$ is a purely functional formula, whereas $Pa \vee Qxy$ and $Px \wedge \exists y Qxy$ are not. For $A \in \mathcal{F}^{\circ}$, let $\forall A$ be the universal quantification over every variable that is free in A. The logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$ treats any formula of the form $\Box\forall(A \supset B)$ with $A, B \in \mathcal{F}^{\circ}$ as an element of the explanatory framework.

The choice for **T** in order to model the explanatory framework has two important consequences. First of all, in view of the rule RN and the axiom K, classical logic consequences of the explanatory framework may themselves be used to generate explanatory hypotheses. For instance, if $\Box\forall x(Px \supset Qx)$ and $\Box\forall x(Qx \supset Rx)$ are premises of a particular abductive problem, not only these formulas but also $\Box\forall x(Px \supset Rx)$ will be part of the explanatory framework. Second, in view of axiom T, a generalization that is part of the explanatory framework is, as such, assumed to be true. This is the formal expression of our restriction to the classical account of abduction, where "*A* explains *B*" implies "*A* implies *B*".

As we will explain below, the logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$ is a non-monotonic extension of **T**, which is itself a monotonic extension of **CL**. Hence, $\mathbf{LA}^{\mathbf{r}}_{\forall}$ provides only sensible consequences under the assumption that the explanatory framework and the set of known facts relevant to the abductive problem are mutually consistent; otherwise, it results in plain triviality.

Our logic for AG is in a sense minimal: iterations of boxes are excluded, and explanation is expressed by rather simple formal tools. It is a topic for further research whether our model can be meaningfully extended to include specific, more fine-grained accounts of explanation (e.g. adding asymmetric axioms to specify causal arguments in the sense of Hausman 1998).

6.3.3 The Dynamics of AG

Apart from the distinction between the explanatory framework and mere generalizations, several other difficulties arise when we try to model abduction in general, and AG in particular.[16] First of all, abduction is a non-monotonic method of reasoning: new information may contradict the hypotheses we have raised. Moreover, it may not always be clear at a certain point whether the currently available information contradicts some of these hypotheses. This requires further inferences, which might not yet have been drawn. As a result, we can discern a double dynamics in abductive reasoning: previously drawn inferences can be retracted in view of additional premises, but also in view of further inferences from the same body of evidence. Therefore, the proof theory of a formal logic for AG should be able to frame this double dynamics, yet the logic still needs to define a sensible and stable output for any given premise set.

Second, every realistic model of ampliative reasoning (such as abduction) should allow us to combine deductive (classical) inferences with ampliative (supraclassical) steps. That is, it should allow the user to draw new inferences on the basis of previously inferred hypotheses, and it should allow the classical consequences of the evidence to falsify such hypotheses (and whatever we derived from them).

This relates to a third important desideratum, i.e. that the hypotheses yielded by a formal logic for abduction should be mutually consistent with the evidence and the explanatory framework. Ampliative reasoning should not only allow us to go beyond the mere deductive consequences of our knowledge, but should also remain within the boundaries of consistency.

The fourth problem is specific to the context of abduction: explanatory hypotheses should be as logically parsimonious as possible. For instance, if Y suffices to explain X, then we should not raise the explanatory hypothesis $Y \wedge Z$. More generally, we should aim to derive only the logically weakest hypotheses that suffice to explain the explananda.[17]

Finally, any logic for abduction should be able to handle cases of multiple explanatory hypotheses (see Section 4.1) in a consistent and uniform

[16] As this chapter is based on a stand-alone article, the following paragraphs may contain some overlap with previous chapters.

[17] As indicated by one of the referees, *logical* parsimony should be distinguished from *expressive* parsimony. For instance, if $Y \vee Z$ explains X, than the explanatory hypothesis Y is expressively more parsimonious because it contains fewer different terms, but logically less parsimonious than the explanatory hypothesis $Y \vee Z$ because Y logically entails $Y \vee Z$.

way.

We chose to use the framework of adaptive logics to formulate the logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$ for AG. Adaptive logics are powerful formal systems that explicate various forms of defeasible reasoning such as reasoning on the basis of inconsistent premises (Batens, 1999), inductive generalization (Batens, 2011), reasoning on the basis of conflicting norms (Van De Putte and Straßer, 2013), etc. Several adaptive logics have also already been developed for singular fact abduction (Meheus, 2011; Gauderis, 2013a) and all of them were shown to meet the above desiderata.

One of the most important developments within the adaptive logics program is the definition of a canonical format, the so-called *standard format* for adaptive logics. This format encompasses a generic dynamic proof theory and a selection semantics. A rich and attractive metatheory has been shown to hold generically for all adaptive logics in standard format (see Batens 2007, n.d.): they are sound and complete, have the reassurance property, their consequence relation is idempotent, cautiously monotonic, etc. Most adaptive logics have been successfully expressed within this format, so it provides a good basis for a unifying study of defeasible reasoning forms in general, and of patterns of abduction in particular.

The main motivation to choose this non-monotonic framework is its dynamic proof theory, which enables us to construct proofs that are very similar to actual human reasoning processes, as will become clear from the examples in Section 4. There we will also argue that each of the other desiderata from the current section are met by $\mathbf{LA}^{\mathbf{r}}_{\forall}$.

6.4 Presentation of The Logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$

6.4.1 The Definition of $\mathbf{LA}^{\mathbf{r}}_{\forall}$

As explained in Section 4.2, an adaptive logic in standard format is characterized by a triple $\langle \mathbf{LLL}, \Omega, \mathbf{x} \rangle$ consisting of a *lower limit logic* $\mathbf{LLL}$, a *set of abnormalities* Ω, and a *strategy* $\mathbf{x}$.

The adaptive logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$ employs $\mathbf{T}$ as its lower limit logic. The set of abnormalities of $\mathbf{LA}^{\mathbf{r}}_{\forall}$ requires a bit more explanation. Consider once more the inference schema of AG introduced in Section 6.3 (p. 129), this time capturing the distinction between mere generalizations and the explanatory framework:

$$\begin{array}{ll}
\text{(P1)} & \forall(A \supset B) \\
\text{(P2)} & \Box\forall(C \supset B) \\
\hline
\text{(H)} & \forall(A \supset C)
\end{array}$$

The $\mathbf{LA}^{\mathbf{r}}_{\forall}$-abnormalities are all formulas which imply that the premises in the above schema are true, whereas its conclusion is false, for a particular A, B and C. To keep the formulas comprehensible, let us first introduce two abbreviating notations. First, let

$$A \nrightarrow_C B =_{df} \forall(A \supset B) \wedge \Box\forall(C \supset B) \wedge \neg\forall(A \supset C)$$

$A \nrightarrow_C B$ can be read as: “although all A are B and C-hood explains B-hood, it is not the case that all A are C”. Second, where $A, B \in \mathcal{F}^{\circ}$, let $A||B$ denote the fact that A and B share no predicates.

Using these two abbreviations, we can now define the set of abnormalities of $\mathbf{LA}^{\mathbf{r}}_{\forall}$:

$$\Omega = \{A \nrightarrow_C B \mid A, B, C \in \mathcal{F}^{\circ}, A||B \text{ and } B||C\}$$

The restrictions $A||B$ and $B||C$ are added to avoid that certain trivial self-explanatory hypotheses block the derivation of other hypotheses.[18]

The strategy of $\mathbf{LA}^{\mathbf{r}}_{\forall}$ is *reliability*, which will be explained in Section 6.4.2.

As for all adaptive logics in standard format, the $\mathbf{LA}^{\mathbf{r}}_{\forall}$-semantics is obtained from the same triple $\langle \mathbf{T}, \Omega, \textit{reliability} \rangle$ – we refer to Batens (2007, n.d.) for a generic definition of the adaptive logics-semantics. In Section 6.4.3, we will present some particular features of $\mathbf{LA}^{\mathbf{r}}_{\forall}$ that show how it meets the desiderata from Section 3.3.

6.4.2 The Proof Theory of $\mathbf{LA}^{\mathbf{r}}_{\forall}$

The $\mathbf{LA}^{\mathbf{r}}_{\forall}$-proof theory is a mere instantiation of the generic proof theory for adaptive logics in standard format (see Section 4.2). In short, lines in adaptive proofs have, compared to standard logical proofs, an extra element, the condition, and can be marked at a certain stage. This happens if the formula of that line is considered to be no longer derivable at that stage of the proof. Adaptive proofs proceed to the next stage by applying one of the three generic rules PREM (for the introduction of premises), RU (for

[18] We refer to Van De Putte (2012, pp. 206-207) for examples that motivate these restrictions.

deductive steps) or RC (for defeasible steps), which results in the addition of a new line to the previous stage.

To get an idea of how this proof theory works for $\mathbf{LA}^{\mathbf{r}}_{\forall}$, consider the formalization of the pineapple-example from the introduction (p. 121):

$$\Gamma_1 = \{\forall x(Px \supset Qx), \Box\forall x(Rx \supset Qx), \exists x Px\}$$

The last premise is added to avoid certain unwelcome results (see Section 6.4.3). In view of the interpretation of the premises, this is a harmless addition: if we want to explain the fact that all pineapples taste sweet, then it seems evident that we also know that pineapples exist.

We start an $\mathbf{LA}^{\mathbf{r}}_{\forall}$-proof from Γ_1 by writing down two of the premises:

1	$\forall x(Px \supset Qx)$	PREM	$\emptyset$
2	$\Box\forall x(Rx \supset Qx)$	PREM	$\emptyset$

As $\forall x(Px \supset Qx), \Box\forall x(Rx \supset Qx) \vdash_{\mathbf{T}} \forall x(Px \supset Rx) \vee (P \nrightarrow_R Q)$, we may apply the rule RU to derive $\forall x(Px \supset Rx) \vee (P \nrightarrow_R Q)$ and, from the latter, that all P are R by RC:[19]

3	$\forall x(Px \supset Rx) \vee (P \nrightarrow_R Q)$	1,2;RU	$\emptyset$
4	$\forall x(Px \supset Rx)$	3;RC	$\{P \nrightarrow_R Q\}$

To illustrate the dynamic flavor of this logic, we have to add more premises to Γ_1. Suppose that we learn about a genetically modified pineapple a, which contains no sugar, but nevertheless tastes sweet because it contains a synthetic type of sweetener. This can be modeled by adding to Γ_1 the premise $Pa \wedge \neg Ra$, which contradicts the hypothesis $\forall x(Px \supset Rx)$. Let us call the extended premise set Γ_2. Since the proofs are dynamic, we need, as explained in Section 4.4, not to start the proof all over again; we can just pick up where we ended our line of thought.

⋮	⋮	⋮	⋮
4	$\forall x(Px \supset Rx)$	3;RC	$\{P \nrightarrow_R Q\}\ \checkmark^7$
5	$Pa \wedge \neg Ra$	PREM	$\emptyset$
6	$\neg\forall x(Px \supset Rx)$	5;RU	$\emptyset$
7	$P \nrightarrow_R Q$	1,2,6;RU	$\emptyset$

[19] Line 3 is added for the sake of clarity. In view of the definition of the rule RC, it is also possible to derive the formula on line 4 directly from those on lines 1 and 2, which we will do further on.

At line 7, we have reached the insight that $P \nrightarrow_R Q$ follows from our premises by **T**. Hence, we need to indicate that there is something wrong with the condition of line 4, which is done, as explained before, by marking it with a $\checkmark^7$-sign.

However, there is a difference with the logic $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$ from chapter 4. For that logic, the language scheme and set of abnormalities were constructed in such a way that if a disjunction of abnormalities could be (unconditionally) derived, also one of these disjuncts could be derived independently (see the formal condition on p. 88).

This is not the case for the present logic. There exist premise sets of which disjunctions of abnormalities can be derived unconditionally. For instance consider $\omega_1, \omega_2 \in \Omega$ for which, for a certain premise set Γ, $\Gamma \vdash_{\mathbf{T}} \omega_1 \vee \omega_2$, but not $\Gamma \nvdash_{\mathbf{T}} \omega_1$ nor $\Gamma \nvdash_{\mathbf{T}} \omega_2$. Were we to use the simple strategy, we would not have to revoke a formula that is derived on condition ω_1. This is clearly naive: it may be that ω_1 is not unconditionally derivable from the premise set, but as it is derived as part of a disjunction, it is unreliable. The *reliability* strategy expresses this idea by marking all conditions that appear as an individual disjunct in a minimal disjunction of abnormalities.

Definition 6.1 (**Minimal** Dab**-formula at stage** s). *A Dab-formula* $Dab(\Delta)$[20] *is a* minimal Dab-formula at stage s *if and only if* $Dab(\Delta)$ *is derived on the empty condition at stage s, and there is no* $\Delta' \subset \Delta$ *for which* $Dab(\Delta')$ *is derived on the empty condition at stage* s.

Definition 6.2 (**Set of unreliable formulas** $U_s(\Gamma)$ **at stage** s). *The* set of unreliable formulas $U_s(\Gamma)$ *at stage* s *is the union of all* Δ *for which* $Dab(\Delta)$ *is a minimal Dab-formula at stage* s.

Definition 6.3 (**Marking for the reliability strategy**). *Line* i *with condition* Θ *is marked at stage* s *of a proof if and only if* $\Theta \cap U_s(\Gamma) \neq \emptyset$.

It is important to remark that, despite the dynamic character of the proofs, adaptive logics are proper proof-invariant logics. Given a Γ and A, the logic defines whether A is a consequence of Γ or not. This does not depend on the way we start a proof or proceed through one. To avoid confusion with formulas that are derivable at a certain stage of a proof (but can be defeated at a later stage), formulas in the consequence set are

[20] Recall, $Dab(\Theta)$ is the (classical) disjunction of the abnormalities in a finite subset Θ of the set of abnormalities Ω.

called *finally derivable*. The final derivability relation of an adaptive logic is defined as follows:[21]

Definition 6.4. *A formula A is* derived from Γ at stage *s of a proof if and only if A is the formula of a line that is unmarked at stage s.*

Definition 6.5. *A formula A is* finally derived from Γ *at stage s of a proof if and only if A is derived at line i, line i is not marked at stage s and every extension of the proof in which i is marked may be further extended in such a way that line i is unmarked.*

Definition 6.6 (Final Derivability). $\Gamma \vdash_{\mathbf{LA}^{\mathbf{r}}_{\forall}} A$ *($A \in Cn_{\mathbf{LA}^{\mathbf{r}}_{\forall}}(\Gamma)$) if and only if A is finally derived in an $\mathbf{LA}^{\mathbf{r}}_{\forall}$-proof from Γ.*

To illustrate the above definitions, consider again our proof from Γ_2. Since the Dab-formula at line 7 is a single abnormality, it will be a minimal Dab-formula in every extension of the proof. Hence, line 4 will remain marked in every such extension. More generally, $\forall x(Px \supset Rx)$ is not finally derivable from Γ_2, i.e. there is no proof in which we can finally derive $\forall x(Px \supset Rx)$ from this premise set.

6.4.3 Some Salient Features of the Logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$

We end this section with a brief survey of the ways in which $\mathbf{LA}^{\mathbf{r}}_{\forall}$ solves some typical problems for any formal model of abduction. First of all, as any adaptive logic in standard format, $\mathbf{LA}^{\mathbf{r}}_{\forall}$ has the Reassurance property (Batens, 2007, Corollary 1):

Theorem 6.7. *If Γ is not $\mathbf{T}$-trivial, then neither is $Cn_{\mathbf{LA}^{\mathbf{r}}_{\forall}}(\Gamma)$. (Reassurance)*

This means that if our explanatory framework and our factual knowledge are mutually consistent, then $\mathbf{LA}^{\mathbf{r}}_{\forall}$ will always yield a consistent set of explanatory hypotheses. This is an immediate consequence of the fact that **CL** is included in **T** and of the axiom T.

Second, as explained in Section 3.3, a logic for abduction should yield only the most parsimonious hypotheses. Consider the following proof from Γ_1:

[21] As explained in Batens (2007), adaptive logics in general lack a positive test. We refer to Meheus (2011) for an extensive discussion of this fact.

1	$\forall x(Px \supset Qx)$	PREM	$\emptyset$
2	$\Box\forall x(Rx \supset Qx)$	PREM	$\emptyset$
3	$\Box\forall x((Rx \wedge Sx) \supset Qx)$	2;RU	$\emptyset$
4	$\forall x(Px \supset (Rx \wedge Sx))$	1,3;RC	$\{P \not\twoheadrightarrow_{R\wedge S} Q\}$ $\checkmark^{9}$
5	$\forall x(Px \supset Sx)$	4;RU	$\{P \not\twoheadrightarrow_{R\wedge S} Q\}$ $\checkmark^{9}$
6	$\exists xPx$	PREM	$\emptyset$
7	$\exists x(Px \wedge \neg Sx) \vee \exists x(Px \wedge Sx)$	6;RU	$\emptyset$
8	$\neg\forall x(Px \supset (Rx \wedge Sx)) \vee \neg\forall x(Px \supset (Rx \wedge \neg Sx))$	7; RU	$\emptyset$
9	$(P \not\twoheadrightarrow_{R\wedge S} Q) \vee (P \not\twoheadrightarrow_{R\wedge\neg S} Q)$	1,2,8;RU	$\emptyset$

As the material implication has the property $A \supset B \vdash (A \wedge C) \supset B$ (*strengthening the antecedent*), the hypothesis on line 5, which states that anything that is P also has the random property S, could be derived. However, using the premise $\exists xPx$, we can derive the *Dab*-formula on line 9 which defeats lines 4 and 5.

Third, the dynamic proof theory and the form of the abnormalities also ensure that no hypotheses can be finally derived from tautologies, and that no contradictions can be finally derived as a hypothesis. The following proof from Γ_1 illustrates how the logic enables us to defeat both self-contradictory hypotheses and hypotheses derived from tautologies.

1	$\forall x(Px \supset Qx)$	PREM	$\emptyset$	
2	$\Box\forall x((Sx \wedge \neg Sx) \supset Qx)$	-;RU	$\emptyset$	
3	$\forall x(Px \supset (Sx \wedge \neg Sx))$	1,2;RC	$\{P \not\twoheadrightarrow_{S\wedge\neg S} Q\}$	$\checkmark^{6}$
4	$\exists xPx$	PREM	$\emptyset$	
5	$\exists x(Px \wedge \neg(Sx \wedge \neg Sx))$	4;RU	$\emptyset$	
6	$P \not\twoheadrightarrow_{S\wedge\neg S} Q$	1,2,5;RU	$\emptyset$	
7	$\forall x(Px \supset (Sx \vee \neg Sx))$	-;RU	$\emptyset$	
8	$\Box\forall x(Tx \supset (Sx \vee \neg Sx)))$	-;RU	$\emptyset$	
9	$\forall x(Px \supset Tx)$	7,8;RC	$\{P \not\twoheadrightarrow_{T} (S \vee \neg S)\}$ $\checkmark^{12}$	
10	$\Box\forall x(\neg Tx \supset (Sx \vee \neg Sx))$	-;RU	$\emptyset$	
11	$\neg\forall x(Px \supset Tx) \vee \neg\forall x(Px \supset \neg Tx)$	4;RU	$\emptyset$	
12	$(P \not\twoheadrightarrow_{T} (S \vee \neg S)) \vee (P \not\twoheadrightarrow_{\neg T} (S \vee \neg S))$	7,8,10,11;RU	$\emptyset$	

The final feature that will be illustrated is how this logic handles multiple explanatory hypotheses. Suppose that we become aware of a property S, which explains Q-hood just as well as the property R does. Hence we have to add the premise $\Box\forall x(Sx \supset Qx)$ to Γ_1, which results in the following set:

$$\Gamma_3 = \{\forall x(Px \supset Qx), \Box\forall x(Rx \supset Qx), \Box\forall x(Sx \supset Qx), \exists xPx\}$$

At first sight, both the hypotheses $\forall x(Px \supset Rx)$ and $\forall x(Px \supset Sx)$ can be derived from Γ_3. But, as shown in the proof below, these two formulas are not finally derivable. The composed hypothesis $\forall x(Px \supset (Rx \vee Sx))$, however, is finally derivable from Γ_3.

1	$\forall x(Px \supset Qx)$	PREM	$\emptyset$	
2	$\Box\forall x(Rx \supset Qx)$	PREM	$\emptyset$	
3	$\Box\forall x(Sx \supset Qx)$	PREM	$\emptyset$	
4	$\forall x(Px \supset Rx)$	1,2;RC	$\{P \not\twoheadrightarrow_R Q\}$	$\checkmark^{9}$
5	$\forall x(Px \supset Sx)$	1,3;RC	$\{P \not\twoheadrightarrow_S Q\}$	$\checkmark^{10}$
6	$\exists xPx$	PREM	$\emptyset$	
7	$\neg\forall x(Px \supset Rx) \vee \neg\forall x(Px \supset (Sx \wedge \neg Rx))$	6;RU	$\emptyset$	
8	$\Box\forall x((Sx \wedge \neg Rx) \supset Qx)$	3;RU	$\emptyset$	
9	$(P \not\twoheadrightarrow_R Q) \vee (P \not\twoheadrightarrow_{S\wedge\neg R} Q)$	1,2,7,8;RU	$\emptyset$	
10	$(P \not\twoheadrightarrow_S Q) \vee (P \not\twoheadrightarrow_{R\wedge\neg S} Q)$	1,2,3,6;RU	$\emptyset$	
11	$\Box\forall x((Rx \vee Sx) \supset Qx)$	2,3;RU	$\emptyset$	
12	$\forall x(Px \supset (Rx \vee Sx))$	1,11;RC	$\{P \not\twoheadrightarrow_{R\vee S} Q\}$	

In view of this last feature, $\mathbf{LA}^{\mathbf{r}}_{\forall}$ models a kind of *practical abduction*: whenever multiple explanatory hypotheses are available, $\mathbf{LA}^{\mathbf{r}}_{\forall}$ allows only for the (undefeated) derivation of a disjunctive combination of these hypotheses. This is opposed to *theoretical abduction*, in which each of the individual hypotheses can be separately derived. For a more thorough discussion of this distinction, see Gauderis (2013a) (Chapter 4).

6.5 A New Problem for the Logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$

Recently, while working on the problem of planning in the field of artificial intelligence,[22] Mathieu Beirlaen found that the logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$ from this chapter and all other so-far constructed adaptive logics that model a form of *practical* abduction (such as the aforementioned $\mathbf{LA}^{\mathbf{r}}_{\mathbf{s}}$ from Meheus 2011) are actually too weak. It appears that if a class (or an object, in the case of singular fact abduction) has two properties that can be independently explained, no explanatory hypotheses at all can be derived for the class (or object). To illustrate this problem for $\mathbf{LA}^{\mathbf{r}}_{\forall}$, consider the premise set $\Gamma_4 = \{\forall x(Px \supset (Qx \wedge Rx)), \Box\forall x(Sx \supset Qx), \Box\forall x(Tx \supset Rx), \exists xPx\}$.

1	$\forall x(Px \supset (Qx \wedge Rx))$	PREM	$\emptyset$

[22] From a logical perspective this problem is very similar to abduction: a goal that has to be reached resembles an explanandum, while different steps that can be taken to reach the goal are similar to the different explanatory hypotheses that explain the explanandum.

2	$\Box\forall x(Sx \supset Qx)$	PREM	$\emptyset$	
3	$\Box\forall x(Tx \supset Rx)$	PREM	$\emptyset$	
4	$\forall x(Px \supset Sx)$	1,2;RC	$\{P \nrightarrow_S Q\}$	$\checkmark^7$
5	$\forall x(Px \supset Tx)$	1,3;RC	$\{P \nrightarrow_T R\}$	$\checkmark^8$
6	$\exists x Px$	PREM	$\emptyset$	
7	$(P \nrightarrow_S Q) \vee (P \nrightarrow_{T \wedge \neg S} R)$	1,2,3,6;RU	$\emptyset$	
8	$(P \nrightarrow_T R) \vee (P \nrightarrow_{S \wedge \neg T} Q)$	1,2,3,6;RU	$\emptyset$	

This is a severe problem. For instance, suppose that we observe that bananas are curved and taste sweet, and we know in general that a sweet taste can be explained by the presence of sugar and that a curved shape can be explained by attraction towards sunlight during growth. Still, in this case, using the logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$, we can neither hypothesize that bananas contain sugar, nor that they grow towards sunlight.

This problem appears not to be resolvable by a simple adjustment to the logic. Mathieu Beirlaen (private communication) is currently working on a (not-yet published) solution in terms of deontic operators, which is in essence applicable to the presented logic. The idea is to enrich the language and indicate at the level of the premises whether different explanations are "allowed" together or not.

This solution might be apt for the context of planning in AI. Yet I am not convinced that this solution is the right way to remedy the logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$, which is designed for the context of scientific reasoning. In fact, I consider this problem rather as an argument against the use of logics for *practical abduction* in the context of scientific reasoning. One should not worry too much about how different hypotheses for different observed characteristics relate formally to each other (whether they should mutually block each other, be in a disjunction, or be in a conjunction), because at this stage of the discovery process this is not very relevant. Different hypotheses are better pursued independently of the other possible hypotheses.[23] Therefore, I think it is better to focus (in the context of scientific reasoning) on logics that model *theoretical abduction*.

[23] It is true that in some cases, there might be good reasons to connect two hypotheses and pursue them together. Yet, such reasons cannot be formal: they are topic- and content-dependent, which put them outside the scope of this logic. Of course, this does not preclude *a priori* that this process of relating hypotheses cannot be studied formally. It just means that such a reasoning process needs another formal framework that can represent reasons to relate hypotheses.

A logic for the theoretical abduction of generalizations can be rather straightforwardly built from the logic $\mathbf{LA}^{\mathbf{r}}_{\forall}$ in a manner analogous with how the logic $\mathbf{MLA}^{\mathbf{s}}_{\mathbf{s}}$ from Chapter 4 (which models theoretical singular fact abduction) is built from the logic $\mathbf{LA}^{\mathbf{r}}_{\mathbf{s}}$ (for practical singular fact abduction). This result, call it $\mathbf{MLA}^{\mathbf{s}}_{\forall}$, would be a bimodal logic in which one modal operator indicates whether the formula is a hypothesis or part of the background knowledge and the other whether the formula is explanatory. Yet it remains to be studied how exactly these operators relate to each other and how premise sets should be formally represented.

6.6 Conclusion

As argued in this chapter, abduction of generalizations (AG) is ubiquitous in everyday and scientific reasoning. We provided a first general analysis of this pattern, and argued that the notion of an explanatory framework should be embodied in any formal model for AG. This idea was implemented in $\mathbf{LA}^{\mathbf{r}}_{\forall}$, which is a well-behaved formal logic that aims to model AG.

An open question remains whether one may obtain a sensible interpretation of the Kripke semantics for this application of (extensions of) **T**. In this way, assumptions about the notion of the explanatory framework may be translated into formal properties of the accessibility relation and vice versa. We focused here mostly on the proof theoretic aspects of our formal model and consider this a topic for future research.

Several enrichments of our formal model can be studied, in order to deal with e.g. probabilistic information, causal arguments, and abductive anomalies.[24] Also, it seems worthwhile to develop ways in which singular fact abduction and AG can be integrated in the framework of adaptive logics to model examples such as those mentioned at the end of Section 6.2.1. Finally, case studies of some of the examples mentioned in Section 2 may shed new light on the relation between AG, unification and other patterns of abduction.

[24] In Aliseda's (2006) terminology, an anomaly is a fact, the negation of which follows from our background theory.

Part III

A Historical Case

motivation

In this part, I will focus on the perspective of individual agents. Studying hypothesis formation from this perspective will give us some insights into the reasons, preferences and circumstances that lead individual agents to form certain hypotheses in answer to particular triggers. In order to understand how these characteristics play their role in hypothesis formation, I will consider an actual (historical) case from science in which the various actors advanced not only quite different hypotheses in answer to a single research question, but also employed different patterns of hypothesis formation.

My main goal is to get some initial understanding of why agents proceed according to one rather than another pattern of hypothesis formation for a given problem by contrasting the reasoning of the various physicists that tried to address one of the big experimental puzzles of early 20th century nuclear physics, the anomalous β spectrum.

The core assumptions at the heart of this part of the dissertation are (1) that given a particular problem, various patterns of hypothesis formation can be applicable; (2) that the choice as to which pattern and which hypothesis looks initially most promising (i.e. before the hypothesis is actually pursued and experimental evidence is collected) is based on the personal preferences of the individual agent; and (3) that such preferences and their motivation for a particular case can be brought forward by means of historical research. Let us turn to the motivation for adopting each of these assumptions.

Patterns of hypothesis formation For the existence of patterns of hypothesis formation, I refer to the motivation of Part II, in which I have argued for this claim. That for some research questions, various of these patterns are applicable is illustrated by the case study of the β spectrum, which is presented in this part. For those who are confused about this claim – the various patterns of hypothesis formation have formally different premises, so how can they be applicable to the same case? – it is important to remember that the classification of hypothesis formation patterns is a classification of descriptions of hypothesis formation instances. A single problem can be interpreted and described in various ways.

Choices depend on preferences That the choice as to which pattern should be employed depends on the agent might at first sound trivial: choices, rational or not, depend (assumed that they are freely made) on the preferences of the agent, and these preferences are in turn shaped by personal characteristics such as experiences, beliefs, values, etc. The question is which of these personal aspects most motivates an agent's preferences. To determine this, for the case of hypothesis formation in science, will be our goal in this part.

Yet, given the nature of our subject, there are at least two good reasons to doubt this assumption: (1) Why should there be a choice or preference? If multiple patterns of hypothesis formation are applicable to a certain problem, why not apply multiple patterns and decide between them only after further evidence is gathered in the process of hypothesis selection? (2) Why should the perspective of the individual agent matter? Is it not the aim of science to reduce personal preferences as much as possible? Why can we not study hypothesis formation by looking exclusively at the field in which the agent is working, with its paradigms, theories and research problems – the macro-structure – and the various methods and patterns available to agent – the micro-structure?[1]

The first question is a pertinent one. To further motivate its importance, consider how humans typically program artificially intelligent agents for research tasks. First, they instruct the agent to generate as many hypotheses as possible, assigning each of them an equal *a priori* likelihood. Next, the agent is instructed to start collecting evidence and update the likelihoods of the various hypotheses accordingly (generally via Bayesian methods). If this is our rational view of how research should proceed, why do we find

[1] See the three broad perspectives distinguished in Section 1.5.

over and over again scientists who single out certain hypothetical positions for which so far no evidence has been gathered? Apparently, history shows us that often some form of initial selection already occurs during the process of hypothesis formation.

The short answer to this question is that it is a matter of resources. While several accounts of rationality tend to neglect this issue, scientists in the actual world are limited beings with limited time, limited funding and limited cognitive energy. At the same time, the pursuit of one particular hypothesis can be a painstaking process requiring many years of research effort of a whole team. Therefore, even if virtually no evidence is available, researchers are obliged to make choices, and to motivate these they can only turn to things such as their experience or theoretical considerations.

Yet this answer, although in line with how we observe actual science develop, is not fully satisfying. Being unable to pay equal attention to a large number of hypotheses is one thing; sticking to only one hypothesis is another. As we will observe in the case study, scientists may tend just to stick to *their* idea, often more than would be rationally justified. Apparently, human agents are not very good at entertaining multiple hypotheses (see also Part I), or, probably related, people find it hard to motivate themselves or their collaborators to conscientiously pursue particular suggestions or to convince funding agencies if no clear hypothesis and research direction is chosen.[2] Increasing awareness of this tendency to psychologically eliminate certain hypotheses prematurely can only be of benefit to science.

This brings us to the second question: if limited resources require early selection during hypothesis formation, why should this choice be motivated by elements of the personal perspective of the agent? If at the micro-level various patterns or methods lead to different hypotheses, why would the collective experience and considerations of the field not unequivocally indicate the most plausible hypothesis or hypotheses given the current state of knowledge?

In fact, we often find a broad consensus concerning the best way forward and the most plausible hypothesis to pursue. Only in this way can multi-billion dollar budgets be freed for the experimental pursuit of recent hypotheses such as the Higgs boson or dark matter. But such consensus on the question of which hypothesis to pursue (and which to fund) is achieved

[2] See also Glass and Hall's (2008) criticism of funding agencies that require research objectives to be stated as factual hypotheses. Some of their ideas will be further discussed in Section 9.3.3.

only after some initial progress, both theoretically and experimentally, has been made. At that point, the problem has often lost its status of being a deep theoretical problem. Deep problems, such as the case of the β spectrum, are problems that trigger many different hypotheses, divide the field, and lead to, in Lakatos's terminology, competing research traditions.

Therefore, if we want to shed light on the process of hypothesis formation, we have to focus on the period before the field has found a consensus concerning the best way forward. In the case of the β spectrum, this means the period between 1927, when the experimental anomaly was found, and 1934, when theoretically a consensus was reached that the existence of Pauli's neutrino was the most plausible option. At this point, the problem shifted to the (mostly) experimental quest to discover this new particle, which happened eventually in 1956 (see Franklin, 2001). Hence, in order to study hypothesis formation, we cannot fail to study how the preferences of individual agents are shaped and how they influence their hypothesis formation processes, all before a group decision is made.

Historical research can reveal these preferences This is an important issue, which I deal with in more detail in the methodological introduction of the chapter on the β spectrum (see Section 7.1). The position I argue for bites the bullet on this issue: I do not assume that historical research can reveal a complete factual account of how the protagonists' preferences were shaped by their experiences, theoretical considerations and values. One can through historical research, however, give a "how possibly" account of how their preferences were shaped and led to the hypotheses they suggested, an account that gives the best explanation of their writings and the historical conditions.

To a certain extent, any form of historical research, especially if it attempts to interpret the motives of historical agents, has to acknowledge the use of *Inferences to the Best Explanation*. So, as long as the full historical context is taken into account, and as long as more speculative passages are clearly marked, this should not constitute a problem for historical research.

For the purposes of this dissertation, it may be asked whether abandoning this assumption does not put in question my goal of understanding the process of hypothesis formation. I do not think that this has to be the case. If certain patterns are found in reasonable and empirally adequate reconstructions, these patterns can be further examined by considering other historical cases or confronting them with present research. Yet, as I have been able to cover extensively only one case study, these options remain for future research, and I can only claim to have looked for some initial insights on the issue.

strengths and weaknesses of the method

Strengths The main advantage of studying historical cases is that it is a method that allows one to look at scientific processes from a perspective that is very closely connected to the actual processes and which reduces *a priori* assumptions to a minimum. Only directly observing scientists (and the possibility to get direct feedback from them) might deliver results that are even more tightly connected to the actual processes. Yet this latter method has the drawback that the process of hypothesis formation is neither planned, nor easily observable, nor easily recognizable by the agents themselves. Also, it misses the clear view that comes with hindsight: not every formed hypothesis is equally interesting. To determine its interest, it has to be connected both with the past (is it a novel approach to the problem?) and with the future (has it proved to be a fruitful approach?). The combination of hindsight and of a close connection to the actual processes is unique to the historical methods.

Related to the previous point, if the research is done sufficiently elaborately, the method of case studies will allow one to grasp most of the full complexity and richness of actual processes. It will allow one to pay attention to all nuances and details, including facets that cannot be covered in any more formal treatment.[3]

Weaknesses The method of historical case studies faces, however, an obvious limitation: how to generalize them? How can conclusions about a general process be drawn from a single case or instance?

I do not think there is a straightforward answer to this question: drawing general conclusions from case studies is always somewhat tricky. Yet a

[3] An interesting example of this, coming from the case of the β spectrum, is the fact that Pauli suggested the particle that later would become known as the neutrino barely one week after his first wife left him. At first sight, common sense instructs us to leave such anecdotes from his private sphere outside any serious analysis of his hypothesis formation process. Yet Pauli was also human, so irrational behavior was most likely not fully unfamiliar to him. So, could it not be that his personal turmoil incited him to make bolder leaps of imagination and to keep less reservations than he usually did? The historian Pais (1986), who personally knew Pauli, thought it mattered.

few things can be considered. The case study of the β spectrum has been chosen specifically because it meets certain desiderata: it had to be a well-documented single problem that triggered at least three different patterns of hypothesis formation within a single community, or at least among contemporaries who had more or less the same methods at their disposal. So I could hope to bring forth some general or structural characteristics, by studying diverging instances within a single context. There are not that many passages in modern physics that meet these desiderata: it had to be a hard and deep problem for the field (otherwise people with a similar background would employ the same patterns of hypothesis formation), it had to be more than just two competing stances (although even this case eventually polarized to a debate between the supporters of Bohr and Pauli) and it had to occur in a relatively short timespan (as the methods of modern physics evolve too fast to claim that stances separated in time by over twenty years have access to the same methods).

Apart from this main drawback, two further issues that I already discussed may also be regarded as weaknesses of the method: the speculative flavor of certain historical reconstructions and the difficulty in drawing clear and precise normative conclusions.

overview of my contributions

In Chapter 7, the case study of the anomalous β spectrum is studied in great detail. No less than six very divergent hypotheses were put forward in a timespan of three years to account for this single experimental anomaly, which was first established in 1927. By identifying the main factors that fueled the hypothesis formation processes of the various actors in this case, it can be stated (for this case) that the protagonists' preferences were mostly shaped by their previous experience and by how they structurally related the different elements of the theory and field of physics. It is also shown that all of the protagonists reused and adapted older methods and ideas.

7 The Curious β spectrum

> *"Would you tell me, please, which way I ought to go from here?"*
>
> *"That depends a good deal on where you want to get to," said the Cat.*
>
> — Lewis Caroll, *Alice's Adventures in Wonderland*, 1865

This chapter is based on the paper "To envision a new particle or change an existing law? Hypothesis Formation and Anomaly Resolution for the Curious Case of the β Decay Spectrum", forthcoming in Studies in History and Philosophy of Modern Physics *(Gauderis, 2013b). I am indebted to Bert Leuridan and two anonymous referees for their helpful comments on earlier drafts. I further want to thank the participants of the various workshops and conferences where I presented (part of) these ideas for their many useful suggestions and questions.*

In this paper, the historical case of the anomalous β spectrum is examined, a puzzle that occasioned the most diverse hypotheses amongst physicists at the time. It is shown that initial preferences for a particular hypothesis are most often implicitly informed by scientists' individual perspectives on the structural relations between various elements of the theory. Also, it is argued that the adaptation of older ideas for new purposes is a far more common practice than is sometimes thought.

The content of the original article is retained, except for the second section, in which an introduction to the literature on hypothesis formation was presented. I have moved this section to the general introduction of the dissertation (Section 1.2). For general consistency with the remainder of this dissertation, small stylistic corrections have been made.

7.1 Introduction

In physics, as in other scientific disciplines, an anomalous experimental result can occasion the formation of formally quite different hypotheses. Confronted with such a result, a scientist has no strict guidelines to help her determine whether she should explain the result by withdrawing or adapting a constraint of the current theory (e.g. a law), or else by presupposing the existence of a hitherto unobserved entity that makes the anomaly fit within that theory (e.g. a particle). But she has more options than this: she can also suggest a new structural model, blame the anomaly on an overlooked feature of the experimental setting, or stretch and modify the theoretical classes that label the observables, among other possibilities.

If a scientist knows in advance which kind of hypothesis would best explain the anomaly, she can employ more efficient heuristics. For instance, when Max Planck was studying the experimental anomalies of Rayleigh-James's and Wien's laws for the spectrum of black bodies, he sought a new formula that fitted the data. Similarly, when Ernst Rutherford was confronted with the backwards scattering of α particles, he knew he had to construct a new structural model for the atom.

As the case study examined in this chapter illustrates, however, the situation is not always so clear: when an experimental anomaly proves perseverant, even the greatest minds in the field can differ strongly in opinion about which kind of hypothesis would lead to a satisfying explanation. Suggested hypotheses can vary so widely primarily because the determination as to which formal kind of hypothesis is needed is in itself an abductive and, hence, defeasible inference. Although often inferred implicitly, this choice is hugely important, as it determines what direction the initial search will take.

A lack of heuristics for this initial choice of hypothesis type presents itself as a problem especially when formal representations are utilized, such as in logic or AI. Because such representations determine in advance what types of hypotheses can be inferred, the choice of the type of hypothesis is (often implicitly) made when the premises are translated into the formal language: there are different ways to describe a (realistic) anomaly in natural language, any of which can lead to a different formal representation.

My aim in this chapter is not to suggest a normative heuristics for this choice, for given the lack of research in this area we lack sufficient knowledge about how scientists in the field decide on this matter (see Section

1.2 for an overview of the philosophical literature on discovery and hypothesis formation). Instead, my more modest goal is to examine how this choice was made in one notable instance, by examining a concrete case study with various diverging hypotheses. It will be shown that this choice is almost always implicitly made in a manner determined by the scientist's previous experiences and specific way of perceiving the problem, and that, moreover, scientists in general are sometimes, due to the strong ontological commitments their particular perspective often entails, very unwilling to accept other kinds of hypotheses.

Between 1927 and 1934, a manifest and persistent anomaly mystified the physics community: while α and γ decay behaved in a manner perfectly accordant with the new quantum mechanics, the energy of electrons emitted in β decay displayed a broad continuous spectrum. This puzzle intrigued the most established and famous physicists of the time, including Bohr, Heisenberg, Pauli, Rutherford, Chadwick, Ellis, and G.P. Thomson, and incited a lively debate among them. Curiously, all suggested hypotheses were of very different formal types: Ellis and Wooster were willing to give up the universality of the quantum postulate, Rutherford and Chadwick thought of varying internal energies, Bohr suggested a restriction of the energy conservation principle, Heisenberg tinkered with a second quantization of space at the scale of the nucleus, and Pauli proposed the existence of a new elementary particle – all these hypotheses being, as we will see, quite radical and highly controversial.

By focusing in detail on how these scientists arrived at their hypotheses, this chapter challenges the somewhat mythical proportions this episode has received in more popular histories of science, which, with its focus on genius and success, typically trace great discoveries back to a single man who enlightens his community by a kind of epiphany. But new ideas do not come out of nowhere; they are related to older suggestions. This debate also cannot be narrowed, as is often done, to Pauli's and Bohr's stances alone: many more ideas were around at the time, and all of them influenced each other.

I start, in Section 7.2, by introducing the case of the β spectrum historically, after which I analyze in detail the reasoning processes of six prominent physicists (or pairs of physicists) who tried to address this puzzle in Sections 7.3 through 7.8. Finally, in Section 7.9, I summarize these results and connect them back to the questions raised in this introduction.

Before we continue, some reservations about the methodology and scope of this chapter are in order.

First, I will not discuss this case in a purely historical or descriptive fashion, as this has been done sufficiently and extensively in other places such as Jensen (2000); Pais (1986); Bromberg (1971); Brown (1978); Hughes (1993); Navarro (2010); Cassini (2012); Guerra et al. (2012). Instead, I will try to reconstruct how the various protagonists could have reached the hypotheses they suggested and show how the choices they made along the way are related to their personal perspectives – a project I have been able to perform only because of the excellent scholarship on this period by historians of science. Their extensive coverage ensures that if the nearly impossible task of a full reassessment of the archival record (given the temporal and spatial scope of this episode) had been executed, it would only have had a minor impact on this project.

Second, in principle, there are at least three ways to study human reasoning processes such as hypothesis formation: from an internal perspective by analyzing direct feedback from the agents (e.g. psychological experiments), from an external perspective by linking the agents' recorded ideas to the historical and scientific context (e.g. historical case studies), or via simulation by trying to reproduce the agent's ideas (e.g. computational or logical approaches). As I do not assume that scientists make a conscious "metachoice" concerning which pattern of hypothesis formation is most appropriate for a particular problem, I believe that the examination of historical cases is the best method to gain some initial insight into how and why different patterns might have been employed in response to a single problem, as we can, by virtue of hindsight, situate these suggestions in their context. Having said this, of course, one should immediately note the drawback that we have no means to gather direct feedback from the agents themselves; we have only our interpretations of their scattered remarks, which are always based on assumptions and might be erroneous. This same problem occurs even when agents are alive and approachable, as agents tend to rationalize and reconstruct their thoughts afterwards.[1]

Therefore, I do not claim to offer a factual representation of the agents' thought processes. I aim rather to offer a coherent interpretation of how

[1] As Franklin (1993, n. 110) reported in his study of the rise and fall of the fifth force hypothesis, protagonists might fail to give an accurate view of their own ideas and positions, even though the interviews were conducted only a few years later. Sometimes, these reconstructions become apparent if they are confronted with external historical evidence, as for instance in Brown (1978), who showed that Pauli's recollections concerning whether he considered the neutrino to be a nuclear constituent were incorrect (see Section 7.8).

the protagonists' recorded ideas could have originated by making reasonable assumptions and specifying the historical surroundings. As Darden (1991, p. 4) has already acknowledged, this kind of research necessarily has a speculative flavor, as it can merely reconstruct "how" the agents "possibly" arrived to their hypotheses. Still, it offers us, as it is generally assumed in the literature on discovery (see Section 1.2), insights into the process of hypothesis formation that cannot be obtained by exploring logical principles or by psychological experiments alone. There is certainly value in trying to provide the best possible explanation of how actual agents in actual historically important debates arrived at their ideas, and such will be my aim in these pages.

7.2 The β puzzle in 1927

This introductory historical section provides the necessary background for the analyses in the following sections, but contains no novel results in itself, aside from making a case for the self-evidence of the *p-e* model. It first summarizes the relevant experiments that led to the β anomaly as it was perceived in 1927 (based on Franklin, 2001; Jensen, 2000; Pais, 1986; Malley, 2011), and completes this background picture with an overview of nuclear theory around 1927 and the various problems it faced (based on Stuewer, 1983; Brown, 2004; Pais, 1986; Hughes, 1993, 1998, 2003; Jensen, 2000; Fernandez and Ripka, 2013).

7.2.1 Experimental History of the β spectrum

The story of the β puzzle goes back to 1896, when Henri Becquerel discovered the phenomenon of radioactivity: some particular substances radiate spontaneously and independently of any interaction with the environment. The discovery of this curious form of radiation was made by mere luck; it revealed itself for the first time in the imprints left by uranium on some photographic plates that Becquerel had stored in a dark cupboard, deprived of all incoming sunlight.

From that moment onward, experimental discoveries unfolded at a steady pace. In 1899, Ernst Rutherford showed that the radiation emitted by uranium consisted of at least two different kinds of radiation, which he labeled α and β radiation. Even though α radiation was identified by Rutherford as helium ions only in 1907, it was already established in 1904 by William Bragg that it had a mono-energetic spectrum, i.e. that α particles of a particular radioactive element are always emitted with the same characteristic amount of energy. For β radiation, it was already suggested shortly after Rutherford's discovery that it consisted of electrons (the elementary particles then recently discovered in cathode rays by J.J. Thomson). By 1902, this thesis was confirmed by experimental evidence provided by Becquerel and Walter Kaufman. Their experiments even hinted at a possible continuity of the β spectrum, though this idea was not accepted by the community at the time. According to Franklin (2001, p. 30), this was a justified call given that their experimental setup was too inaccurate to draw such a conclusion. The main reason why this idea was not taken seriously at the time, however, was the general expectation that β decay would prove analogous to α decay, and so produce a mono-energetic spectrum.

In 1909, William Wilson argued that β rays could not be a homogenous stream of mono-energetic electrons, given that, in matter, β particles had an exponential absorption curve, while homogenous electron streams (such as cathode rays) had a linear curve. Hence, the electrons found in β rays must have a range of energies, a variety that could not be explained by analogy to α decay.

Shortly after this, improved energy spectra for β radiation showed the occurrence of multiple lines, which suggested that there existed a discrete set of possible emission energies. As such, one suspected that β sources consisted of multiple unstable elements, still all decaying with a characteristic energy and, therefore, resulting in a single line in the spectrum. But, as line spectra grew more detailed as the quality of spectral photography improved, more and more lines appeared, and it came to be understood that it was "impossible to assume a separate substance for each beta line" (Otto Hahn, as cited in Franklin, 2001, p. 43). Apparently, β radiation was truly heterogeneous.

In 1914, while theoretical explanations for these line spectra were still lacking, James Chadwick and Hans Geiger tried to count the distribution of electrons in these lines with an improved particle counter. To their great surprise, they found hardly any line. For the first time, they established the continuous spectrum of β radiation on a solid experimental basis. The earlier observed complex line spectra proved to be just a secondary effect of the process of spectral photography.

Experimentalists were left perplexed. The idea that β decay was emitted with a continuous spectrum seemed impossible. Many, most prominently Lise Meitner, Otto Hahn, and Chadwick himself, put forward a long list of hypotheses to explain this surprising result, such as secondary radiation of the electrons, the production of recoil electrons, influence by γ rays, etc. What all these hypotheses had in common was that all supported the initially mono-energetic emission of β particles, and ascribed the continuity to subsequent secondary processes, somewhere between the radioactive source and the measurement of the spectrum.

This speculation came to an end in 1927, when Charles D. Ellis and W. A. Wooster from the Cavendish laboratory in Cambridge (which had been led by Rutherford since 1919) constructed a direct test to determine whether energy was lost between the β ray source and the location of measurement. By determining the average heat increase per β particle emission, they found that the energy needed for this increase was the average and not the maximum of the β spectrum. This means that no energy was lost after the emission, and that, hence, the β particles left the source with a continuous spectrum.

This experiment did not immediately settle the question, however. Although starting to question her own secondary origin hypotheses, Lise Meitner came to doubt whether Ellis and Wooster had controlled for all these possible secondary effects in their experiments, as a result of which certain continental physicists, by contrast with their colleagues in England, did not have much confidence in the Cavendish results. Until 1929, Pauli, for example, thought that non-detected γ rays were the cause (Jensen, 2000, pp. 137-143, Rueger, 1992, p. 315). Only after Meitner and Orthmann replicated, improved and confirmed the Ellis and Wooster experiments in the late spring of 1929 did a general consensus arise concerning the continuity of the β spectrum. In a famous letter to Ellis (in July 1929), Meitner admitted that:

> It seems to me now that there can be absolutely no doubt that you were completely correct in assuming that beta radiations are primarily inhomogeneous. But I do not understand this result at all. (Meitner, as cited in Franklin, 2001, p. 59).

Before examining the various proposals to explain this counterintuitive result, I will complete the background picture by sketching the contours of nuclear theory in 1927. More particularly, we must consider the nuclear model which prevailed at the time and its difficulties.

7.2.2 Nuclear theory in 1927

In 1927, the prevailing model for the constitution of atomic nuclei was the so-called *p-e* model. In this model, the nucleus of an atom with mass number A and charge number Z consists of A protons (p) and A−Z electrons (e), kept together by the electromagnetic force. For example, according to the *p-e* model, the α particle, identified as the nucleus of $^4_2$He, consists of four protons and two electrons. By comparison, in our current understanding, this particle consists of two protons and two neutrons, held together by the residual strong force.

While this *p-e* model became hugely problematic around 1927, it was the core assumption of virtually all nuclear models proposed since the famous Rutherford-Geiger-Marsden experiments led to the discovery of the nucleus in 1911.[2] The tenacity with which the problematic *p-e* model was adhered to is related to the inevitability of its adoption. Before we consider the details of the various problems related to this model, something must be said as to why its adoption appeared so self-evident to so many at the time (Pais, 1986, p. 231) and why this model was so deeply entrenched in the minds of physicists of that era (Stuewer, 1983, p. 32); or, as Brown (2004, p. 309) has put it, why electrons in the nucleus were taken for granted until the discovery of the neutron in 1932. This reconstruction will help us understand the mindset of the physicists discussed in the next sections.

A constitution model is expected to specify in a reductionist fashion the various elements of the target phenomenon and the relations among them. In the case of a model for the nucleus, a specification of the various elementary particles and forces must explain the observed properties: that these particles stick together inside the nucleus, that they allow for radioactive radiation,[3] and that each element has a specific mass and charge

[2] For an overview of this exotic assembly of often quite speculative models, see Stuewer, 1983, pp. 22ff.; Hughes, 1998, n. 17; Pais, 1986, pp. 230ff.. Despite their wide variety, these models all had in common that they presupposed the existence of electrons in the nucleus, and except for a few exceptions (e.g. Van Den Broek's 1913 model took α particles and electrons as the fundamental constituents), most of these models conjectured, already well before the experimental liberation of H nuclei (protons) from heavier nuclei (Rutherford, 1919), the existence of some kind of particle with positive elementary charge in the nucleus. These fundamental constituents were generally combined, however, into larger stable substructures such as α or $^4_2$He particles and other speculative entities (e.g. Rutherford's $^3_2$X particle (1920, pp. 392ff.), see Section 8.3). Although Rutherford complained of the excess in speculative models, he seems to have taken some part in it too (Hughes, 1998, p. 346).

[3] Around 1911, the peripheral electrons orbiting the nucleus were sufficient to explain

number.[4] At the time, few elementary particles were known. Since J.J. Thomson's discovery, the negatively charged electron was best-known, and its charge was considered to be the elementary unit of electrical charge. Regarding positively charged particles, until Rutherford's discovery of the nucleus there was no need to presuppose the existence of "corpuscules", because in Thomson's old "plum pudding" model the positively charged matter was spread uniformly within the atom. It was the insight that most of the atom is void, except for a dense material nucleus that concentrates the positive charge, that led naturally to the idea that nuclei are a kind of positively charged (composite) particles. In particular, two types of nuclei were well-known: the nucleus of the lightest element, hydrogen, and of the second lightest, helium, which had been identified as the constituent of α radiation. Finally only two forces were known at the time, electromagnetism and gravity, though the effects of the latter are negligible on an atomic scale.

As both α and β particles were observed in radioactivity, it was natural to assume that both were present in the nucleus. Yet as the internal mass of all elements was always an integer multiple – the atomic weight number – of the mass of the H nucleus, and not always of the He nucleus (which is four times heavier), it made more sense to take the H nucleus (which Rutherford called the "positive electron") as the fundamental nuclear constituent, a hypothesis first conjectured by Rutherford (1914) and later confirmed by his discovery of artificial disintegration and liberation of H nuclei from nitrogen nuclei (Rutherford, 1919). As electrons do not add up to the atomic weight number (their mass is only about $1/1000^{\text{th}}$ the mass of a proton), assuming the presence of that number of H particles was the only way to account for the weight of a nucleus. Yet as the nuclear charge is in general about half the nuclear weight, the presence of electrons in the nucleus seemed the only logical explanation to compensate for this positive charge. After all, did one not observe their emission in radioactive β decay? Moreover, the electromagnetic attraction between negative and "positive" electrons explained the stability of nuclei.

most thermal, optical, elastic, magnetic, and chemical properties of atoms; the only exception to this idea appeared to be the phenomenon of radioactivity: "Radioactive phenomena form a world apart, without any connection with the preceding phenomena. It seems therefore that radioactive phenomena originate from a deeper region of the atom." (Marie Curie, as cited in Pais, 1986, p. 223)

[4] These numbers were summarized in the table of Mendeleev, originally in a table in which the elements were ranked according to increasing atomic weight. In 1913, Van Den Broeck conjectured that the rank in the table actually matches the nuclear charge Z.

Seen in retrospect, this model is the simplest possible given one important ontological commitment. One had to consider elementary particles as truly fundamental, i.e. indestructible and permanent, much in the way that the ancient Greeks had regarded atoms as the smallest building blocks of the universe: they are never created and never destroyed, nor can they transform into each other. This ontological assumption seemed natural at the time (see Brillouin's testimony, quoted in Navarro, 2004, p. 451). The idea that the electrons found in β decay were created in the process itself seems not to have crossed these physicists' minds. While already in 1924 Eddington had spoken of particle creation (in the context of cosmic ray research), and Dirac, in 1928, became the first to adopt it in mathematical quantum mechanics (Bromberg, 1976), only in 1933 did it come to be understood that elementary particles could be created from and transformed into radiation, that they were unstable and decayed into other particles, and that they were not only the building blocks of matter but also the vehicles for nuclear interaction (Navarro, 2004, pp. 451-455).

In other words, the *p-e* model appeared self-evident: it was a simple, elegant and visual model that explained all the observed data (as the father of this model, Rutherford, preferred them to be (Hughes, 1998, p. 343)); no further existential assumptions were needed about unobserved particles; and the ontological commitment on which it was based was fully in line with the conception of elementary particles prevalent at the time. Any other model would have had to introduce radically new categories of particles and forces or drastically change existing concepts, which would require extensive theoretical elaboration or, at least, some experimental evidence that challenged the elegant and straightforward *p-e* model.

This apparent self-evidence of the *p-e* model explains scientists' relatively long adherence to this model, even as new discoveries gradually changed the theoretical framework, such as the first glimpses of the strong nuclear force in 1921 or the concept of wave-particle duality. Only by 1932, when the neutron was discovered, did physicists start to understand that the *p-e* model, by that time hugely problematic, was obsolete. Yet still, although Heisenberg was able to complete a new nuclear model constituted of neutrons and protons within just four months (Bromberg, 1971), it took several years for the neutron to be truly accepted as an elementary particle and not merely as a close proton-electron combination (Stuewer, 1983, pp. 46-56; Navarro, 2004, pp. 442-443; Fernandez and Ripka, 2013, pp. 263-270).

7.2.3 Problems for nuclear theory around 1927

Around 1927, the *p-e* model started to face various difficulties. Apart from the discovery of the continuous β spectrum, at least three other important problems were pointed out. These differed slightly from the β puzzle, as they were specific to the *p-e* model, whereas, in the case of the β puzzle, it was not clear where the problem was situated.

Two problems were pointed out by Ralph Kronig, an American physicist who suggested the electron spin before George Uhlenbeck and Samuel Goldstein (Stuewer, 1983, pp. 34-35). In 1926, he showed that, unless the spin of the various nuclear electrons exactly cancelled each other out, the magnetic moment of the nucleus would be much larger than the observed effects in spectral photography. Nuclear electrons should produce splitting levels of the same size as peripheral electrons (the so-called fine structure), whereas experimentally the magnetic moment of the nucleus produces effects at a smaller scale (the so-called hyperfine structure).

While some, like Owen Richardson, tried to explain this anomaly by assuming that nuclear electrons radiate part of their spin, Kronig, in 1928, found another, even more vexing anomaly. Observing the spectra of nitrogen nuclei, he discovered that these nuclei obeyed Bose-Einstein statistics, an indication that they have an integer spin. But both electrons and protons were known to have a spin of ½. Therefore, nitrogen nuclei, which according to the *p-e* model consist of 14 protons and 7 electrons, should have in total a half-integer spin and, therefore, obey Fermi-Dirac statistics – a contradiction. Kronig concluded that “in the nucleus protons and electrons do not maintain their identity in the same way as in the case when they are outside the nucleus” (cited in Stuewer, 1983, p. 35).

The final problem that troubled the nuclear electron hypothesis was the so-called Klein paradox. This paradox, formulated by one of Niels Bohr’s close collaborators at the end of 1928, was intended to attack the Dirac equation and its negative energy solutions. According to the Dirac equation, electrons confined to a region the size of the nucleus would have such a high probability of escaping (with negative energy) through the nuclear potential barrier that they could not be confined to the nucleus at all (Stuewer, 1983, p. 39). It is significant that this paradox was at the time mostly considered as a paradox for the Dirac equation, while according to our present understanding it is a problem for the *p-e* model, more particularly, for the presence of electrons in the nucleus.

7.2.4 Theoretical Attempts to Account for the β Puzzle

In the next sections, we will consider six hypotheses meant to account for the continuous emission spectrum of β decay, as experimentally demonstrated by Ellis and Wooster in 1927 and verified by Meitner in 1929. Each suggestion was in its own right an original idea that could provide the explanatory link to the β anomaly; all sketch in a more or less programmatic way how the initial idea might lead to a full explanation, as well as how the intended explanation should be understood in relation to the theoretical framework the researchers had in mind. Although three of the six hypotheses seem very similar, i.e. all three suggest to restrict the energy conservation principle, I still consider them as distinct hypotheses because they are formed differently and so connect their basic idea differently to the theoretical framework.

7.3 Ellis and Wooster: Non-Universality of the Quantum Postulate

At the end of their seminal paper, in which they experimentally demonstrated the continuous β spectrum, Ellis and Wooster offered “a simple hypothesis by which these facts can be reconciled” (1927, pp. 122-123). But, while their experimental results are today part of the canon of nuclear physics, these last pages have gained little, if any, traction in the physics community, mainly because – even if the Cavendish laboratory in Cambridge was not noted in the 1920s for its openness to developments in mathematical physics (Hughes, 1998) – it shows an almost surprising misunderstanding of the basic quantum postulate: it is taken to be a consequence of undisturbed classical particle motions. Although their idea did not leave any mark on the further course of events, it has some interest for our specific purposes.

Ellis and Wooster’s hypothesis is based on Rutherford’s satellite model of the nucleus. This was Rutherford’s version of the *p-e* model, which, by 1927, he had developed from an early explanation of his discovery of artificial disintegration (1919, pp. 589-590) into a highly sophisticated and structured visual model that enabled him to explain both the artificial disintegration of light elements and the radioactivity of heavy elements (Stuewer, 1986).

In the final version of this semi-classical model (Rutherford, 1926, pp. 370-371), which is the version referred to by Ellis and Wooster (1927,

p. 122), the nucleus is said to be composed of three distinct regions. Surrounding a positively charged inner nucleus, one could first distinguish, at a distance, a number of electrons, and then, at a further distance, a number of neutral satellites circulating the system. These neutral satellites were α particles (He nuclei) that had gained two electrons in a close bond, kept in stable orbits by polarization or magnetic forces.

Based on this model, Ellis and Wooster put forward the following hypothesis:

> There is no reason why the outer satellite region should not be quantised, and so give the possibility of ejection of alpha particles of definite energy, but yet the electronic region unquantised in the sense that the electrons have energies varying continuously over a wide range. (Ellis and Wooster, 1927, p. 122)

The pattern according to which this hypothesis is formed is very straightforward. On the one hand, it is observed that there is a qualitative difference between the discrete α spectrum and the continuous β spectrum; on the other hand, the employed nuclear model is cited to establish that the particles emitted in α and β decay first reside in different regions of the nucleus. Hence, via a simple instance of abductive reasoning, it becomes reasonable to suggest that the same qualitative difference applies to these two regions. As Ellis and Wooster (1927, p. 123) took the essence of quantum theory to be the quantized orbits model (nowadays generally referred to as the *Bohr-Sommerfeld atomic model* or the *old quantum theory*), they regarded the discrete emission spectrum of α decay as an indication that the neutral satellites (containing the α particles) orbit the inner nucleus in quantized orbits (analogous to the electrons in the old Bohr-Sommerfeld model). Based on these assumptions, Ellis and Wooster were able, in a quite straightforward way, to conceive the inner orbit containing the β particles as not quantized or continuous.

It was not that Ellis and Wooster were unaware that the quantum postulate, i.e. that there is a quantum of action, was to be universally applied on the atomic and subatomic levels. They simply did not regard this universality as a necessity, but rather as a contingent fact, though one which had to that point been consistently confirmed by experiment. This can be seen in the following passage, where Ellis and Wooster anticipate some suspicious frowns as they continue their discussion:

> It is interesting to enquire whether this picture of the free electrons in the nucleus existing in unquantised states is contrary to modern views. At first sight it would certainly appear to be so, but this is not necessarily the case. (Ellis and Wooster, 1927, pp. 122-123)

They explained this by stating that, for a particle to occupy a quantized orbit, it must be able to "describe many complete orbits without disturbance" (an analogy with the classical quantization of, for example, standing waves might have played a part here). While this condition might be fulfilled for the outer shell of neutral satellites, one can "scarcely expect undisturbed electronic orbits" so close to the positively charged nucleus.

In short, Ellis and Wooster did not consider the quantum postulate as a genuine postulate; for them, it was an emergent phenomenon that arises from particles describing stable orbits, which could be described by classical mechanics and electromagnetism (although they gave no account of how this exactly happens). Given this perspective, though they did not question the applicability of this postulate to existing atomic theory or to the outer nuclear layer, they did not feel the need to retain its apparent universality, which allowed them a straightforward solution for the β spectrum within the contours of Rutherford's nuclear satellite model.

As might be suspected, this idea had a very brief history. By 1927, quantum mechanics and the dominant Copenhagen interpretation had been developed and started to spread quickly (Kojevnikov, 2011; Heilbron, 1985). Gamov (1928) applied these new ideas to the nucleus and α decay, and succeeded in providing a quantitative explanation of the Geiger-Nutall relation between the decay constant and the energy of the emitted particle, something Rutherford's semi-classical qualitative satellite model was totally unable to do (Stuewer, 1985, see Section 7.6). Yet as Gamov's calculations confirmed Rutherford and Chadwick's experimental results in the Cambridge-Vienna controversy, Rutherford realized he had to accept Gamov's model over his own, even if this took him some time (Hughes, 1998; Stuewer, 1985, pp. 349-352). Also, Ellis appears to have shifted only slowly, as, in 1929, he was still defending this early view in a letter to Meitner (Jensen, 2000, p. 134).

7.4 Rutherford and Chadwick: Identical Nuclei with Varying Internal Energies

After Ellis and Wooster's paper, experimentalists generally shied away from advancing much speculation concerning what might explain the β spectrum. An exception can be found, however, in some remarks made by Ernest Rutherford and James Chadwick (1929) in an article on artificial disintegration, published in early 1929, in which they suggested that it might be possible that not every nucleus of a given element has the same internal energy.

Rutherford and Chadwick's article was certainly not an attempt to solve the β puzzle. Their goal was simply to report some unexpected results from their artificial disintegration experiments. They had discovered that, after inducing the artificial emission of protons from aluminum nuclei by shooting them with α particles, the energy of these emitted protons varied widely and continuously,[5] a surprising result that could not be ascribed to inaccuracies in the measurements. After verifying that this result was not caused by hitherto unobserved particles,[6] they declared that:

> The process of disintegration of an aluminium nucleus by an α particle of given energy is not exactly the same for each individual nucleus. [...] The variation in energy change must be due to variations in the internal energy either of the initial aluminium nucleus or of the final nucleus. (Rutherford and Chadwick, 1929, p. 190)

After expressing the need for further evidence and experiments, they repeated this hypothesis in their conclusion, adding that:

> This suggestion, [...], is supported by evidence from the natural disintegration of the radioactive elements. The disintegration

[5] Variation in the range of emitted protons had been observed earlier in artificial proton emission from nitrogen nuclei. But in their 1929 article, Rutherford and Chadwick argued that this earlier variation can be ascribed to the variation in momentum of the incident α particles. The variation they found now for aluminum nuclei, however, is considerably larger than the variation found earlier for nitrogen nuclei.

[6] Before coming to their conclusion Rutherford and Chadwick verified that no other particles such as "neutrons" were present. As Rutherford and Chadwick were trying to unravel the nucleus, they were prepared to find some hypothesized composite substructures, such as "neutrons", which Rutherford (1920, pp. 396-397) had conjectured to be close proton-electron combinations (see Section 7.8).

> electrons from β ray bodies are emitted with energies varying over a relatively wide range and in some cases at least, e.g. radium E, the energy balance is not restored by the emission of an appropriate amount of γ radiation. (Rutherford and Chadwick, 1929, p. 192)

In other words, by ascribing their results to the same cause as the similar continuous β spectrum, they suggested that the β puzzle might be explained by assuming that the internal mass or energy of an element can vary.

This idea is very radical: it infringes the *principle of identity* for chemical elements, which states that two atoms of the same element and isotope have identical properties. In their conceptual framework, however, the idea can be formed in a very straightforward way. Given the most basic assumptions about the process of disintegration – it is a nuclear process resulting in the emission of an observable particle – the observed continuity of emission energies can only have originated in a limited range of places: (1) either it was already present at the start of the process; (2) or it was created at the moment of disintegration; (3) or it entered somewhere between the disintegration and the place of measurement. Before I explain in more detail why Rutherford and Chadwick preferred the first option, let us take a closer look at how this disjunction is formed.

At first sight, the disjunction seems to be the result of an elementary abductive reasoning step, which can be modeled by existing logics for abduction.[7] But the disjunction does not mention just a couple of possibilities: anyone considering this disjunction (see e.g. also G.P. Thomson (1929, p. 405) or Pauli (Brown, 1978, pp. 22-24)) was convinced that it covered all (initial) possibilities. It was, in other words, an exhaustive disjunction. This relates to how these physicists' knowledge of disintegration processes is structured: not as a set of propositions (the building blocks for logics), but as a coherent spatiotemporal model of the process, which synthesizes their knowledge.[8] The basic assumptions mentioned above constitute the outline of such a model, which can be represented visually by drawing

[7] See, for instance, the logics for abduction developed in the adaptive logics framework (see Part II).

[8] In this chapter, I understand models as they are commonly understood in the philosophical literature on the use of models in science, as "a representation of a system with interactive parts and with representations of those interactions" (Nersessian, 2008, p. 12). These imaginary functional or structural analogues of the target phenomena allow for determining future states by mentally simulating the model by means of the interactive parts. In the case of

the experimental setup or by a more abstract sketch. In this type of process model, one can derive a purportedly exhaustive disjunction of possible origins for a property observed at the end of the process by covering the possibility of each spatiotemporal region in the model.

This is a pattern of abduction which draws on our intuitions about causal processes. In Salmon's terminology (1984), a characteristic observed at the end of a causal process must either have been uniformly present throughout the process, or else introduced into the process as a mark by means of a single local interaction at a certain space-time point, and the characteristic must have remained present at all subsequent stages until the space-time point of the observation. Exactly because the disintegration process is considered to be a causal one, these physicists assumed the exhaustiveness of the disjunction.

Of the three possible options, the third, i.e. that the continuous variety in energies appears in the model after the particle has left the nucleus, had been investigated by the experimentalists over the fifteen preceding years (see Section 7.2.1), a quest that ended with the Ellis and Wooster's caloric experiment and the consensus among Ellis, Wooster, Meitner and Geiger that the electron leaves the nucleus with a continuous spectrum.

This left two options open: either the variety in energy is present from the start, or it is introduced into the process at the exact space-time point of the disintegration. The first option is interchangeable with the thesis that nuclei of the same element can vary in internal energy. The second option is (in this model) equivalent to the assumption that energy is not conserved in a single disintegration (otherwise nuclei with fixed internal energies before and after cannot emit electrons with varying energies).[9] Hence, these physicists were caught between Scylla and Charybdis and had to give up either the principle of identity for chemical elements or the principle of energy conservation – both of them cornerstones of a physicist's worldview. This dilemma also explains the arduous focus of Meitner and others in earlier years to find a hypothesis that would fit the third option, and their perplexity when the caloric experiments of Ellis and Wooster excluded this

the model for disintegration, the visual picture represents the system, the various variables (which can be adjusted) represent the interactive parts, and the mathematical formulae (that specify the relations among the variables) represent the interactions. See also Part IV of this dissertation in which I examine the nature of scientific models in greater detail.

[9] Of course, this equivalence pertains only to this particular model of disintegration, which assumes that the nucleus emits a single particle. At the time, this was, however, the model that most physicists had in mind.

possibility.

This leaves us with the question of why Rutherford and Chadwick preferred to give up the principle of atomic identity, a route taken by no other protagonist in this history. In my opinion, the answer should be sought in the fact that Rutherford and Chadwick were in the first place experimentalists.[10] Experimentalists tend to have, as Franklin (1999, p. 97) has put it, an instrumental loyalty. While their ideal might be to look for "the best physics experiment in their field that can be done", and consequently build the appropriate apparatus, in reality they tend to look for the best experiment that can be done with their existing equipment (or with a minor modification). In that way, they recycle their expertise time after time, and become more and more experienced in employing the existing apparatus and its underlying models.

In the case of these Cavendish researchers, nuclear reactions were typically elicited by smashing small particles (mostly α particles) into nuclei, and then an effort was made to determine the properties of the remnants by observing them in electromagnetic fields – experiments and calculations that crucially depend on the conservation theorems. By performing this type of experiment over and over again, the theoretical models on which they were based became ever more deeply ingrained in their minds. As Franklin (1999, p. 149) has claimed that there are probably no anti-realists in the laboratory,[11] it is perhaps unavoidable that experimentalists form a deep belief in the veracity of their underlying models, i.e. that these models, which they manipulate mentally each time they perform physical experiments, have a true functional or structural analogy to reality. The only time Rutherford and Chadwick mention the conservation theorem in their paper of 1929 is when they explain the model of artificial disintegration on which their calculations and experiments are based. Also in their book on radioactivity (Rutherford, Chadwick and Ellis, 1930), though published at a time when they must already have heard of Bohr's proposals to limit energy conservation (Jensen, 2000, p. 160), they hardly mention the conservation theorem, except when they use it to explain their models. Clearly,

[10] See Hughes (1993, 1998) for a thorough discussion of the relation between the experimentalists at the Cavendish laboratory and theoretical physics.

[11] Franklin's claim is a strong version of *entity realism*, as it is proposed by Hacking (1983), which takes the manipulability of an entity as the criterion for belief in its existence. Franklin claims that experiments can also give us reasons to believe in the truth of some laws between these entities. I only use the descriptive part of his claim here, i.e. that experimentalists tend to form such beliefs.

due to their experimental bias, they did not question the validity of the conservation theorem, for it was an implicit and inherent part of the underlying models for the experiments they performed every day.

On the other hand, the internal or rest mass of an element can be measured. From an experimental point of view, it is perfectly conceivable that what was thought to be equal might turn out to allow for small variations. In fact, it was at the Cavendish laboratory that precisely a decade earlier Aston had discovered, with his newly devised mass-spectrograph, the isotopes suggested by Soddy: chemically identical elements with varying masses (Hughes, 2009; Fernandez and Ripka, 2013, pp. 166-171, Soddy, 1921, p. 369). Due to this long history of research on isotopes at the Cavendish, it must have appeared quite reasonable to Rutherford and Chadwick to expect that further variations at the level of individual nuclei might be measured. As a result, they supported the thesis of non-identity until 1932 despite the lack of any experimental proof, such as a varying lifetime for radioactive elements (Jensen, 2000, p. 161).[12]

7.5 G.P. Thomson: Non-Conservation of Energy as a consequence of Quantum Mechanics

As mentioned in Section 7.2.1, theoretical physicists, certainly on the continent, did not immediately appreciate the seriousness of the problem (Jensen, 2000, pp. 137-143). According to Pais (1986, p. 309), only one reference to the Ellis and Wooster paper can be found in all the literature of 1928: a short note from George Paget Thomson in *Nature*. In this first article (1928) and a more substantial article published a year later (1929), Thomson explained the β spectrum in terms of the non-conservation of energy for the emitted electrons.

The single most interesting feature of this account, which will turn out to be incoherent, is that Thomson described the non-conservation of energy

[12] They did try, however, to minimize this radicalism by situating the variation in the binding energy between the disintegration electrons and the nucleus, thus leaving the stable part of the nucleus identical (Rutherford, Chadwick and Ellis, 1930, p. 410). Opposition to their proposals was, given the lack of experimental evidence, very fierce; consider the following quote from Bohr's Faraday Lecture: "Although the corresponding variations in mass would be far too small to be detected by the present experimental methods, such definite energy differences between the individual atoms would be very difficult to reconcile with other atomic properties." (Bohr, 1932, p. 382)

in β decay not as an anomaly but as a result that was "to be expected on the new wave mechanics" (1928, p. 615).

Shortly before the fifth Solvay conference in 1927, a consensus had emerged concerning the mathematical equivalence of the formalisms of Born-Heisenberg and Schrödinger,[13] offering the field a new and versatile set of formal tools that could be applied to many open problems in theoretical physics and that would lead, in subsequent years, to a substantial list of successful explanations. Thomson regarded the experimental β anomaly as just one of the many puzzles to be solved by means of this new formalism.[14] Given the broad consensus and quick succession of solved puzzles, this expectation is understandable. It might even explain, to a certain extent, Pais's observation that the severity of the β puzzle was not directly appreciated: in these first years of quantum mechanics, the number of puzzles that could be addressed was still large indeed and the frontiers of formalism's application were still vague. As such, it was not immediately clear which puzzles would turn out to be a challenge for the new theory. In the case of the β spectrum, this took at least a year.

Thomson's account is clearly in contradiction with the *orthodox* or *Copenhagen interpretation* of the quantum formalism,[15] as he did not accept two

[13] See for instance the discussion between Heisenberg and Schrödinger after the latter's talk at the 1927 Solvay conference (Bacciagaluppi and Valentini, 2009, pp. 471-472); for a detailed history of the reconciliation of the two formalisms, see Longair (2013, ch. 15); for some recent discussion about the actual equivalence of the original formalisms, see Muller (1997); Perovic (2008). Of course, the agreement on the formalisms' equivalence was only a footnote to the real discussion at the Solvay conference, which concerned the interpretation of the quantum formalism (for an introduction, see Heilbron, 1985; Bacciagaluppi and Valentini, 2009; Mehra, 1975). Therefore, and due to the heterogeneity of the group of physicists involved, it would maybe be premature to label this episode as the installment of a new Kuhnian (1962) paradigm (see also Beller, 1999, ch. 14 and Bokulich, 2006 for an interesting analysis of Kuhn's notion of a "paradigm" as reminiscent of Heisenberg's notion of a "closed theory", a concept Heisenberg used to mount his rather dogmatic defense of the Copenhagen interpretation of quantum mechanics). Apart from this discussion (see e.g. Massimi (2005) for a reading of this period that is more sympathetic towards Kuhn's ideas), I do think that Kuhn's description of "normal science" fits the period from 1927 onward, as it should not be forgotten that, while senior professors continued their interpretational debates, most contributions to the field were made by a large group of younger researchers that employed the new mathematical formalism to address a wide variety of problems in the field (Kojevnikov, 2011, pp. 346-348).

[14] G.P. Thomson, following C.G. Darwin, was attracted by Schrödinger's wave formalism mostly because it allowed, as both Thomson and Darwin believed, for more "mechanical" explanations (Navarro, 2010). In the present case, this preference for mechanical explanations would lead to his faulty assumptions.

[15] Recent scholarship has shown that the Copenhagen interpretation is a far less coherent

of its central theses: the *completeness* of the wave function and the *complementarity principle* (which, as originally formulated by Bohr (1928),[16] states that the principles of conservation of energy and momentum are complementary to the space-time description of elementary particles). Thomson claimed that "the conception of a particle in motion is almost meaningless unless it can be supposed to have a definite velocity at a definite time" (1929, p. 413), meaning that he disbelieves that the wave function – inherently probabilistic in nature – provides a complete description of the electron. In fact, he adhered to the *pilot-wave interpretation* of quantum mechanics: wave functions exist physically as accompanying or guiding pilot-waves of particles, while the particles themselves have a definite but "hidden" trajectory. The standard formulation of this interpretation, also known as the *de Broglie-Bohm interpretation*, manages to be empirically equivalent with the Copenhagen interpretation by restricting the epistemic access to this definite trajectory to what is known in a statistical way via the wave formalism – hence the hidden character of this trajectory. But Thomson was misled by this idea of definiteness, and tried to gain knowledge about the electrons' trajectories by other means: by stating that "the equality of the particles emitted and atoms transformed is exact and not statistical" (1929, p. 406), he took the principle of identity (in which he firmly believed, unlike Rutherford and Chadwick) to require that the properties of the hidden trajectories must be exactly the same for all emitted electrons. This is in clear contradiction with the (empirically adequate) orthodox interpretation, which takes this principle only to require that identical systems can be described by the same wave function; measurements of identical systems are still uncertain and distributed according to the probabilities specified by the wave function. Thomson, however, fallaciously inferred that "the initial velocity is the same in all cases." (1929, p. 415) – being apparently unaware that the attribution of a definite speed would prevent any knowledge about the position of these particles.[17] At the same

(Beller, 1996) and unified (Howard, 2004) view than has traditionally been thought, as Bohr's and Heisenberg's views on complementarity diverged quite seriously. This does not need to concern us here, as Thomson rejected any notion of complementarity.

[16] As Camilleri (2007) shows, Bohr originally conceived of this concept as the complementarity between the description of the stationary unmeasured state (for which conservation of energy and momentum applied) and the description of this state in terms of position measurements (a space-time description). It was only in the wake of his dispute with Einstein that he extended this view around 1935 to our current understanding of the complementarity principle in terms of mutually exclusive experimental arrangements.

[17] Heisenberg's uncertainty relations state that the uncertainty in position is inversely proportional to the uncertainty in momentum. Physically realistic quantum models allow, there-

time, he employed the Gaussian model for a free particle (a well-known exemplar of the quantum formalism that describes position probabilities in terms of moving Gaussian curves, also called wave packets) to describe the evolution of the electrons' wave function in time.[18] As the variance of moving Gaussians grows over time, this model predicts that the uncertainty for position measurements will rise proportionally. At this point, Thomson had no other option than to accept that the electrons (which are, on his pilot-wave view, spread according to this Gaussian distribution inside the wave packet) can speed up or down from their initial velocity to move to the front or back of the wave train. Hence, the energy of a single emitted electron is not conserved in free space.[19]

Despite his confused assumptions, it is for our purposes still interesting to investigate the pattern of discovery by which Thomson arrived at these ideas. The initial step in his reasoning is rather easy to retract: he found that by taking a Gaussian with appropriate parameters, one could get a "fair fit to Ellis and Wooster's result" (1928, p. 616). The visual resemblance between the somewhat skewed experimental curve of the β spectrum (Figure 7.1) and the mathematical shape of a Gaussian, which figures prominently in the quantum model for a free particle,[20] led him to assume that the emitted electrons in β decay behaved as free particles with the same wave function. Once he had formed this initial hypothesis via visual identification, he could then apply this model to the data and calculate the properties of this wave function: it should be in a large superposition of momenta, and have a rather small initial wave length that increased with time. Combined with his faulty assumption of the exact and equal initial velocity, this led him to his thesis of "straggling" electrons. The calculated

fore, for uncertainty in both momentum and position, treating a system as in a superposition of both momenta and positions.

[18] Thomson appears to make a peculiar categorical distinction between *velocity*, a property of particles which has a definite nature, and *momentum*, a property related to the wave formalism which can be superposed. As such, he takes the emitted electrons to have initially a definite velocity, while at the same time allowing them to be in a superposition of momenta (as prescribed by the Gaussian model of a free particle).

[19] The assumption that the electrons had at first a definite and equal velocity is the real problem in Thomson's reasoning. Pais's remark that "his conclusions resulted from inappropriate manipulations with phase velocities and group velocities" (1986, p. 312) refers only to the consequences of this initial confusion. Bromberg traces Thomson's mistake to the fact that he employs a pilot-wave model (1971, p. 311), but this assumption would not be problematic if he adhered to, for instance, the de Broglie-Bohm interpretation.

[20] Thomson refers to a presentation of this model by C. G. Darwin in 1927, but the Gaussian model for a free particle is still a textbook exemplar of quantum mechanics.

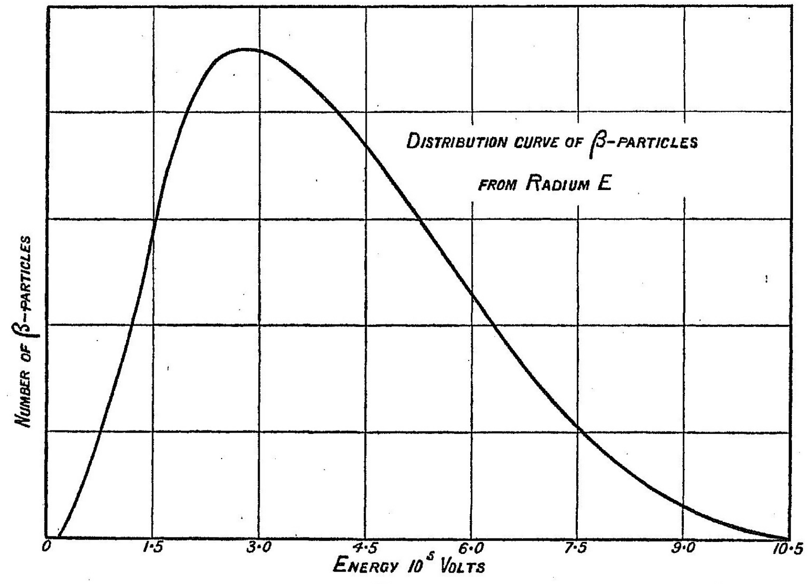


Figure 7.1: The experimental β spectrum visually resembles a Gaussian distribution. (reproduced from Ellis and Wooster (1927, p. 111), permission granted by Royal Society Publishing)

small initial wave length – he compared it to the sound pulse produced by a firing gun – allowed him to explain why non-conservation was observed only for β decay and not for other forms of decay, as the wave length he calculated for β decay was much smaller than other observed wave lengths in those days.

Thomson's account is a clear example of how the counterintuitive aspects of quantum mechanics misled even renowned physicists in the early years.[21] Willing to embrace the new formalism and maybe blinded by its attractive fruitfulness, Thomson thought it also contained the key to the β puzzle. In order to solve this puzzle, he actually used a very straightforward pattern of discovery: the visual recognition of a well-known mathematical model in the experimental data. Despite the coherence of his reasoning, he unfortunately relied heavily on an assumption based on classical intuitions, which made his contribution inconsistent.[22] However, the

[21] G.P. Thomson received a shared Nobel prize for physics in 1937 precisely for his work on the wave character of the electron.

[22] According to Navarro (2008, 2010), part of the trouble for his transition to quantum

idea that energy is not conserved in β decay would prove persistent.

7.6 Bohr: Non-Conservation of Energy as part of a Theory for Elementary Particle Constitution

The essence of Niels Bohr's stance on the matter is generally reduced to his embrace of the idea of energy non-conservation (e.g. Franklin, 2001, p. 68; Pais, 1986, p. 309). Yet scholarly work by Bromberg (1971, p. 309) and Jensen (2000, ch. 6) shows that we cannot evaluate Bohr's suggestion independently of the much broader scope he had in mind. By 1929, Bohr (and Heisenberg) became convinced that nuclear and atomic systems differ profoundly, and that a new theory must be constructed to address the various problems at the nuclear scale – a theory without energy conservation. By the time Bohr felt assured enough to first publish his ideas, however (Bohr, 1932), the β puzzle and the other nuclear problems had already polarized the field between his supporters and opponents. As the criticisms of the latter were aimed particularly at his views about energy non-conservation, Bohr seemed to stress this thesis in a more autonomous way beginning around 1932.

To understand why Bohr was on the lookout for a new physical theory, we have to trace his views to his first attempts to solve the β puzzle. These can be found in an unpublished, programmatic note written in June 1929,[23] which is particularly interesting for our purposes because, taken together with some short remarks in his correspondence, it displays the formation of his ideas. He starts this note with a reference to Thomson, whom he credits for connecting the idea of a limitation of the energy principle and the β puzzle. But, in his well-known gentle way,[24] he informs

mechanics can be related to the old continuous and aetherial worldview of his father, J. J. Thomson, and his classical Cambridge training, influences he struggled considerably to get free of. It took him until 1930 to come to "understand that the new physics was totally alien to the old notion of explanation by way of mechanical models." (Navarro, 2008, p. 250)

[23] This note is included in Vol. 9 of the Bohr Scientific papers (Bohr, 1929/1986). In the introduction to this volume, Peierls (1986) states that the content of this note must be "in substance" the same that he sent to Pauli on 1 July, accompanied by the words: "The other is a little piece about the beta-ray spectra, which I have had in mind for a long time, and which has been typed in the last few days, but I have not yet made up my mind to send it off, since it yields so few positive results and has been written so sketchily."

[24] Bohr's contemporaries did not always understand his well-known hesitation to use confrontational language; consider the start of Pauli's answer to this 1929 note: "It already starts so depressingly with a reference to the nonsensical remarks by G.P. Thomson, and from this the people in England will only draw the erroneous conclusion that you regard these remarks

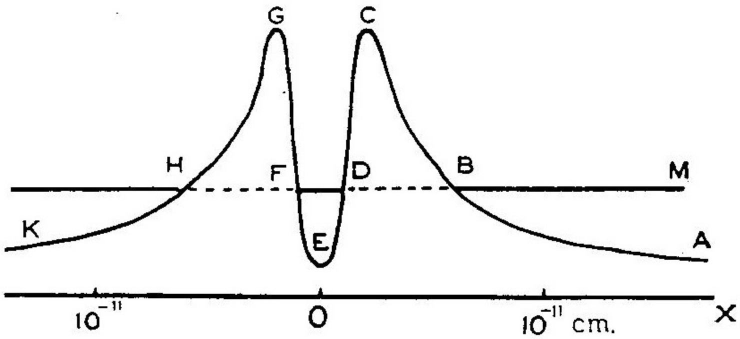


Figure 7.2: Electrons with energy M and captured between F and D have, according to quantum mechanics, a non-zero probability of "tunneling through" the nuclear potential barriers HGF and DCB without changing their energy level. (reproduced from Gurney and Condon (1928, p. 439), permission granted by Nature Publishing Group)

us that Thomson's view is wrong and resulted from a misunderstanding of the complementarity principle. Bohr considered this principle to be a crucial insight of quantum theory, a view he saw confirmed by the successful explanation of α decay by Gurney and Condon (1928) and, independently, by George Gamov (1928) (Stuewer, 1985).[25]

The reason why Bohr deemed the α decay explanation so "striking"– it was the first explication of a quantum tunneling phenomenon – is that it used the quantum formalism and the complementarity principle to explain a curious aspect of radioactivity, i.e. that the disintegration time of a nucleus is independent of its previous history, and subject only to a fixed chance. On a classical account, this is inexplicable: either the energy necessary to overcome the nuclear binding energy remains constant (in which case there would be either instant disintegration or no disintegration at all) or else it changes over time (in which case disintegration would be dependent on the previous history of the nucleus). But quantum mechanics allows for the description of another type of behavior. To obtain this, Gurney, Condon and Gamov modeled the nucleus by a constant potential well (Figure 7.2; the well is formed by the peaks G and C), in which an α

as important." (as cited in Peierls, 1986, p. 5)

[25] Gamov stayed at the Bohr Institute in Copenhagen during the academic year 1928-1929, where he unsuccessfully tried to apply the same style of reasoning to the β spectrum; this failure is one of the factors that led Bohr to consider the β puzzle.

particle, having energy M, is captured in an orbit between F and D, or, in other words, its energy is lower than the potential energy peaks (i.e. the energy required to overcome the nuclear binding energy). Classically, this particle cannot escape without external energy, but in quantum mechanics (the measurement of) the position of a particle is subject to a certain uncertainty. This means that there is a non-zero probability that the particle, confined between F and D, is actually located past the potential energy barrier (e.g. on the right-hand side of B) while maintaining the same energy level M. In other words, the particle has “tunneled through” the potential peaks raised by the nuclear binding energy. As Bohr noted, this is a “particularly instructive example” (1929/1986, p. 87) of the complementarity between the principle of energy conservation and the space-time description in terms of position measurements. In fact, it would be exactly this satisfying explanation that led Bohr to think that β decay could not be explained by quantum mechanics, but that a new theory was needed.

This α decay explanation also led to a change in the predominant conception of radioactivity. As we have seen in the earlier work of Thomson (1929, p. 412) and Ellis and Wooster (1927, p. 123), radioactivity was often considered to be a violent explosion in which a particle is hurled away. But Gurney and Condon say that it is better to change this in light of the present explanation and speak rather of particles that are “slipping away” (1928, p. 439), while Gamov speaks of “leaking” particles (1928, p. 805). This new metaphor for radioactivity led to drastic changes in the predominant conception: while radioactivity had long been regarded as an abrupt change in the history of the involved particles, including possible deformations or alterations, the new image of leaking drops suggests that it is a purely mechanical process and so must be described as such.

These observations allow us to understand why Bohr put forward the idea of a new theory. If we follow the new metaphor and take radioactivity to be a purely mechanical process, β decay should be modeled analogously to α decay, because the mechanical constraints are similar: a decay ratio which is fixed, a single observed and identified particle, and a change in nuclear constitution that conserves total atomic mass and charge. But as Gamov (1928) showed, the quantum tunneling model, which provides a successful explanation of α decay, also allows for a calculation of the rate of radioactive decay as a function of the velocity of the emitted particles (the so-called Geiger-Nutall relation). In other words, the decay rate and the energy of the emitted particles are directly related. If this is so, the continuous β spectrum would also force the rate of β decay to vary continuously amongst the different nuclei, which contradicts the observed fixed decay rate. Hence, Bohr concluded that:

> The existence of a well-defined rate of decay of β ray disintegrations would exclude any simple explanation of the continuous β ray spectra based on the ordinary ideas of wave mechanics. (Bohr, 1929/1986, p. 87)

Bohr regarded the quantum mechanical framework as inapt to explain the β decay observations. Another theory was needed, and in fact a conceptual niche presented itself quite directly: if β particles do not "leak" from the nucleus, then they cannot be present inside the nucleus beforehand (otherwise quantum mechanics and the quantum tunneling model would apply); and if they are not present in the nucleus beforehand, this means that the β particles are created in the process of decay itself – a type of behavior for which no theory had yet been formulated. But before Bohr could substantiate this new theory for the constitution of elementary particles, he realized that his conclusions – apparently logical consequences drawn from quantum mechanics and the observational data – presented a severe challenge to the *p-e* model, for which the presence of electrons in the nucleus is a basic assumption. First, then, he had to address this theoretical conflict.

It is well-known that Bohr approached this kind of theoretical conflict by scrutinizing the concepts involved and specifying how they apply to the experimental observations.[26] It is this style of reasoning that Heisenberg hints at when he describes that Bohr's insight was not so much

> [...] a result of a mathematical analysis of the basic assumptions, but rather of an intense occupation with the actual phenomena, such that it was possible for him to sense the relationship intuitively rather than derive them *[sic]* formally. (Heisenberg, 1967, p. 95)

More particularly, we can describe Bohr's pattern of reasoning as follows: by analyzing the meaning, preconditions and implications of the con-

[26] There exists a vast literature on Bohr's work and philosophy. For an introduction to the role of (classical) concepts and language in his reasoning, see Howard (1994); Favrholdt (1994); Bokulich and Bokulich (2005); on scrutinizing the conditions of the various concepts, see Favrholdt (1994, pp. 83, 94-95); Folse (1994, pp. 134-137); Camilleri (2007, pp. 520-524); on the importance of experimental observations, see Shomar (2008); Tanona (2004, p. 685).

cepts involved, he could identify which minimal conceptual assumptions were needed to describe the experimental data. In this way, he could, by restricting the meaning of these concepts, carve out the necessary conceptual space to resolve the contradicting elements – if his attempt was successful. Of course, this pattern was not a straightforward algorithm that Bohr could easily execute: explicating the many, often merely intuitive aspects of and assumptions related to the meaning of a concept was a painstaking process, in which Bohr succeeded only through numerous discussions with his collaborators, friends and visitors to his center. The most important result he reached in this process is his famous complementarity principle (Camilleri, 2007, pp. 520-524; Shomar, 2008, pp. 329-334): faced with the apparent contradiction between energy conservation and uncertainty of position, he realized that the idea of energy conservation makes sense only for isolated states (which, as such, are not observed), while the meaning of the concept of position inevitably involved measurement, i.e. the observation of this position. Hence, we obtain complementary descriptions for the same phenomenon by limiting the contradictory concepts in such a way that they can co-exist. As Heisenberg notes:

> This concept of complementarity fitted well the fundamental philosophical attitude which he had always had, and in which the limitations of our means of expressing ourselves entered as a central philosophical problem. (Heisenberg, 1967, p. 106)

This preoccupation with the meaning of concepts is indeed a constant in Bohr's writing, such as when he described the task of physicists as:

> [...] not to penetrate in the essence of things, the meaning of which we don't know anyway, but rather to develop concepts which allow us to talk in a productive way about phenomena in nature. (Bohr in a letter to H. Hansen, as cited in Pais, 2000, p. 23)

Bohr employed this same pattern of reasoning in the present case. Scrutinizing the role of electrons in the *p-e* model, he realized that they are needed only to ensure the correct atomic charge and the electromagnetic attraction that keeps the nucleus together.[27] For these roles, the only prop-

[27] From 1921 on, serious doubts arose concerning whether it was the electromagnetic force that kept the nucleus together. Still by 1928, no valid alternative had been proposed (Pais, 1986, p. 240).

erty of (nuclear) electrons that one has to assume is that they have a negative elementary charge – a non-mechanical aspect. On the other hand, the many problems related to the presence of electrons in the nucleus (the β spectrum, the spin of the N nucleus, the absent total magnetic moment, the Klein paradox) all have to do with the mechanical properties of these electrons: their momentum, energy and spin. These problems led Bohr to conclude that :

> The behavior of electrons bound within an atomic nucleus would seem to fall entirely outside the field of consistent application of the ordinary mechanical concepts, even in their quantum theoretical modification. (Bohr, 1929/1986, p. 88)

By separating the mechanical and electric properties of (nuclear) electrons, Bohr was able to resolve this conflict. On the one hand, electrons do not exist as individual particles inside the nucleus; only their total charge (a multiple of the elementary charge) exists and is somehow distributed inside the nucleus to ensure the *p-e* model. On the other hand, it is only in the process of β decay that an electron is created as a "dynamical individuality" (Bohr, 1929/1986, p. 88), while a negative unit charge attaches itself somehow to this newly formed particle. By restricting the assumptions about electrons in both cases, Bohr carved out conceptual space between the two existing theories for a yet to be constructed theory of "the constitution of elementary electric particles" (p. 87). For him, it was clear that this new theory could only account for the various energetic puzzles surrounding the electron concept – Bohr was at the time also puzzled by the classical problem of the infinite self-energy of electrons, and even connected the energy production in stars to this newly projected theory (see Section 9.4) – if it was not subjected to the principle of energy conservation.

To this analysis, we can add a further observation. As Pais (1986, p. 312, n. 20) and Jensen (2000, p. 149) have already remarked, this proposal is not related to the earlier BKS theory (Bohr, Kramers and Slater, 1924), in which he had already proposed the idea of energy non-conservation in order to remedy the old quantum theory.[28] In fact, as Darrigol

[28] For a historical introduction to the BKS theory, see Darrigol (1992, ch. 9) and Longair (2013, pp. 194-197). It is common in the literature to suggest a relation between Bohr's ideas of 1924 and those of 1929 (e.g. Franklin, 2001, p. 68). Lakatos (1970, pp. 168-173)) regards the non-conservation of energy even as the central thesis of a research program that ran from 1924 (the Bohr-Kramers-Slater paper) until 1936 (the Shankland experiments).

(1992, p. 214) has noted, Bohr had come upon the idea to limit the conservation of energy and momentum even earlier. But it has not been sufficiently stressed in the literature that at each of these three times there was a different reason to consider the non-conservation of energy. Previous to the BKS paper, Bohr privately held the opinion that the idea of momentum conservation would be impossible to reconcile at a micro level with discontinuous quantum jumps (Darrigol, 1992, p. 214). In the BKS paper, which presented a probabilistic theory for the first time, the energy conservation theorem had to be sacrificed to ensure the statistical independence of the atoms:

> It may be emphasized that the degree of independence of the transition processes assumed here would seem the only consistent way of describing the interaction between radiation and atoms by a theory involving probability considerations. This independence reduces not only conservation of energy to a statistical law, but also conservation of momentum. (Bohr, Kramers and Slater, 1924, pp. 792-793)

Energy conservation would imply that each time an atom emitted a quantum of energy, another atom would absorb this quantum – a contradiction with the assumption that these processes are statistically independent. Bohr acknowledged that this theory was mistaken following the 1925 experiments by Compton and Simon (Franklin, 2001, pp. 65-68). The reason why he proposed energy non-conservation for the third time in our case study in 1929 is clearly different, and this time, as Gamov noticed, “he now goes even further and stresses that the energy need not be conserved even in the mean” (as cited in Jensen, 2000, p. 149).

However, there is a parallel to be observed. Clearly, maintaining the energy theorem was not high on Bohr’s list of priorities. He was willing to sacrifice it if necessary, and, although his previous suggestions to abandon it had been unsuccessful, he still thought that this revision could help him solve the many puzzles of nuclear theory.

This brings us to our main questions: why was Bohr so willing to withdraw the energy conservation theorem? And why did he not take seriously Pauli’s suggestion to acknowledge a new particle (see Section 7.8) when the debate about the β spectrum narrowed to these two suggestions around 1932? In specifying his method of reasoning, we saw that Bohr gave absolute priority to the observational data, which he tried to account for using the physical concepts at his disposal. He had always been rigorous in this. In the BKS paper he stated, concerning photons, that:

> Although the great heuristic value of this hypothesis [...], the theory of light-quanta can obviously not be considered as a satisfactory solution of the problem of light propagation. (Bohr, Kramers and Slater, 1924, p. 787)

At a moment when many physicists were starting to accept the physical existence of Einstein's photons (certainly from the 1922 Compton experiments onwards), Bohr continued to disbelieve in their existence until the mid-1920s (Stachel, 2009), due to, as he put it, a lack of experimental observation. So it should be no surprise that Bohr was also reluctant to assume the existence of another new particle, of which, by 1929, there was not the slightest experimental trace.

This analysis concurs with Shomar's (2008) characterization of Bohr as a "phenomenological realist", i.e. someone who has a realist position about low-level phenomenological models, which are seen as a kind of theoretical descriptions of reality, but an instrumentalist position about high-level theories, which are seen to have merely the status of conceptual tools. This explains in the same way why Bohr was hesitant to accept the reality of new particles (phenomenological models that link very closely to experimental observations), but willing to sacrifice the energy conservation principle (a high-level theoretical principle).

This leaves us with the question of why the energy conservation theorem should have been the law that needed to be sacrificed, and not another law or principle. In my opinion, the main reason was that Bohr projected that the new theory he had in mind could address all energetic problems at once, and that therefore it would be better at first not to include such a strong theorem about which he already, earlier, had doubts.

This summarizes his views on the matter. Unlike Thomson, Bohr realized that this problem was beyond the scope of the new quantum formalism. This did not, however, lead him to doubt the new quantum mechanics; rather he thought that the best one could do in describing phenomena was to assemble a patchwork of various theories, joined by correspondence principles and common concepts and stripped of contradictions by specifying and restricting these concepts. Still, one should also keep in mind that Bohr had hardly any "positive results" to support this new theory, and his hesitation to put these ideas in print demonstrates that he was aware of the radical nature of his suggestion.

7.7 Heisenberg: Non-conservation of Energy as part of a Second Quantization at the Scale of the Nucleus

We will touch only briefly on Werner Heisenberg's ideas, mainly because his stance on the matter was never published and he held it for only a few months. Also, his idea, however short-lived, must be understood in relation to his broader research project at the time: to cope with the many infinities and paradoxes associated with the construction of a relativistic quantum electrodynamics (QED), especially at distances on the order of the size of the electron (Cassidy, 1981; Rueger, 1992).

Heisenberg's suggestions concerning the β spectrum were brought to light by Bromberg (1971) and recorded in two letters to Bohr, dated in February and March 1930, available at the Niels Bohr Archive. In these letters, as Bromberg tells us, Heisenberg proposed to construct a mathematical lattice world with grid cells of nuclear dimensions. If the scale of the system was large with respect to these cells, normal quantum mechanics would apply; but within these cells, phenomena obeyed new laws. He obtained these by turning the Klein-Gordon differential equation into a difference equation tailored to these new cell dimensions. This had as a consequence that the energy of particles became periodically dependent on the wave number. By further supposing that particles near the maxima behaved as electrons and those near the minima as protons, Heisenberg constructed a first picture of the nucleus without "real" nuclear electrons. The price of this idea was high: within the cells, neither charge, energy nor momentum was conserved, which made him ask Bohr whether he regarded "this radical attempt as completely crazy" (as cited in Bromberg, 1971, p. 325). Yet, as we have seen, Bohr was on much the same track (except for charge non-conservation) and also on the lookout for a new nuclear theory. Still, Heisenberg's idea was short-lived: already in April 1930, after a meeting with Bohr, Pauli and Gamov in Copenhagen, he jettisoned it because he realized that the introduction of a fixed cell grid length could not be relativistically invariant.

The key to understanding how Heisenberg arrived at such a radical theory can be found in the following passage from a letter he sent to Bohr in December 1929, in which he commented on Ellis's findings that protons emitted in the artificial disintegration of nitrogen also showed a continuous spectrum, a result which later turned out to be mistaken:

> I find Ellis' claim that also the H particles from the disintegration of N show a continuous spectrum dreadful; for how shall one then understand the sharp α ray spectra? (Heisenberg, as cited in Jensen, 2000, p. 148)

As he later acknowledged in a personal interview with Bromberg (1971, p. 328), the many problems associated with nuclear electrons in particular and with the nucleus in general made him wonder why the α spectrum was the lucky exception. In other words, instead of β decay, he thought that α decay might be an anomaly; in doing so, he presupposed at the same time the existence of a nuclear theory, to which α decay was the anomaly and which explained all the other nuclear problems. He especially hoped that it could solve the infinite self-energy of a point electron, which proved such a hurdle for QED (Cassidy, 1981, p. 8). By thus inverting the anomalies, he reached in a much more straightforward way the same conclusion as Bohr, i.e. that a new (nuclear) theory was needed. But unlike Bohr, who expected this to be a theory of elementary particle constitution, Heisenberg foresaw a more general theory of all nuclear phenomena (which would reveal α decay to be the real exception).

Heisenberg observed that the main difference between quantum mechanics and the nuclear problems was one of scale. At the same time, he realized that the scale of the nucleus was of the same dimension as both the classical electron radius (which proved in QED to be the scale of the electron, below which the theory diverged to infinity) and the Compton wavelength of the proton (as the proton was the heaviest particle known at the time, this was the smallest length in which the uncertainty relations allowed a particle to be localized). These coincidences must have led him to hardcode this dimension as the dimension of grid cells, for which he had the freedom to alter laws within their boundaries.

This quantum-nuclear divide is clearly constructed in a way analogous to the classical-quantum divide: processes of large scale with respect to the pivotal distance can be described by the former theory, while phenomena at the scale of this distance obeyed new laws.

At this point Heisenberg needed some formal tool to start exploring this new level. He did this using the relativistic Klein-Gordon equation for spin-less particles[29] and hardcoding his cell axiom directly into this equation by changing the differential equation into a difference equation. It was this formal “point of attack”, as he called it himself, that led him to deduce the various results he proposed (Bromberg, 1971, p. 328).

Interestingly, Heisenberg’s strategy, which is part of what Cassidy (1979) has called his “professional style”, had proven fruitful before: the results of his first paper on matrix mechanics in 1925 were obtained via a similar procedure, i.e. using the correspondence relations and hardcoding the model of virtual oscillators in these formulae (MacKinnon, 1977; Miller, 1984/1986, pp. 135-138). Yet this time, his results were less lasting, and after realizing that his theory could not stand the test of relativity, he turned his focus away from nuclear physics, because he lacked a new formal point of attack (Bromberg, 1971, p. 329). This changed when the neutron was discovered in 1932, after which he needed only four months to construct the first proton-neutron model of the nucleus.

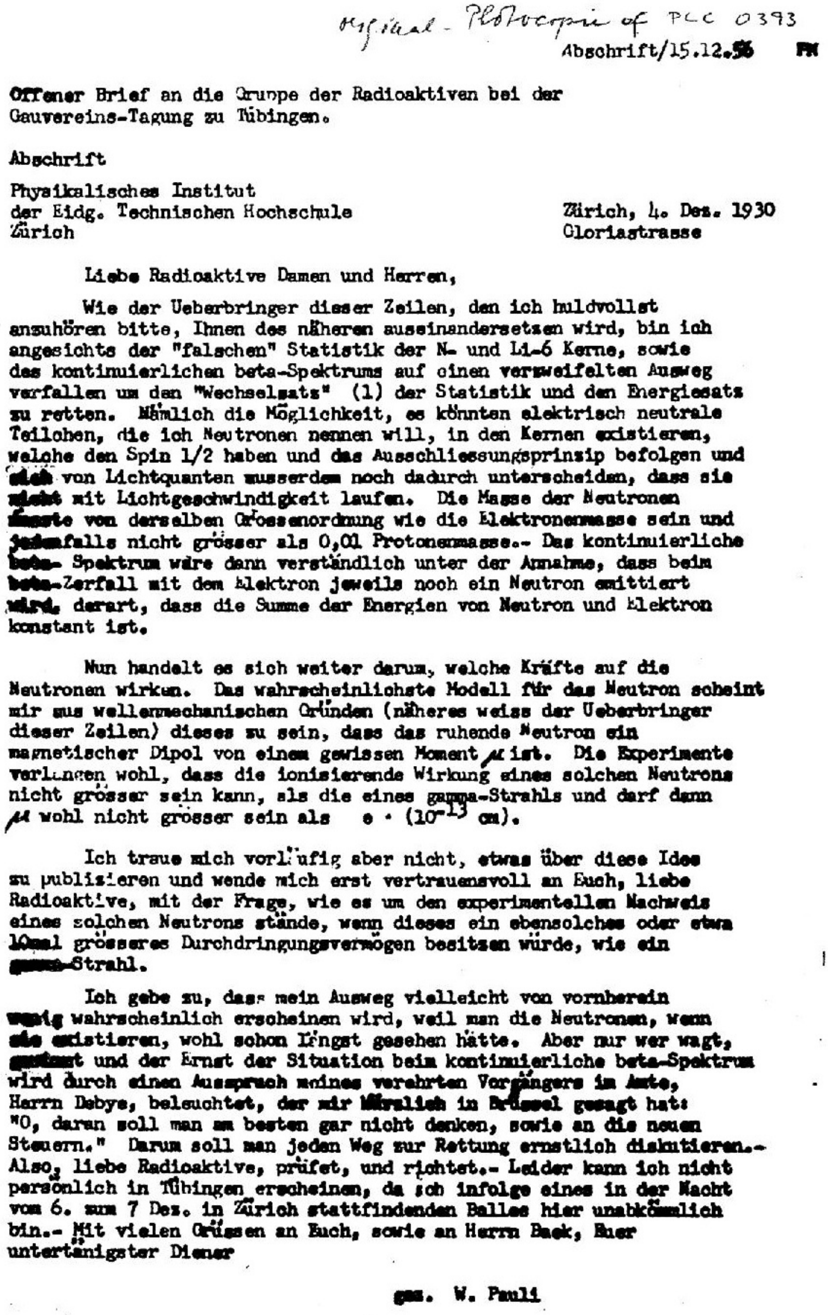
7.8 Pauli: a New Elementary Particle as a Nuclear Constituent

Wolfgang Pauli’s suggested solution for the β puzzle is quite famous because the canonical history of modern physics has equated the new particle he envisioned with our neutrino, making of Pauli’s idea a highly original feat of epiphany that provided the key to the β puzzle – a story spiced up by the anecdotal details from his original letter addressed to the *“Liebe Radioaktive Damen und Herren”* from Tübingen (Pauli, 1957/1964, p. 1316). This outsiders’ perception has, however, been successfully challenged: the particle Pauli first had in mind was in fact different from what is currently understood as the neutrino (Brown, 1978; Pais, 1986). Furthermore, while it is correct to credit him for hypothesizing an as-yet unobserved particle to solve the β puzzle, the idea he had in mind was not necessarily so new as is commonly thought. As I will show in this section, it could well have been an adaptation or variant of Rutherford’s original neutron idea.

Pauli’s famous letter, dated December 4, 1930 (see Figure 7.3),[30] in

[29] Heisenberg at the time, just like Bohr, had issues with Dirac’s infinite sea interpretation of the relativistic Dirac equation for particles with spin, which might be why he turned to the older Klein-Gordon equation for this theory.

[30] Pauli himself made this letter public in a lecture on the history of the neutrino (Pauli, 1957/1964, pp. 1316-1317). The idea must have come to his mind only shortly before this letter: the first known written reference to it is in the form of a letter from Heisenberg to Pauli



Original - Photocopie of PLC 0393

Abschrift/15.12.56 FN

Offener Brief an die Gruppe der Radioaktiven bei der Gauvereins-Tagung zu Tübingen.

Abschrift

Physikalisches Institut
der Eidg. Technischen Hochschule
Zürich

Zürich, 4. Dez. 1930
Gloriastrasse

Liebe Radioaktive Damen und Herren,

Wie der Ueberbringer dieser Zeilen, den ich huldvollst anzuhören bitte, Ihnen des näheren auseinandersetzen wird, bin ich angesichts der "falschen" Statistik der N- und Li-6 Kerne, sowie des kontinuierlichen beta-Spektrums auf einen verzweifelten Ausweg verfallen um den "Wechselsatz" (1) der Statistik und den Energiesatz zu retten. Nämlich die Möglichkeit, es könnten elektrisch neutrale Teilchen, die ich Neutronen nennen will, in den Kernen existieren, welche den Spin 1/2 haben und das Ausschliessungsprinzip befolgen und sich von Lichtquanten ausserdem noch dadurch unterscheiden, dass sie nicht mit Lichtgeschwindigkeit laufen. Die Masse der Neutronen müsste von derselben Grossenordnung wie die Elektronenmasse sein und jedenfalls nicht grösser als 0,01 Protonenmasse.- Das kontinuierliche beta- Spektrum wäre dann verständlich unter der Annahme, dass beim beta-Zerfall mit dem Elektron jeweils noch ein Neutron emittiert wird, derart, dass die Summe der Energien von Neutron und Elektron konstant ist.

Nun handelt es sich weiter darum, welche Kräfte auf die Neutronen wirken. Das wahrscheinlichste Modell für das Neutron scheint mir aus wellenmechanischen Gründen (näheres weiss der Ueberbringer dieser Zeilen) dieses zu sein, dass das ruhende Neutron ein magnetischer Dipol von einem gewissen Moment μ ist. Die Experimente verlangen wohl, dass die ionisierende Wirkung eines solchen Neutrons nicht grösser sein kann, als die eines gamma-Strahls und darf dann μ wohl nicht grösser sein als $e \cdot (10^{-13}$ cm).

Ich traue mich vorläufig aber nicht, etwas über diese Idee zu publizieren und wende mich erst vertrauensvoll an Euch, liebe Radioaktive, mit der Frage, wie es um den experimentellen Nachweis eines solchen Neutrons stände, wenn dieses ein ebensolches oder etwa 10mal grösseres Durchdringungsvermögen besitzen würde, wie ein gamma-Strahl.

Ich gebe zu, dass mein Ausweg vielleicht von vornherein wenig wahrscheinlich erscheinen wird, weil man die Neutronen, wenn sie existieren, wohl schon längst gesehen hätte. Aber nur wer wagt, gewinnt und der Ernst der Situation beim kontinuierliche beta-Spektrum wird durch einen Ausspruch meines verehrten Vorgängers im Amte, Herrn Debye, beleuchtet, der mir kürzlich in Brüssel gesagt hat: "O, daran soll man am besten gar nicht denken, sowie an die neuen Steuern." Darum soll man jeden Weg zur Rettung ernstlich diskutieren.- Also, liebe Radioaktive, prüfet, und richtet.- Leider kann ich nicht persönlich in Tübingen erscheinen, da ich infolge eines in der Nacht vom 6. zum 7 Dez. in Zürich stattfindenden Balles hier unabkömmlich bin.- Mit vielen Grüssen an Euch, sowie an Herrn Back, Euer untertänigster Diener

gez. W. Pauli

Figure 7.3: Transcript of the original letter in which Pauli suggested for the first time the existence of a hitherto unobserved particle. (Retrieved October 21, 2013, from `http://physics.stackexchange.com`)

which he presented his "desperate remedy", was written to a group of experimentalists – the most important among them being Meitner and Geiger – who were to hold a seminar in Tübingen three days later, which Pauli was unable to attend.[31] In this letter, Pauli hypothesized an electrically neutral particle, named the "neutron", which he conceived as a permanent constituent of the nucleus with spin of ½. Its velocity, he said, was somewhat below the speed of light and its mass was relatively small, on the order of the mass of an electron. This particle was to be kept in the nucleus by electromagnetic forces, and so must have a magnetic moment.

In the introduction to this letter, Pauli expressed his hope that the discovery of this particle would solve both the β puzzle and the anomalous spin of the nitrogen nucleus. In other words, just like Bohr, he tried to solve several nuclear puzzles at once. This might be the reason why his proposal had characteristics in common with both our present neutron (a neutral spin-½ constituent of stable nuclei) and our present neutrino (a very light spin-½ particle that carries away the remnant energy in β decay). It was not until the experimental discovery of the 'heavy' neutron by Chadwick in 1932 (which explained the spin of the N nucleus) that Pauli would finally consider his proposal – renamed the neutrino by Fermi – no longer as a nuclear constituent but solely as the key to the β puzzle (Brown, 1978, pp. 24-28).

Pauli was at first particularly hesitant about his ideas and well aware that the lack of experimental evidence could be held strongly against him. In a letter to Klein one week after his original letter, he wrote that:

> So, if the neutrons really existed, it would scarcely be understandable that they have not yet been observed. Therefore, I also do not myself believe very much in the neutrons, have published nothing about the matter, and have merely induced

three days earlier (Jensen, 2000, p. 153; Pais, 1986, p. 315).

[31] As the letter states, he was expected to attend a ball the night before – according to Pais (1986, p. 315), the Italian student ball – at which his presence was "indispensable". Pais spices up this story further by revealing that Pauli wrote this letter only one week after his first wife, to whom he had been married for less than a year, left him. Pauli seems further to have once referred to the neutrino as that "foolish child of the crisis of my life", which led Pais to stress the importance of this anecdotal evidence as follows: "I tend to regard Pauli's association between his time of personal turmoil and the moment at which he stated his new postulate as highly significant. Revolutionary steps were out of line with his general character." (Pais, 1986, p. 314)

> some experimental physicists to search in particular for this sort of penetrating particles (Pauli, as cited in Jensen, 2000, p. 154).

However, as he received a "positive and encouraging" answer to his original letter from Geiger – Pauli puts great emphasis on this support in his recollections (Pauli, 1957/1964, p. 1317) – and given the severity of the problems, he kept toying with his idea and started lecturing about it on a trip across America the next summer. In October 1931, while attending the first nuclear physics conference in Rome, he must have sparked Fermi to develop his own β decay theory (Brown, 1978, p. 27). However, until the discovery of Chadwick's neutron, Pauli's proposal remained a minority position, the majority of physicists being convinced by Bohr's ideas (Jensen, 2000, p. 155). Only after the 1933 experimental results of Ellis and Mott did Pauli finally allow the first printed publication of his – since the discovery of the neutron evolved – idea (in the report of the 7th Solvay Conference in 1933, reprinted in Brown, 1978, p. 28). Ellis and Mott's experiments favored Pauli's suggestion because they found that the β spectrum had a sharp upper limit, indicating that something (a particle) carried away the difference in total energy rather than that the electron energies were distributed randomly around an average emission energy (which would be the case in case of non-conservation of energy).

Let us now try to understand how Pauli originally came to his idea. Clearly, his motivation stems from a serious discontent with Bohr's thoughts concerning the non-conservation of energy. In his letter to Klein, quoted earlier, he made a more elaborate argument against Bohr's proposal in the form of a thought experiment:

> Imagine a closed box in which there is radioactive β decay; the β rays would then somehow be absorbed in the wall and would not be able to leave the box. [...] If the energy law thus would not be valid for β decay, the total weight of the closed box would consequently change. (This conclusion seems quite compelling to me.) This is in utter opposition to my sense of physics! For then it has to be assumed that even the gravitational field – which is itself generated by the entire box (including the radioactive content) – might change, whereas the electrostatic field, which is measured from the outside, should remain unchanged because of the conservation of charge. (Yet both fields seem analogous to me; that, incidentally you will re-

> call from your five-dimensional past.) (Pauli, as cited in Jensen, 2000, p. 153)

The five-dimensional past to which Pauli refers is his own early career in relativity. At the age of 21, Pauli had written a state of the art overview of general relativity, which impressed even Einstein (Pais, 2000, p. 215). Given this, although he does not mention it explicitly in this article, he must have been aware of Noether's theorems, which state the correspondence between conservation laws and the (differentiable) symmetries of fields. In fact, for this article about relativity Pauli made use of Klein's notes, which called attention to these theorems several times (Kosmann-Schwarzbach, 2010, p. 93). Seen from this perspective, Pauli found it unacceptable that the analogical treatment of the various field symmetries was broken.

This perspective differed significantly from that of most quantum physicists at the time. As Heisenberg recalls in an interview with Thomas Kuhn in the 1960s:

> Much later, of course, the physicists recognized that the conservation laws and the group theoretical properties were the same. And therefore, if you touch the energy conservation, then it means that you touch the translation in time. [...] But at the time, this connection was not so clear. Well, it was apparently clear to Noether, but not for the average physicist. (Heisenberg, as cited in Kosmann-Schwarzbach, 2010, p. 85)

Pauli was clearly ahead of his time. As one of the few protagonists in quantum physics, he adhered already to a modern ontology that considered particles and fields (with their symmetries) as the unifying ontological entities.[32] For him, conservation laws were not just empirical laws, but structural relations grounded in his ontology. As such, they could not be refuted by simple empirical observations.

This analysis concurs with De Regt's (1999) analysis of the heuristic methodologies of Pauli and Heisenberg. Based on their earlier work in the mid-1920s (on the Zeeman effect and matrix mechanics, respectively), De Regt interpreted the difference between their methods at the level of their personal philosophies: Pauli was an ontological realist whose opera-

[32] The present-day day view that even these two basic ontological entities coincide should, according to Weinberg (1999, p 241), be ascribed to Feynman.

tionalist methodology placed consistency with other theories and simplifying unification above empirical adequacy, while Heisenberg would be best described as a kind of pragmatist (although not fully anti-realist), whose principal aim was to forge mathematical theories that were empirically adequate, even if to do so he had to employ *ad hoc* strategies (see, for example, his suggestions concerning the lattice world in Section 7.7).

This ontological necessity of the conservation laws must have triggered Pauli to think over the nuclear problems himself:

> I tried to connect this problem of the nuclear spin and statistics with the other problem of the continuous beta-spectrum by the idea of a new neutral particle without abandoning the energy theorem. (Pauli, 1957/1964, p. 1316, my translation)

In essence, the energy conservation theorem is an equation in which the sum of the measurements before must be balanced with the sum of those taken after. The abnormal statistics for the N nucleus, too, are basically an unbalanced spin equation with the sum for the theoretically predicted particles on one side and the observed total spin on the other. Unbalanced equations cannot be balanced in so many ways: if one has confidence in the terms of the equation and their values, the only way to balance it is by adding to the picture something with appropriate values. In Pauli's case, as charge was already conserved, this meant an electrically neutral, spin ½ nuclear particle (in addition to the already present protons and electrons) that is, together with an electron, released in β decay with appropriate momentum and energy.

However, the idea of a new neutral particle was not new: Rutherford had already, in his Bakerian Lecture (1920), mentioned the possibility of "an electron to combine much more closely with the H nucleus, forming a kind of neutral doublet". Such a neutral particle, which was just one of the many speculative nuclear composite particles he suggested (Hughes, 2003, p. 362), would have "very novel properties": "it should be able to move freely through matter, and its presence would be difficult to detect" (1920, p. 396). Rutherford, at the time unaware of the strong nuclear force, thought that such particles were requisite to explain nuclear constitution:

> The existence of such atoms seems almost necessary to explain the building up of the nuclei of heavy elements; for unless we

> suppose the production of charged particles of very high velocities it is difficult to see how any positively charged particle can reach the nucleus of a heavy atom against its intense repulsive field. (Rutherford, 1920, pp. 396-97)

It is my thesis that it is well possible that Pauli, realizing that the presence of a neutral particle could restore the conservation laws, thought of Rutherford's idea and understood that it could, with slightly modified properties, offer a solution to the problems he was working on. In the remainder of this section, I will develop several arguments for this speculative thesis.

First, Rutherford's "neutron" idea remained very much alive before Chadwick's discovery and was part of the research program of the Cavendish laboratory (Stuewer, 1983, p. 27; Fernandez and Ripka, 2013, p. 253; for a list of early references to the neutron, see Stuewer, 1983, n. 150); this notwithstanding that Chadwick's (1932, p. 698) claim – namely that the particle he discovered was precisely the particle discussed by Rutherford in his Bakerian lecture – was also an attempt to gain some prestige for the Cavendish laboratory in the field of theoretical nuclear physics and raise some much-needed funds (Navarro, 2004, p. 443; Hughes, 2000, p. 46).

Consider for instance Rutherford and Chadwick's article of 1929, discussed in Section 7.4. There, they proposed a hypothesis for the observed continuous spectra (i.e. that identical nuclei have varying internal energies) only after

> [...] the liberated protons were examined in order to test whether any particles other than protons were present; for example, whether the particles of very long range might possibly be 'neutrons'. (Rutherford and Chadwick, 1929, p. 189)

This particular article is mentioned by Bohr to Fowler in a letter concerning the β problem (Jensen, 2000, p. 147), and Heisenberg, too, mentioned these experiments in a letter to Bohr (Jensen, 2000, p. 148). It is very plausible that Pauli, who was in both Bohr's and Heisenberg's inner correspondence circle (e.g. Bohr's first attempt to solve the β spectrum was sent to Pauli) and who regularly visited Copenhagen to discuss the problems of the day, knew this article, written only one year earlier, or, at the very least, was familiar with the quest of the Cavendish Laboratory to find a neutral nuclear constituent or neutron. This puts the following quote from his original letter in perspective:

> [...] i.e. the possibility that there could exist electrically neutral particles in the nucleus, which I want to call *neutrons*, and which have a spin of ½, obey the exclusion principle, and distinguish themselves from light quanta in the fact that they do not move at the speed of light. (see Figure 7.3; Pauli, 1957/1964, p. 1316, my translation, emphasis added)

Second, Pauli tells us the following about his views at the time he lectured about them on his American trip in the summer of 1931:

> I did not hold them anymore to be nuclear building blocks; *therefore,* I no longer called them neutrons, and used no particular name for them. (Pauli, 1957/1964, p. 1316, my translation, emphasis added)

Apparently, the reason why he called them neutrons was that he thought they were nuclear constituents. But, even more importantly, Brown (1978, pp. 24-27) has shown that this statement is wrong and in contradiction with the recollections of the participants of the 1931 Rome conference and the newspaper articles detailing Pauli's American travels. Brown situated the moment when Pauli changed his mind about whether his particle was a nuclear constituent in 1932 or 1933, yet did not draw the obvious conclusion: that Pauli thought that his 'neutron' was a nuclear constituent until Chadwick discovered the (heavy) neutron in February 1932. This would mean that, although he thought that Rutherford was mistaken in regarding the neutron as a composite particle, he was convinced of its presence in the nucleus.

Finally, even in his recollections, Pauli links his idea to Rutherford's by describing Rutherford's suggestion in the historical paragraphs before the earlier quotes. Furthermore, Pauli criticizes Rutherford for taking the neutron to be a close combination of a proton and an electron, and informs us that this was the reason why Rutherford had no experimental success in finding neutrons in hydrogen discharges (Pauli, 1957/1964, p. 1315). While this is a correct analysis from a present-day point of view, it could also reflect Pauli's thinking in 1930: the main difference between his and Rutherford's ideas was that Pauli's neutron was an elementary particle, which allowed for a lower mass and explained the half-integer spin (which all known elementary particles had at the time). It seems possible that Pauli adopted this idea, i.e. that neutrons, if they existed, would have to be elementary particles, in light of the failure of Rutherford's experiments up to 1930.

In summary, although I do not deny its somewhat speculative nature, there is evidence for the thesis that Pauli, consciously or not, thought that Rutherford's idea, given its neutral charge and high penetrability, could be used as the fitting piece to solve the nuclear problems, on condition that it was not considered to be a proton-electron combination but instead a truly elementary particle (hence, having spin ½) of smaller mass. The idea that there might be two different neutral particles, the neutron and the neutrino, most likely occurred to him only after the discovery of the (heavy) neutron by Chadwick in 1932. This thesis, however, demystifies one of the many stories about epiphany that have entered the canonical history of science, and supports the more credible view that many new ideas originate from adapting old ideas for new purposes. After all, the only difference between Rutherford's and Pauli's original idea was a difference of mass and of its elementary nature (being a spin ½ particle); Pauli's idea had to undergo many more adaptations before it became our current idea of the neutrino.

7.9 Summary and Conclusions

Let me start by summarizing the six attempts to form an explanatory hypothesis for the anomalous β spectrum discussed above. This will enable us to draw some general conclusions about these processes and so to link this case study back to the questions raised in the introduction.

First, I discussed Ellis and Wooster's suggestion in their seminal paper about the β spectrum. Making use of Rutherford's nuclear satellite model, they stated that the difference between the discrete α spectrum and continuous β spectrum could be traced back to the nuclear layer in which the particles originated. Therefore, they put forward the idea that, in addition to the quantized orbit in which the α particles resided, there was an unquantized orbit of β particles and questioned, as a result, the universality of the quantum postulate. For Ellis and Wooster, this was justified as they regarded the quantum postulate not as a genuine postulate, but rather as a phenomenon arising from particles that describe stable orbits.

Second, we considered Rutherford and Chadwick's ideas. These experimentalists suggested that the continuous β spectrum was caused by variations in the internal energy of otherwise identical β nuclei. This idea seriously infringed the identity principle, which states that equal particles (or atoms) are indistinguishable. To reach this radical hypothesis, they evaluated a spatiotemporal process model of β decay, and realized that the continuous variations must either have entered in the decay itself or else have been present from the start. As the former option implied (in the interpretation of their model) that energy was not conserved – something inconceivable because of their "experimental bias" – they regarded the latter option as the more plausible one. Most probably, an analogy with Aston's discovery of isotopes a decade earlier at the Cavendish laboratory played a role in Rutherford and Chadwick's reasoning.

Next, I discussed Thomson's account, which is a clear example of how the counterintuitive aspects of quantum mechanics misled even renowned physicists in early years. Fully embracing the new formalism and maybe somewhat blinded by its attractive fruitfulness, he assumed that it also contained the key to the β puzzle. To solve this puzzle, he actually followed a very straightforward pattern of discovery: he visually recognized a Gaussian curve in the experimental data, and calculated the appropriate parameters for a maximal fit with the quantum mechanical model of a free particle (which is formulated in terms of Gaussian curves). This identification and calculation led him to the conclusion that energy was not conserved in the process (which is, in his account, a natural consequence of quantum mechanics). Unfortunately, he relied heavily on certain classical assumptions that rendered his contribution contradictory. However, the idea that energy is not conserved in β decay proved persistent.

The fourth physicist discussed was Niels Bohr. By considering the successful quantum mechanical explanation of α decay (which results in a mono-energetic spectrum) and recognizing the mechanical equivalence of α and β decay, Bohr understood that if the β electrons were present in the nucleus beforehand, then in their case quantum mechanics would also predict the occurrence of a mono-energetic spectrum. Because this was (experimentally) not the case, he concluded that the electrons must have been created in the process of decay. As no physical theory yet existed for such a process of elementary particle creation, he foresaw the development of a new theory, which, if he did not impose energy conservation on it, would allow him to solve multiple energetic problems in the nucleus at the same time. Yet as the nuclear model prevalent around 1929, the *p-e* model, required the presence of electrons in the nucleus, Bohr first had to address this apparent contradiction, which he did by means of a typical Bohr-style conceptual analysis. By this process, which was also at the heart of the formulation of his complementarity principle, he was able to reconcile apparent contradictions and link observations back to classical concepts.

In discussing Heisenberg, we noted that he turned the debate on its head, regarding solved problems as in fact anomalies and vice versa. Overwhelmed by the many problems concerning the nucleus and the formulation of QED, he thought that α decay (for which a sound quantum mechanical explanation existed) was the anomaly for a yet to be constructed theory of nuclear physics that would explain all problems of the nucleus. Inspired by the quantum-classical divide, he proposed a new divide between the nuclear and the quantum (in his view, the atomic) levels, which allowed him to construct new laws for this new level via an appropriate correspondence principle – exactly the same process that led him earlier to the formulation of his matrix mechanics.

The sixth and final physicist discussed was Pauli, who was put off by Bohr's and Heisenberg's denunciations of energy conservation. Given his personal history in relativity, he recognized the unifying power of fields as ontological entities and the consequences of Noether's theorem. Conservation laws, therefore, were at the heart of his ontology: as they seemed to be violated in experiment, he understood that the only option was to add something to the picture (something not yet observed) in order to balance them. At the time, Rutherford's early proposal of the neutron was already in the air, and it could well be that Pauli, as argued above, used this earlier idea in a slightly adapted form.

Having thus reviewed the various attempts to solve the β puzzle, and so completed a case study of genuine variation in different patterns of hypothesis formation, I will now draw some general conclusions by answering the following two questions, based on the evidence brought forth by the case study, which reflect the questions raised in the introduction. How did the scientists in this case study determine which pattern of hypothesis formation they would employ? And do the patterns of hypothesis formation employed in this study have any common features that tend to be overlooked in the literature on hypothesis formation?

The main conclusion of this study is that, in the examined case, the scientists' choice of pattern of hypothesis formation was always implicitly made and directly determined by their personal perspective on their field and on how the problem at hand should be situated in it. Even when scientists work with the same formalisms and theories, they sometimes have different perspectives on how the various elements structurally hang together, and it is this perspective, which is often based on their personal experiences, that implicitly determines which patterns of discovery the scientist will judge suitable. In our case study, we saw that Ellis and Wooster's idea that the quantum of action resulted from classical stable orbits led them to doubt the universality of the quantum postulate; that Rutherford and Chadwick's experimental bias prevented them from questioning the laws of conservation, which were at the core of their experimental models; that G.P. Thomson's continued adherence to his classical intuitions concerning electrons confused him, and led him to an incoherent conclusion; that Bohr's total perspective on science as describing the observational phenomena in everyday concepts informed and motivated his method of tinkering with higher-level concepts, yet led him to remain hesitant of allowing new low-level phenomenological models (such as particles); that Heisenberg's reversal of solved problems and anomalies cleared an entire field for him, for which a theory could be constructed; and that Pauli's ontological perspective made him suspicious of all proposals to limit conservation laws.

This idea concurs to a certain extent with Henk De Regt's (1996) point that physicists' philosophical remarks are not of much importance for the philosophy of science, but can be understood as the justificational grounds for their research heuristics in confrontation with their contemporaries.[33]

As a second conclusion, the case study also suggests that individual real-life scientists apparently do not employ different patterns of hypothesis formation when approaching a single puzzle: they tend to stick to a pattern that best fits their perspective. This adherence to a certain method has led many of the involved scientists to important results: Bohr's complementarity principle, Rutherford's idea of the neutron and Heisenberg's matrix mechanics were all obtained by the same pattern of hypothesis formation as they used in this case. Still, this tendency to adhere to a particular pattern is certainly not absolute; and none of the scientists involved had any problem (eventually) acknowledging the subsequent results of others.

Finally, as different as the patterns and motives of these scientists were, two properties of hypothesis formation patterns tend to appear prominently that are not always adequately appreciated in the literature: the adaptation of old ideas and the use of visual and intuitive models.

[33] See also Kojevnikov 2011 on the role of philosophizing for physics professors in the 1930s, the protagonists in this story.

First, none of the scientists discussed presented a completely new idea; all adapted an old idea for new purposes or drew an analogy with an existing idea. Ellis and Wooster reinterpreted Rutherford's nuclear satellite model; Rutherford and Chadwick drew an analogy with the research on isotopes conducted in their laboratory; G.P. Thomson employed the existing Gaussian model for a free particle; Bohr had already relied several times previously on the rejection of energy conservation; Heisenberg constructed his nuclear-quantum divide by analogy with the quantum-classical divide; and Pauli could well have been influenced by Rutherford's older idea of the neutron.

Second, all of the scientists discussed relied on visual or intuitive models. Ellis and Wooster's model was clearly a visual nuclear model; Rutherford and Chadwick used a causal process model, which allowed them to derive an exhaustive list of possibilities; G.P. Thomson started from a visual identification of the β spectrum as a Gaussian curve, yet it was also his visual classical intuitions about trajectories of particles that led him to misunderstand quantum mechanics; Heisenberg introduced grid cells as a form of lattice theory; and Pauli adhered to an ontology based on the symmetries of fields. The only exception here might be Bohr, but if we understand how he tried to apply (restricted) everyday concepts to physical phenomena, we realize that what he was doing was exactly re-introducing intuitive images and concepts in an overly mathematical and formal theory.

Part IV

Thinking about Models

motivation

In this part, I want to tie the research on hypothesis formation more closely to the recent literature on scientific methodology and the philosophy of science in general. In this literature, substantial progress has been made on the topic of scientific discovery by paying close attention to the use of models in science and the practice of model construction. Therefore, if research on hypothesis formation aims to be relevant for explaining scientific discovery, we first need to understand how these two practices actually relate in model-based science.

My main goal is to find out how model construction and hypothesis formation connect in actual scientific practice. I will further show how one of the conclusions of the previous part, i.e. that scientists adapt old ideas for new purposes, can be better understood if we describe this practice in terms of models.

The core assumption at the heart of this part of the dissertation is that the use of models is a ubiquitous phenomenon in scientific practice, or, in other words, that it is hard to make sense of science without the concept of scientific models.

The assumption of the ubiquity of models coincides with the general consensus in the literature. In Section 1.2, I already noted how scientific models have received increasing attention in the literature of the philosophy of science in recent decades, and the recognition of their use in science has elicited many fruitful philosophical questions about their ontological nature, the nature of their representational relation to the world, and by virtue of which characteristics models are epistemically capable of leading us to new scientific knowledge. A further introduction to this general literature on the use of models is included in Chapter 9, the more general chapter of this part.

strengths and weaknesses of the method

For this part, it is not that easy to list in a simple way the various strengths and weaknesses of the methods applied, as this part does not employ one specific method but rather a mixture of conceptual analysis and actual case studies.

Mixing these two methods is clearly a trade-off. No conceptual analysis will ever cover every case in actual practice, and neither will case studies alone ever produce concepts sufficiently coherent to allow for conceptual analysis. Somehow, sufficiently coherent concepts need to be formulated that still cover as many actual cases as possible.

Yet by mixing these methods in this way, the strongest disadvantages of the individual methods of conceptual analysis and of historical case studies (such as, for instance, those mentioned in the introductions to Parts I and III) are at least somewhat tempered. But the problem of generalization from individual cases can also now not be ignored. I deal with this issue differently in each of the two chapters of this part. Chapter 8 further elaborates a part of the case study from Chapter 7, which, as has been extensively argued, can be considered as an exemplary case. Chapter 9 attempts to show that its analysis of the relation between models and hypotheses is generally applicable, by taking three cases from the same field (astrophysics) with very distinct types of models.

A further advantage of connecting hypothesis formation to the use of models is that it shows how forms of creative abduction can occur. If one considers creative abduction merely from the point of view of hypothesis formation, the newly hypothesized concepts or objects often seem to appear out of nowhere. But in bringing hypotheses in connection with models, it becomes more clear how, in model-based science, new ideas are formed.

overview of my contributions

In chapter 8, I elaborate on the case study examined in the previous part and show how one of its conclusions, i.e. that scientists adapt old ideas for new purposes, can be better understood if we look at how scientists use models and how these models can lead them in particular directions. In the included case study, I show how the gap in Pauli's reaction model for β decay could be filled by adapting Rutherford's older neutron idea, which was developed in the context of a completely different model. The concept of scientific model employed is, although less developed, fully compatible with the more detailed view of the latter chapter.

In chapter 9, I analyze in great detail the relation between hypotheses and models in science. After delineating how I understand these concepts, I start by identifying the various stances that can be found in the present literature and consider various objections to these. Next, by drawing on three cases from astrophysics, I first distinguish between two types of hypotheses, heuristic hypotheses and fully interpretable hypotheses, and show how the relation between models and these two types of hypotheses in actual scientific practice can be understood.

8 The Adaptation of Old Ideas

This piece of history illustrates a general maxim: that any hypothesis, however absurd, may be useful in science, if it enables a discoverer to conceive things in a new way; but that, when it has served this purpose by luck, it is likely to become an obstacle to further advance.

— Bertrand Russell, *History of Western Philosophy*, 1945

This chapter is based on the paper "Pauli's Idea of the Neutrino: how Models in Physics allow to Revive Old Ideas for New Purposes", published in L. Magnani (ed.), Model-Based Reasoning in Science and Technology: Theoretical and Cognitive Issues *(Gauderis, 2013c). I am indebted to Bert Leuridan and two anonymous referees for their helpful comments on earlier drafts.*

In this paper, I examine why models have proven to be a key catalyst for many new ideas and hypotheses in science. More in particular, I identify a third reason why models can perform such an important heuristic role, apart from the two main reasons stated in the current literature, namely, that they allow one to mentally simulate many scenarios and that they facilitate a wide cross-fertilization across various disciplines. What these two reasons do not focus on is the functional design of models: the fact that they are designed for a certain purpose. Hence, gaps in models are functional gaps, which motivate scientists to vigorously explore older ideas (formulated for different purposes) that can be adapted to fill these gaps. This idea is illustrated by explicating Pauli's neutrino suggestion in terms of the used models.

The content of the article is largely retained; although, in order to prevent overlap with the previous chapter, I have reduced and adapted some parts (especially Section 8.4) and further made some small stylistic corrections.

8.1 The Heuristic Role of Models in Science

Models perform an important heuristic role in scientific investigation (Redhead, 1980; Morgan and Morrison, 1999). Their success in this role is typically accounted for by reference to two widespread practices. On the one hand, because of their dynamic nature, models allow one to extensively explore and mentally experiment with existing theories. As such, one can simulate various scenarios and identify and mediate lacunas or anomalies in a given theory (Morgan and Morrison, 1999, p. 19; Nersessian, 2008, p. 185). On the other hand, many simple abstract models have been applied to a variety of contexts outside of their original field, a process that has led to an extensive interdisciplinary cross-pollination.[1] This shows that scientists are actively looking for useful models that can be applied to problems in their own field.

To this analysis, I want to add a third important heuristic practice involving models. Models have a typical functional structure or, in other words, they are designed with a certain purpose in mind. This means that lacunas or gaps in a model are functional, i.e. that the model misses something that can fulfill a subfunction required by the rest of the model. This invites the designer to actively explore old ideas (originally formulated for different purposes) that might serve in a slightly adapted form to fill these gaps and play the role of the missing cogwheel for the total model's purpose.

My aim in this chapter is to explicate and illustrate this third practice (and its relation to the other two) by further exploring Pauli's 1930 suggestion of the particle later called the neutrino. In Section 7.8, I have argued that this idea may not have been so original as Pauli thought it was. In fact, it is perfectly possible to see it as an adaptation of Rutherford's original idea of the neutron, proposed in 1920. I will further substantiate this claim by explicating the problems that both Rutherford and Pauli were working on, as well as the models they employed for their purposes. This analysis will interpret this history as an example of how an idea that arose in a certain program managed to stay alive, even though the program grew obsolete, and to be picked up ten years later as the missing piece in a model for

[1] There are numerous examples of this cross-fertilization. Some of the more spectacular examples are the so-called genetic algorithms in artificial intelligence, which are based on models of natural evolution (e.g. Goldberg, 1989), the use of Markov chain-models to identify authors in philological studies (e.g. Khmelev, 2000) and the use of phase transition models from physics to address problems in the social philosophy of science (e.g. De Langhe, 2013).

another and much more prominent puzzle in the field.

I will start, in Section 8.2, by expanding on the role of models in scientific discovery in order to show how the three heuristic practices identified above can be understood in generic terms, although the scope of this analysis will be restricted to the heuristic use of models in physics. This will provide us with a conceptual framework to analyze the case study, which naturally falls into two distinct parts: Rutherford's reasoning and models in 1920 (Section 8.3) and Pauli's reasoning in 1930 (Section 8.4).

8.2 Models and Scientific Discovery in Physics

As some scholars have noted, it is very hard to give a precise definition of a model, even if we restrict ourselves to models in physics (Hartmann, 1995, p. 52; Nersessian, 2008, p. 12). Such a definition is here not really needed, however, and one can content oneself, as Nersessian proposes, with a loose definition that is sufficient to capture the way physicists think of models.[2]

From the most general point of view, a *model* can be conceived as an abstract imaginary system of interrelated parts that has as a whole certain distinguishing characteristics. The most important characteristics of models are (1) their *functional design*, (2) their *representational potential* and (3) their *susceptibility to manipulation*, all of which are uncontested in the relevant literature.

Models in science are in the first place functional or, as Morrison and Morgan state it, are designed or constructed to "function as tools or instruments" (1999, p. 11). The purpose of this design can be a variety of scientific activities such as theory construction, explanation, prediction, suggesting which data should be collected, etc.[3]

[2] In Chapter 9, I craft a more precise definition of a scientific model. Yet, the view developed in this chapter is compatible with and fits the more refined definition of the next chapter.

[3] Epstein (2008) distinguished seventeen different reasons why one should model. Apart from the most straightforward reasons prediction and explanation, he identifies models also as a key method for e.g. finding gaps in the data and formulating new research questions. While models are often constructed for several of these reasons, Epstein convincingly argues that e.g. models for prediction and explanation are typically of a different nature. Not all of his seventeen reasons, however, will necessarily motivate directly the design of particular models. His paper aims mostly to convince scientists in fields where models are less common, and some of his reasons such as "teaching us a scientific outlook" are little more than interesting qualities that result from the regular use of models.

The main reason why models can function as a tool for all these purposes is their second property: they are meant to represent certain features of the world, or as Nersessian explains it, "they are designed to be structural, functional or behavioral analogues of their target phenomena" (2008, p. 12).

Finally, the reason why models can be considered to be tools is their susceptibility to manipulation, and in this they distinguish themselves from mere representative descriptions. The design of a model is such that one can interact with it mentally by manipulating certain features, adjusting certain parameters or adding or removing certain parts, all of which represent interventions in the target field (Morrison and Morgan, 1999, p. 12).[4]

With this characterization in mind, we can explain how models play a role in the three heuristic activities identified in the introduction. When a scientist is confronted with a new target phenomenon or a collection of experimental data,[5] she can try to structure this data by constructing a model. She does not have to start this activity from scratch. Generally, some initial constraints are available from a general theory or from some related models. For example, if a researcher tries to construct a model for a particular type of nuclear reaction, initial constraints are raised by her model of nuclear constitution and by certain general theories, such as quantum mechanics. Still, it is obvious that there are no clear algorithms at this stage and model construction, here, is more a matter of skill. In the literature, this activity has been described as constructing with bits from different sources (Morrison and Morgan, 1999, pp. 15-16), as matching representations to mathematical structures (Czarnocka, 1995, p. 30) and as constructing a hybrid between target and source domains (Nersessian, 2008, p. 28). The common denominator is that modelling is a piecemeal assembly process. This assumes that scientists have certain simple blueprints at hand,[6] simple mathematical models and structures that they have acquired over the years and which they can combine with theory, experimental data and various representations. The so-called model constructing skills consist, then, in maintaining and expanding such a set of blueprints; and applying

[4] Because of this feature, the use of models has received a central place in the interventionist view of science, going back to Hacking and Cartwright (Cartwright, Shomar and Suárez, 1995).

[5] For the distinction between phenomena and data, see Bogen and Woodward (1988); Woodward (1989); Glymour (2000); Massimi (2007, 2011); Woodward (2011).

[6] My use of the notion 'blueprint' is not exactly the same as that of Cartwright (2007), when she uses this notion in the context of models.

this set to various problems is exactly the second heuristic practice that I have identified.

Second, the susceptibility of models to manipulation allows scientists to simulate and explore within the constraints imposed on the model both by its internal (formal) coherence and by their knowledge of the target domain. This hybrid construction gives the models a relative autonomy, which allows scientist to identify lacunas and anomalies both in the theory and in the model itself – the first heuristic practice stated above.

Now we can specify the third practice that makes models such a useful heuristic tool. The functional design of a model ensures that every part of the model has its own function. As such, if a lacuna is identified in a model, it will be a functional lacuna, i.e. the model will be missing something that can fulfill a function required by the rest of the model. Researchers can try to form their own original ideas to fill these gaps, but, as the case study will exemplify, researchers generally browse their field for ideas that have the requisite properties. This is a different activity than the use of various models from other fields. Where, in this latter activity, the abstract structure of the model is borrowed and completed by adjusting the representational elements to objects in the field the researcher is working in, in the former activity, the researcher actively pursues ideas from her own field that were proposed for different purposes or problems, some of which may have already become obsolete.

To sum up, I have identified three scientific heuristic practices involving models, which explain why the use of models is heuristically so successful. First, models can be applied, by virtue of their partly formal structure, to many problems situated in other fields, or at least shed some initial light on these problems. Second, they allow for dynamic simulation on the basis of which researchers can explore the various combinations of the model's parameters. Third, because of their functional design they invite scientists to actively reconsider old ideas in order to spot an idea that has the right characteristics to fulfill a particular function in the grand design of the model.

In order to apply these concepts to the case study, I will first explain my view on the two types of models that will be discussed in it, i.e. constitution models and reaction/process models. *Constitution models* are the oldest type of physical models and relate to ancient philosophical questions about the nature of things. The main purposes of these models are explanatory: by specifying the various parts and the total structure of the target phenomenon, one aims to explain certain properties of the whole, such as, for instance, its stability or fluidity. As *process* and *reaction models* started to emerge only since the scientific revolution, they are much younger types of models. Their main purpose has been not so much to explain the nature of the represented changes, but rather to explicate the necessary conditions and the results to be expected. Experimentalists have used them as a guideline to manipulate and control physical reality. In other words, these types of models are primordially designed for prediction, not explanation. At the same time, the scientific revolution and these experimental models put more stringent conditions on the older constitution models: they had to become compatible with (most of) the process and reaction models of the target phenomenon and their descriptions had to be limited to qualities that are in principle testable. Therefore, constitution models in physics are generally limited to describing the various subentities and specifying the forces or mechanistic properties that keep them together. Still, their main purpose remains explanatory as they are not strictly needed for prediction.

Both types of models are dynamic in nature and allow the scientist to interact with their various parts. For constitution models, this dynamics lies mostly in the possibility of exploring what combinations of subentities can possibly exist. To discuss the dynamics, I need to distinguish between process and reaction models. I view *reaction models* as models that take the represented change to occur instantaneously, e.g. models for radioactive decay or chemical reactions. In contrast with process models, which represent a gradual or stage-based change of the target phenomenon, reaction models represent only the *initial situation* and the *resultant situation*, considering the change as something that has happened at a certain point in between. The main goal of these models is to specify, apart from the conditions under which the reaction can take place, which characteristics and entities are conserved and how the non-conserved properties change. Their main dynamics lies in the fact that one can mentally explore various situations to picture what the result of the reaction would be. *Process models*, which do not occur in the case study, draw one's attention rather to the change of the target phenomenon itself, and enable scientists to simulate various scenarios of how they might control or accommodate this process.

By ascribing different purposes to these different types of models, while still assuming their compatibility, I take models to be more or less autonomous but related to each other. The autonomy of the models' purposes is also one of the reasons why Morgan and Morrison (1999) consider models to be independent of physical theories, the other reason being that also the construction of models happens more or less autonomously from theory, or, as Cartwright (1999) states it: "Theories do not provide us with algorithms for the construction of models, they are not vending machines into which one can insert a problem and a model pops out". This relative autonomy will also help us to understand the role of theories such as quantum mechanics and classical electrodynamics in the case study. The *semantic view* (in which theories are superfluous families of models) and *syntactic view* (in which the logically structured theory carries all scientific value) are both too restricted to capture how theories and models function in the endeavors of scientists.[7] Cartwright (1983) has described the laws in a theory as "schemata that need to be concretized and filled in with the details of a specific situation, which is a task that is accomplished by a model"; and models, that may have been initially constructed in the framework of a theory, can develop their own dynamics, which might lead to the suggestion to withdraw a certain aspect of the theory (e.g. Bohr's suggestion to withdraw the energy conservation theorem, see Section 7.6).

8.3 Case Study A: Rutherford's idea of the Neutron

In this and the next section, I take on the challenge of explicating how my claim of Section 7.8, i.e. that Pauli's suggestion might well have been an adaptation of Rutherford's older neutron idea, should be understood in terms of the characteristics of the models both physicists employed. The case study naturally falls into two parts. I will first, in this section, explain Rutherford's project around 1920, and show how the idea of the neutron emerged from his model. In the next section, I will then consider Pauli's idea, which was presented in a completely different context, and show how the model he had in mind led him to think that the neutron might be the solution.

As described in Section 7.8, Rutherford suggested the idea of the neutron for the first time in his Bakerian lecture (1920). The main reason why he believed this idea to be valuable was that he thought that its existence "seems almost necessary to explain the building up of the nuclei of heavy elements" (p. 397). Translated into our present conceptual framework, Rutherford perceived an incompatibility between his constitution model of atomic nuclei and the theory of classical electromagnetism, because the laws of the latter do not allow for the building up of the more heavy nuclei. This had led him to investigate further the constitution of nuclei,

[7] See Frigg and Hartmann (2012) for an excellent summary of these two points of view.

partly by actual experiments, partly by mental simulation of the model. It was exactly this type of mental simulation of the various possibilities that convinced him that there possibly existed a neutron, although his conception of it was totally different from our current understanding. The fact that this idea, yielded by simulation of the model, could fill the functional gap in the model convinced him of the soundness of this idea, a conviction that inspired him to look tenaciously for experimental proof over the next ten years. Finally, in 1932, his close collaborator Chadwick managed to assemble sufficient evidence to confirm its official discovery.

Let us first explain Rutherford's nuclear model. Like many of his contemporaries, he believed that the nucleus consisted of "electrons and positively charged bodies" such as helium and hydrogen nuclei (p. 377). But, as he had already suggested in 1914, the assembly of these positively charged bodies can ultimately be considered a combination of positively charged hydrogen or H nuclei (which gradually came to be called "positive electrons" or protons) and negatively charged electrons, which were kept together by the electromagnetic force – the so-called *p-e* model (see Section 7.2.2).[8]

Yet although this model was at the time the only viable one, given the common ontology of the day, it faced severe difficulties.[9] As Rutherford mentions, the apparent lack of magnetic moment of the intranuclear electrons hints that these electrons must be somewhat "deformed" (1920, p. 378) and that they are in no sense comparable to the extranuclear electrons orbiting the nucleus. But his main problem was the constitution of large nuclei. As soon as a nucleus contained a certain number of protons, the combined repelling Coulomb force would be just too large to let another proton come close enough to swallow it. Rutherford was vividly aware of this problem, as observed it on a daily basis in his experiments. While he found it possible to shoot lighter elements with α particles or He nuclei, initiating a collision, he found it impossible to penetrate larger

[8] Hanson (1963, pp. 157-159) explains the fact that for a long time scientists refused to consider any other elementary particle besides protons and electrons by pointing to the fact that these two particles were at the same time considered to be the elementary subunits of the two types of electrical charge. As there was no other type of electricity, there was no reason to presuppose another elementary particle. See also my analysis in Section 7.2.2 of the self-evidence of this model.

[9] At this point, I only mention problems that were already known in 1920. The more famous problems for this *p-e* model, which I have discussed in Section 7.2.3, such as its use of the wrong type of statistics for the nitrogen nucleus and the Klein paradox, arose only during the 1920s.

nuclei due to their high electrostatic repulsive forces.[10] It was exactly because part of the α particles were repelled from the gold foil in the famous Rutherford-Mardsen-Geiger experiments of 1909 that Rutherford had inferred the existence of the nucleus in the first place.

Rutherford thought he could cope with these problems by assuming certain substructures in the nucleus. Nuclei were not just a heap of protons and electrons that all attract each other more or less equally. He thought that protons and electrons were bounded in small stable substructures, which in turn were grouped together to form the full nucleus. The reason why he (and the physics community in general) had this idea was the remarkable stability of the α particle. In experiments it turned out to be impossible to break up this element by collisions (1920, p. 379). It was also observed as an independent structure in α decay, which led several physicists to assume that it was part of the nucleus as such. Around 1920, it was Rutherford's main experimental program to find more stable combinations such as the α particle to complete the nuclear constitution model. Because he was not able to reach the nucleus of heavier elements with α particles, he conducted experiments mainly on the lighter elements (nitrogen, oxygen, carbon) in order to produce collisions and study the remaining parts. His first discovery were H nuclei or protons (Rutherford, 1919). This was important because, although it was generally assumed that protons existed independently in the nucleus, it was "the first time that evidence had been obtained that hydrogen is one of the components of the nitrogen nucleus." (p. 385), and that, hence, the *p-e* model had some experimental ground. Second, he discovered a certain atom, which he called ${}^{3}_{2}\mathrm{X}$, with atomic mass 3 and nuclear charge 2, which made it "reasonable to suppose that atoms of mass 3 are constituents of the structure of the nuclei of the atoms of both oxygen and nitrogen." (1920, p. 391)

In order to figure out the substructure of this X atom, Rutherford reasoned that "from the analogy with the He nucleus, we may expect the nucleus of the new atom to consist of three H nuclei and one electron" (p. 396), which made this atom a snug fit in the *p-e* model. But when he realized that this means that a single intranuclear electron can bind three protons,[11] it appeared to him "very likely that one electron can also bind

[10] This also nicely illustrates that Rutherford perceived experimental data, models and theories all as more or less autonomous entities that should be made compatible with each other.

[11] An atom with a single intranuclear electron had not yet been observed so far: hydrogen had, according to the *p-e* model, no intranuclear electrons, while the next element in the periodic table, helium, already had two.

two H nuclei and possibly also one H nucleus" (p. 396). In other words, by mentally exploring what is also reasonable to expect according to the *p-e* model, he came to the idea of a close binding of one proton and one electron, an "atom of mass 1 and zero nucleus charge". He expected this combination, which he started to call the *neutron* only later on, to be a very stable entity with "very novel properties". Because there would be hardly any electromagnetic field associated with this neutral combination, it would be able to travel rather freely through matter. Therefore, it might reach the nucleus of heavy atoms without suffering from a repelling force, where "it may either unite with the nucleus or be disintegrated by its intense field" (1920, p. 396).

The thought process by which Rutherford came to the idea of the neutron is highly intriguing, because hardly any part of it is still acceptable according to our present standards: the *p-e* model is plainly wrong; later experiments did not confirm the existence of the X atom; the whole idea that there exist certain substructures in the nucleus is flawed; and above all, according to our present understanding, it is absolutely untrue to consider a neutron as a combination of a proton and an electron.[12] Still, judged in light of Rutherford's background knowledge, his thought process is a very sane and sound piece of reasoning in which he improved his constitution model by combining experimental data with mental simulation of his model. And, although Pais (1986, p. 231) claims that this whole search program for atomic substructure has left no mark on physics, I have argued that this program did indeed lead to a valuable idea, which is not only the forerunner of our current neutron, but possibly also, as I will show in the next section, of our current neutrino.

8.4 Case Study B: Pauli's idea of the Neutrino

As explained in Section 7.8, Pauli's suggestion was an attempt to answer two experimental anomalies: the continuous β spectrum, the severity of which led Debye to say at the time that "it is better not to think about it at all, just like new taxes" (cited in Pauli, 1957/1964, p. 1316; see Figure 7.3), and the integer spin of nitrogen nuclei.

[12] The idea that neutrons were not close combinations of protons and electrons took some time to gain traction. Even when Chadwick discovered the neutron in 1932, and Heisenberg published four months later the first proton-neutron constitution model of the nucleus, they did not consider the neutron already as an elementary particle (Bromberg, 1971).

Let us first try to understand these problems in terms of the models in which they were formulated and of Pauli's perspective on them. This will allow us then to explain how Rutherford's older idea could be adapted to fit the gaps in Pauli's models.

Like all of his contemporaries, Pauli saw radioactive decay in terms of a reaction model (as defined in Section 8.2) in which an unstable nucleus (the initial situation) decayed spontaneously into a remnant nucleus and an emitted α or β particle plus some γ radiation (the resultant situation). For α particles, this model preserves both energy and electric charge during the reaction, but for β decay, the surprising continuity of energies in the resultant situation had already led to some very radical ideas such as that this continuity already existed in the initial situation (Rutherford and Chadwick, see Section 7.4) or that this model should be relieved of its energy conservation constraint (Bohr, see Section 7.6). It was this final suggestion that triggered Pauli's engagement with this problem, as he considered, given his ontological views (see Section 7.8), retracting the energy conservation constraint as a bridge too far.

In 1930, the accepted model for atomic constitution was still the *p-e* model, yet its problems had only grown larger since 1920 (see also Section 7.2.3). In 1926, experimental research proved the spin of nitrogen nuclei to be integer. Yet, according to the *p-e* model, which defines the total spin of a nucleus as the sum of the spin of its constituents, the $^{14}_{7}$N nucleus consists of 14 protons and 7 electrons, and should, hence, have in total a half-integer spin, as both protons and electrons have spin ½. This anomaly hit the *p-e* constitution model right at its core: it seemed impossible that nuclei were constituted only by protons and electrons.

As Pauli stated, he tried to address these problems by connecting them (Pauli, 1957/1964, p. 1316). From an experimental perspective, this is no mere trivial connection. β decay was studied in the context of radioactivity, which focused typically on heavy elements. For instance, the initial β spectrum by Ellis and Wooster was established for what was called at the time Radium E (nowadays $^{210}_{83}$Bi, an unstable isotope of bismuth). Yet from the perspective of a theoretical physicist like Pauli, who focused on the involved models, the sizes or types of the involved nuclei do not matter much. As radioactivity was conceived as a reaction model involving nuclei in both the initial and resultant situations, it could easily be brought into connection with the constitution model for any nucleus, as the latter model must apply to all nuclei in the reaction model.

Each of these connected models had a gap, which could be solved in only a very limited number of ways. For the radioactivity model, given that for Pauli the conservation of energy must apply, either the energy of the nuclei must be continuously varying in the initial situation, or else something energetic was not included in the resultant situation. As Pauli deemed the first option (Rutherford and Chadwick's road) to be impossible, he was obliged to accept a lacuna in the resultant situation of his model. Concerning the constitution model for atomic nuclei, Pauli could, by mentally manipulating the model, see that there was, mathematically speaking, no possible combination of protons and electrons that could yield the right nuclear mass, charge and spin for all types of nuclei. As both protons and electrons have been experimentally discovered in nuclear experiments, and as elementary particles were considered to be indestructable, Pauli could only acknowledge that this constitution model was missing something that carried some spin. Combining these models, he was left with a process model for radioactivity that had two gaps: some spin was missing in the initial situation and some energy was missing in the resultant situation. The simplest assumption was to assume that it was the same that was missing.

It now appears straightforward to conclude that, if Pauli had come to realize all this, it would have been just a small step for him to hypothesize an electrically neutral particle (as charge was conserved) with the requisite spin, energy and momentum to fit these models. However, considering this as a simple move is viable only from our present-day perspective; it would be a historical fallacy to characterize it as 'simple' in the case of Pauli. It is true that in our present-day ontology of fields and particles as their manifestations, a missing or unobserved particle is a straightforward hypothesis for any deviating characteristic. But in Pauli's time, there was a great reluctance to accept the idea of new particles and, even though Pauli had a quite modern field concept, the connection between particles and radiation was not yet fully understood (as the concepts of fields and particles had not yet been unified).

That Pauli filled these connected lacunas by suggesting an electrically neutral particle makes a lot more sense if we know that, at the time Pauli was confronted with these incomplete models, there was a well-known hypothesis in circulation about an electrically neutral constituent of the nucleus, which was hard to detect and the object of a great deal of experimental research effort. Hypothesizing a neutral nuclear constituent could fill both gaps in Pauli's models if it was, unlike the Rutherford neutron, conceived as an elementary particle, and hence as having spin ½. In Section 7.8, I presented various reasons why this may indeed actually have been how Pauli came to his idea, although I do not claim this to be factually proven.

But even without decisive factual evidence, if we explain the problem that Pauli was trying to solve in terms of the functional gaps in his models, it becomes clear that making the mental leap to the idea of a fitting neutral particle would have been easier if he had already encountered and thought about the other hypothesized neutral particle being considered at the time, which proved to be so hard to detect. Although it had been suggested in a completely different context and was aimed at another purpose, its properties made it suddenly a viable candidate to fill the functional gap of these new problems. The only thing Pauli had to do was conceive it as an elementary particle, such that it had spin ½ and could have a mass that was lower than a proton-electron combination.

Aftermath In 1932, Chadwick, a close collaborator of Rutherford at the Cavendish laboratory, announced the experimental discovery of the (heavy) neutron. In his article (1932) he stated explicitly that what he had found was the particle Rutherford envisioned in his Bakerian Lecture in 1920. This discovery was directly accepted by the physics community, and gave rise to the first proton-neutron models of the nucleus later that year, which were able to explain the anomalous statistics of nitrogen nuclei. Pauli's hypothesized particle, however, remained alive as a hypothesis for the β-spectrum, and was dubbed the "neutrino" by Fermi in 1933 to distinguish it from Chadwick's discovery. As it was no longer necessary that the neutrino was a nuclear constituent (as the problem of the anomalous spin had been solved by the neutron), it could now be conceived as a product of β decay. Also, an experimentally improved measurement of the β spectrum showed a shape that favored Pauli's hypothesis over Bohr's suggestion to give up energy conservation (Brown, 1978). From this point on, Pauli's solution drew more and more adherents, while Pauli himself connected this idea also with the conservation of angular momentum in β decay and the problems of Dirac's 'hole theory' (Massimi, 2005, pp. 133-134). Finally, in 1934, Fermi's model for β decay, which incorporated Pauli's neutrino, provided a solution to these various related problems. Consensus soon followed, and Bohr himself admitted defeat in 1936. Experimental evidence for the neutrino, in any case, was not found until 1956.

8.5 Conclusion

In explaining the utility of using models for heuristic purposes, the functional design of models is often left out of the picture. In this chapter, I have shown by means of a conceptual analysis and consideration of a detailed historical case that precisely this functional design of models forces researchers to vigorously explore old ideas in order to adapt them for their current purposes. In light of this, I have identified a third practice – in addition to the often-mentioned mental simulation of various scenarios and the wide cross-fertilization between different fields – that explains the heuristic success of models. At the same time, it provides a template for understanding how certain forms of creative abduction can occur.

Old ideas are often reused, generally adapted or employed as analogies or metaphors. The case study in this chapter explains in detail how Pauli could fill the functional gap in his model for radioactive β decay by adapting an old idea that figured in Rutherford's atomic constitution model. But not only entities or objects can serve as ideas to be adapted for new purposes; this is illustrated by Bohr's suggestions that the energy conservation principle be retracted. At least twice before, he had used this same idea to solve a certain puzzle, each time with a completely different purpose.

Although it is in many cases impossible to tell, without a detailed case study, whether a given case of hypothesis formation concerns the use of an old idea or whether one came independently to the same idea, there is no reason to suspect that most of these ideas were original. In light of this, if we want to understand how scientists use models and reuse them, it is important to be aware of how models invite scientists by their functional structure to actively explore old ideas in order to adapt them for their own purposes. Further case studies and formal analyses are needed, however, to understand the implications of this for our current methodologies.

9 Models and Hypotheses

Imagination creates events.
— Giovanni Francesco Sagredo,
letter to Galileo, 1612

This chapter is based on the article "Models and Hypotheses", which is currently under review (Gauderis, 2014b). I am indebted to Bert Leuridan and various other members of the Centre for Logic and Philosophy of Science for their helpful comments on earlier drafts and presentations.

In this article, I have studied in depth the relation between models and hypotheses in scientific practice. This analysis is based on an overview of the various stances that have been developed in the literature as well as on three case studies from astrophysics, which have allowed me to develop my own view on the matter, i.e. that, in model-based science, hypothesis formation and model construction are mutually dependent and supportive practices for scientific discovery.

The content of the original article is largely retained, except for some minor stylistic adaptations.

9.1 Introduction

As a result of a shift in focus in the philosophy of science from dealing largely with issues of scientific confirmation towards studying actual scientific practices and the questions they give rise to, philosophical interest in the use of models in science has steadily increased in recent decades. Although early interest was mostly fueled by adherents of the so-called *semantic* and *structuralist views of theories*,[1] who tried to tailor their formal

[1] Until the 1970s, the received view of theories (also called the *syntactic view*) maintained the Euclidean or Aristotelian ideal of a theory as a set of axioms and a suitable logic to infer

analyses towards the type of models actually used by scientists, it is now recognized that the structural set-theoretical meaning of models is best not equivocated with the actual practices of *model-based science*,[2] which elicit many ontological and epistemological questions in their own right. The study of these questions in relation to many actual scientific cases has led many to appreciate that much of science can be adequately described as model-based science, which should not be seen so much as a division between the various disciplines as rather a strategy that any discipline can employ to address theoretical scientific research (Godfrey-Smith, 2006). The construction, manipulation and refinement of models are also now generally considered to be key scientific practices (Frigg and Hartmann, 2012).

The recognition of the use of models in science has elicited a substantial amount of research to clarify the relation between this rather new addition to the jargon of scientific methodology and older inhabitants of this conceptual jungle, such as the relation between models and theories (e.g. the semantic view, Giere 1988), models and discovery (Redhead, 1980; Morrison and Morgan, 1999), models and laws (Cartwright, 1983; Giere, 1999a) and models and data (Suppes, 1962; Harris, 2003). Yet, no substantial attention has been paid so far to the relation between models and hypotheses in science. The main reason why this relation has been left unattended may have to do with the fact that models and hypotheses are generally considered to belong to the jargon of two mutually exclusive conceptions of the scientific method, i.e. the inductive (model-based) view and the hypothetico-deductive view.[3]

In this chapter, I investigate how hypotheses and models relate in actual model-based scientific practice, show that both are necessary concepts in

all true sentences in an ideal scientific language, supplemented with a set of correspondence rules to link theoretical terms to empirical observations. The heavy language-dependency of this view has led various scholars to develop the so-called *semantic* view of theories, in which theories are equated with a class of models, abstract mathematical structures for which the theory is true (Suppes, 1960; Suppe, 1977, 1989; Van Fraassen, 1980). A related, *structuralist* view of scientific theories was developed by, among others, Balzer et al. (1987). For a recent paper incorporating these structuralist ideas, see Leuridan (2013).

[2] The two main arguments for this distinction are that many models from actual scientific practice cannot be accommodated within the set-theoretical view of models (Downes, 1992) and that, while the semantic view aims to analyze all of science in terms of models, not all actual scientific practice relies on the manipulation of models (Godfrey-Smith, 2006; Weisberg, 2007).

[3] This position is advanced, for instance, in the article of Glass and Hall (2008) that I discuss in Section 9.3.3.

understanding this practice, and show that they are mutually supportive. Apart from touching upon recent debates in the literature on models, such as those concerning the nature of their representational function and their construction, this research will reinstate a modernized concept of a scientific hypothesis, in line with model-based scientific practice, by shrugging off some of the unrealistic intuitions with which it has been burdened by the old Popperian hypothetico-deductive view.

After delineating my precise usage of the main concepts of this chapter (Section 9.2), I will identify four stances on the relation between hypotheses and models by examining the scattered remarks that have been made in the literature and considering what objections might threaten these stances (Section 9.3). Then, I will look into actual scientific practice and present three case studies to expose the nature of the interplay between models and hypotheses (Section 9.4). This will allow me to develop my own account of how hypotheses and their role should be understood in the context of model-based science (Sections 9.5 and 9.6).

9.2 Some Conceptual Issues

Before I present the main arguments of this chapter, some preliminaries are in order concerning its scope and topic. More specifically, as ‘model’ and certainly ‘hypothesis’ are often used as umbrella terms and as their meaning is often thought to be more or less self-evident, I need to specify more precisely the kind of hypotheses and models this chapter will deal with. Unavoidably, this requires a trade-off between catching as much as possible of the actual usage of these concepts in scientific practice and defining sufficiently coherent concepts to allow for analysis.

Scientific Hypotheses I take *scientific hypotheses* to be (1) statements (2) about the empirical world (3) that have an unknown or underdetermined truth status and (4) are advanced as a tentative answer to a particular research question.

Let me expand on each part of this characterization. First, scientific hypotheses are linguistic statements or propositions, by virtue of which it always makes sense to talk about their truth status.

Second, this chapter focuses only on hypotheses that make reference to the empirical world. This excludes, along with mathematical conjectures, also hypotheses that refer exclusively to parts of and relations within a particular model. Studying the internal properties of scientific models is an important aspect of theoretical science, but conjectures of this kind are generally not what scientists refer to with the notion 'scientific hypothesis'.[4]

Third, although it makes sense to speak (typically in retrospect) of confirmed hypotheses, it is assumed that hypotheses are not known to be true. Yet this does not exclude that scientists can have a firm and even justified belief in them, certainly in later stages of research. Also, and although this move may be less commonly accepted as it drastically extends a Popperian notion of hypothesis, I do not assume that hypotheses are fully determined or have an unambiguous reference. As the case studies in this chapter show, many actual hypotheses in early stages of research unavoidably have ambiguous or vague references. It is only afterwards, when the conceptual apparatus, requisite models and governing conditions have been developed in subsequent stages of research, that the intended hypothesis can be formulated unambiguously.

Finally, scientific hypotheses are not mere conjectural statements; they are advanced in an attempt to answer particular research questions. In other words, they are *truth-purposive*. Scientists advance them with the purpose of finding the answer to a research question by trying to determine the suggested hypotheses' truth value, even if they know that any particular hypothesis can be rejected or refined later on. Importantly, it is not required that hypotheses be compatible with the agent's background knowledge: many valuable truth-purposive hypotheses presented in history have firmly contradicted large portions of the adopted set of beliefs or (assumed) knowledge of those who suggested them. In such cases, the agent thought that pursuing the truth value of the hypothesis he had in mind might anyway lead to certain answers to his research question, even when he was well aware that parts of his background knowledge would need revision if this particular hypothesis turned out to be true.

With this final condition, I have excluded a large class of hypotheses

[4] This relates to Contessa's (2010) distinction between *external sentences* (e.g. "The emission spectrum of hydrogen can be calculated with the Bohr model") and *internal sentences* (e.g. "In the Bohr model of the atom, electrons orbit around the nucleus in well-defined orbits"). I consider scientific hypotheses to be external, while internal sentences belong to the model itself or a description of it. (On a side note, I disagree with Contessa (2010, p. 223) when he states that the electrons orbit in well-defined orbits in Rutherford's 1911 model: Rutherford assumed only that the extra-nuclear charge was uniformly distributed throughout a sphere (Rutherford, 1911, p. 671); well-defined electron orbits appeared only earlier in Nagaoka's speculative 1904 model (also called the Saturnian Model) and afterwards in the 1913 Bohr model.)

from my characterization of scientific hypotheses: explicit counterfactuals and belief-negating hypotheses. Although these *truth-denying* hypotheses have their role in science by virtue of, for instance, thought experiments (De Mey, 2006), I consider them fundamentally different from the *truth-purposive* hypotheses this chapter deals with, as (a) their truth value is explicitly known or believed to be false and (b) they neither provide a direct answer to any particular question, nor are they aimed to determine their truth value (as it is already assumed to be false). Their purpose is generally to set up a line of reasoning that can lead to certain sought-for answers via a detour, such as a thought experiment or a *reductio ad absurdum* argument.[5]

Scientific Models I take *scientific models* to be (1) abstract or concrete artifacts (2) purposefully created in order to be manipulated to perform particular scientific tasks (such as prediction or explanation) by exploiting certain representational relations.

Although this characterization is in line with much of the actual use of the notion '(scientific) model' by scientists and with the contemporary literature on models,[6] I have made some restricting choices.

First, ontologically, I consider models to be either concrete or abstract models, yet my focus will be on the abstract type. It is commonly accepted that the human imagination can create such things as abstract objects and that many scientific models, such as the ideal pendulum or the Bohr model of the atom, should be understood as such. But the consensus ends here,[7] and the concept 'abstract artifacts' is still burdened with

[5] My distinction between *truth-purposive* and *truth-denying* hypotheses relates to Rescher's (1964) classic distinction between hypotheses with an unknown truth status, on the one hand, and belief-negating hypotheses and counterfactuals, on the other. However, there is one *caveat*: Rescher operates in a logical framework (which assumes logical omniscience). Therefore, for Rescher, it makes no difference whether the agent explicitly believes (or knows) that the hypothesis is false, or that this is only a consequence of her set of beliefs (or knowledge). For my purposes, this distinction does matter. A hypothesis is *truth-denying* only if the agent *explicitly* believes (or knows) that it is false. When the agent thinks it might be true, it is *truth-purposive*, even if it is in contradiction with his set of beliefs (or knowledge). This situation actually occurs frequently in science: as many problems are overdetermined, scientists are often willing to accept that part of their set of beliefs (or assumed knowledge) is wrong in advancing a new hypothesis.

[6] This characterization is inspired by, amongst others, the views of Giere (2004, 2010); Hughes (1997); Teller (2001); Bailer-Jones (2003); Nersessian (2008); Knuuttila (2011), and fits the accounts of actual scientists reporting on their use of models (Bailer-Jones, 2002).

[7] Current debates focus mostly on the relation between scientific models and other abstract

metaphysical riddles.[8] However, as we are concerned with scientific practice, we can content ourselves that, even if an adequate account of abstract artifacts is hard to achieve, abstract artifacts do exist as such in the folk ontology of scientists. Actual scientists do talk about models as if they were abstract artifacts. So, even if it turns out that our analysis is confined to the folk ontology of scientists, it can still give us some insights into scientific practice itself, precisely because this practice is framed in terms of this folk ontology.

As it is my purpose to determine the relation between models and hypotheses (which I take to be abstract propositions, rather than concrete utterances or sentences, see Part I), my analysis will unavoidably focus on abstract models. In principle, this would exclude from the analysis any tangible model, such as plastic models, diagrams, descriptive texts or annotated drawings. But this should not unduly concern us, as most such tangible concrete models can be straightforwardly interpreted as representations of a particular abstract model,[9] while tangible models used for the direct representation of real target phenomena, such as a wooden bridge model, are not of concern to us here.

Second, functionally, I take models to be used to represent some target

objects such as fictional objects (Godfrey-Smith, 2009; Giere, 2009; Contessa, 2010). On the other hand, this discussion has also been rejected as a non-issue, which deviates attention from the more pressing questions about the use of models in scientific practice (Teller, 2001; French, 2010).

[8] Although not yet fully explored in the context of scientific modeling, the *artefactualist view* of abstract objects has some obvious advantages over the *nominalist view* (as it recognizes abstract objects) and the *Platonic view* (as it sees them as having originated in time and as ontologically dependent on the existence of a concrete creator). Yet, it also gives rise to important questions such as the identity of models (e.g. are the Bohr model and the Bohr-Sommerfeld model the same model?), the contours of models (what are the essential and non-essential parts of a model?) and the nature of the creative act, as scientific models are often constructed in successive stages by varying groups of scientists (See also the discussion on *creationism* or *artefactualism* in Kroon and Voltolini (2011)).

[9] Interpreting concrete or tangible models as representations of abstract models makes sense only if one adopts an (at least) *three-place analysis* of the *representation relation*: representation is not purely a relation between a model M and a target T, but a relation of an agent S who uses a model M to represent a target T for some purpose (Giere, 2004). As such, concrete tangible models, such as a double helix made from cardboard, can be used in two ways: either to represent directly a target phenomenon (actual DNA), or to represent an abstract model (the Crick and Watson double helix model), which can itself be used to represent that same initial target (actual DNA). Although the particular form in which an abstract model is represented does influence the scientist's actual manipulations (Knuuttila, 2011; Vorms, 2011), I will pay no further attention to individual (tangible) models in the present chapter.

system in the real (or empirical) world.[10] This is what Giere (1999b) has called the *representational conception* of models, as opposed to the *instantial conception* of models used in the semantic and structuralist analysis of theories. In the representational conception, the intended representational relations can be exploited for predictive or explanatory purposes by manipulating the model. In the case of abstract models, this kind of manipulation can be done in several ways, for instance, by varying the model's variably designed parts or contrasting various concrete representations of it. This representational conception also resembles to a certain extent the notion of a model used in the model-based reasoning community (e.g. Nersessian, 2008), yet does not focus exclusively on the psychological issue (Godfrey-Smith, 2006).

The target system the model is used for to represent can also be a set of data points or measurements. Such models of data (Suppes, 1962) or phenomenological models, which are generally constructed via statistical methods of data analysis, are sometimes seen as temporary models requiring further explanation by deeper explanatory or constitutive models (e.g. the 1885 Balmer formula for the hydrogen emission spectrum lines was explained by the 1913 Bohr model of the hydrogen atom). Yet, this type of model is often employed in actual scientific practice, especially for predictive purposes (consider, for instance, the importance of the discipline of data analysis) and is highly esteemed by scientists with a strongly inductivist mindset (see e.g. Glass and Hall in Section 9.3.3). Therefore, it is important that our analysis of the relation between models and hypotheses should apply to this type of model as well.

9.3 Four Stances on the Relation between Models and Hypotheses

In this section I review four stances that can be found in the literature. However, it should be kept in mind that none of the authors I will associate with these stances were explicitly concerned with specifying the relation between hypotheses and models. In each case, the characterization was embedded in a broader research goal.

[10] Although this characterization fits large classes of models in science, it does not fit all models (Downes, 1992).

9.3.1 Models are (a particular form of) Hypotheses

Although the stance that models are just a form of hypotheses is never explicitly articulated in the current literature, it is often implied when authors use the terms 'model' and 'hypothesis' more or less interchangeably. The idea is that models are just a particular form of hypotheses: they are a bit more elaborate, and often have some figurative elements, but in essence they are just hypothetical suggestions which can be tested to confirm whether they conform to reality. This view is particularly appealing to people focused on explanatory and mechanistic models, as for this kind of models the parts of the model are intended to have an accurate one-to-one correspondence relation with the parts of the target system.

This stance, however, fails to take into account the important and currently hot issue of the representational relation between models and the world. The representational relation between hypotheses and the world is easier to specify: hypotheses are linguistic entities. Therefore, whether they represent the world can be determined by finding out whether they are true or false. But models are not linguistic entities.[11] Therefore, one cannot determine whether a model is *literally* true or false. When a model is called true (or false), this attribution normally has to be understood in a *metaphorical* or *pragmatic* sense: it indicates that the model meets the purposes for which it was designed, such as accurate prediction or explanatory power, not that it consists of literally true sentences.[12] Even if one replaces truth with a gradual notion such as accuracy, for many models it makes no sense to assess whether or not they are accurate, both because they were never intended to be so and because of their use of idealizations, simplifications and fictional entities.

The representational relation between models and the world is cur-

[11] There is a minority position that does take models literally as linguistic entities (Frigg and Hartmann, 2012). This view, which is embedded in a syntactic view of theories, takes models (just like theories) to be sets of statements about a target system, simplified or idealized for certain purposes (Achinstein, 1968; Redhead, 1980). This position, however, has to cope with similar concerns as the syntactic view of theories. For example, it faces the obvious objection that there can be many different linguistic descriptions of the same model. How should the canonical description be determined? As a matter of fact, I have found no recent adherents of this position.

[12] Mäki (2011) has, however, tried to define a literal truth relation for models (see also Perini (2005) on the possibility of such a truth relation for pictorial representations), but, in essence, Mäki's proposal boils down to defining the truth of a model as the truth of the assertion that the driving mechanism of the model is the same as its target mechanism (which makes him rather fit the stance discussed in Section 9.3.2).

rently still under debate. But, although most contributors have understood that any analysis of the representational relation must take the user and her intentions into account,[13] the discussion has somehow arrived at an impasse between structural accounts (following Van Fraassen, 1980, 2008), which seem too strict to capture all models used in science, and similarity accounts (following Giere, 1988, 1999b, 2010), which seem too minimalist to explain the epistemic value of reasoning with models (Knuuttila, 2011, p. 264; Downes, 2011). Although the view developed in this chapter inclines more towards a similarity account, we need not go further into the details of this debate.

So, while most would nowadays agree that models are not pure linguistic entities, it could still be argued that my characterization of scientific hypotheses (in Section 9.2) is too narrow. Why could the concept of hypothesis not be stretched to include particular non-linguistic entities, such as models? After all, the predicate 'hypothetical' can be sensibly applied to other objects.

In answer to this line of reasoning, it should be noted that the requirement that hypotheses be linguistic entities (or even propositions) is not in fact a restriction of the concept. The core feature of any hypothesis is that its truth value is either uncertain or underdetermined (in the case of a truth-purposive hypothesis) or else that it is known or believed to be false (in the case of a truth-denying hypothesis). Such truth values can be *literally* ascribed only to linguistic entities such as stories, descriptions or claims.[14] Of course, it is a natural stretch to attribute the predicate 'hypothetical' to objects such as dark matter (truth-purposive) or Earth's second moon (truth-denying). But in these cases, the predicate refers to the existence of these objects, and such attributions are equivalent to stating that "Dark matter exists" has an uncertain (or even underdetermined) truth value and that "Earth has a second moon" is known to be false.

Let us now return to the question of whether models can be called hypothetical. Clearly, models cannot be hypothetical in the same sense as dark matter or Earth's second moon. If one has a particular (abstract)

[13] See Footnote 9.

[14] For now, I leave aside attempts to expand the truth concept to non-linguistic entities. My main reason for this is that it is hard to do so in an unequivocally accepted way. Definitions of such an expansion (e.g. Mäki, 2011) are based on an intuition of what should be considered true, and these might differ from agent to agent. For instance, concerning Mäki's definition, one might argue that the *literal* truth of models should be defined in terms of accurate prediction, not in bringing forth the correct mechanism.

model in mind, then it exists. Earth's second moon does not, and about the existence of dark matter one can read a mountain of scholarly articles (see Section 9.4.3). If models are called hypotheses, this ascription refers to their content, not their existence.

So, could it not be that models are linguistic after all, or, even more minimalistically, that they maybe have a characterization defined in purely linguistic terms? After all, many models in science are known purely from a textual description, and Craver (2006), borrowing ideas from Railton (1981), has introduced in the mechanism literature the notion of *the ideally complete description of a mechanism* as the ideal for a mechanistic model. Let us grant this for a moment, and assume that there exists for each model in science an ideal, fully linguistic description that fully characterizes the model. Such a description of a model would indeed have a truth value. But it would be true only by reference to the model itself. If we were to determine its truth value by reference to the world, it would always be false. Models include fictitious entities (e.g. point masses or frictionless planes) or describe unreal and simplified conditions (e.g. no air resistance or uniform mass density) and even if a model is very descriptive, as are particular mechanism models in biology, its (ideally) full description would be false by reference to the world because of the simplifications and abstractions it incorporates. For instance, the description might state that one body part is directly adjacent to another part, while in reality there are blood vessels, tissues and fat cells in between.

Or, turning the argument around, if the ideally full description of a model were to be completely true with respect to the world, there would be no model defined, because the description would be just a direct description of this part of the world. In conclusion, we can state that if such a thing as the (ideally) full description of a model existed, it would be true only by reference to the model, and it would be literally false with respect to the world. Hence, it would be counterfactual.[15] Therefore, if we were to use this construction to call models hypotheses, they would be truth-denying hypotheses and not truth-purposive hypotheses, as their creators had likely intended.[16] In addition, such a construction (of an ideal descrip-

[15] This analysis relates to the analysis of the falsehood of models in Cartwright (1983) and Wimsatt (1987/2007).

[16] An exception to the general idea that modelers aim to be truth-purposive might be the construction of *toy models*, which are purposefully built not to represent much but rather to experiment with the theoretical tools themselves. Toy models can therefore truly be characterized as counterfactuals, and allow thus for analysis as a thought experiment.

tion) also conflicts with the idea that we are looking at how models are actually used in scientific practice.

9.3.2 Hypotheses are Statements about the Relation between Models and their Target Systems

This is the idea Giere has been arguing for since his book *Explaining Science: a Cognitive Approach* (1988). According to him, (theoretical) hypotheses (which, he claims, overlap considerably with the use of the notion by scientists themselves) are assertions of some sort of relationship between a model and the system it is intended to represent. In his more recent work (2004; 2008; 2010), Giere specifies this notion of hypothesis further, holding that hypotheses are claims that a fully specified and interpreted model (a model of which each element is provided with a physical interpretation) fits a particular real system more or less well, or any generalization of such claims.

If one has come to appreciate that the relation between models and the world is not simply a matter of truth (or falsehood), but may include a plenitude of possible representational relations depending on the purposes of the agent, it is quite natural to understand scientific hypotheses as specifications of the nature and fit of these representational relations. For instance, many hypotheses state that the values calculated using a particular model fit particular measurements of the target system of the model (within certain error margins), or that the mechanism represented by a particular simplified and idealized model is the same mechanism driving a real target system. Perhaps because it is natural to understand hypotheses in this bridging role, I have found no dissenting voices on this issue amongst scholars working on scientific models.

However, although this analysis is compelling and very suitable to account for a number of hypotheses used in actual scientific practice,[17] it does not fit the majority of hypotheses advanced and argued for in scientific practice. The reason for this is straightforward. Giere's characterization of a hypothesis depends on the existence of a model that can be fully interpreted. This means that this kind of hypotheses can be stated only once a fully interpretable model has been developed, which is typically only in the closing stages of the discovery process. Giere is not to blame for this. His project is to analyze how accomplished science is structured – the starting

[17] In Section 9.5, I will call this class of hypotheses the *fully interpretable hypotheses*.

point of his 1988 investigation was a mechanics text book. But if we want to understand the role of hypotheses in scientific practice, we should take into account that hypotheses are much more closely linked to the discovery process than to the presentation of well-established science. In the process of scientific discovery, advanced hypotheses are seldom well-specified and fully interpretable (as the case studies in the next section show). Therefore, although we can use Giere's account for a subclass of scientific hypotheses, i.e. the *fully interpretable hypotheses*, we must supplement it with an account of hypotheses used in the actual process of scientific discovery.

9.3.3 Radical Inductionism: Hypotheses should be avoided in Model Construction and Refinement

Recently, Glass and Hall (2008) launched a well-argued attack in the top-ranked journal *Cell* on the use of hypotheses in scientific practice. The use of hypotheses, they argue, is a relic from the old hypothetico-deductive perspective on science, which denied induction as a valid form of reasoning. According to them, the latest articulation of this obsolete view, Popper's Critical Rationalism, was successfully challenged in the second half of the twentieth century by, amongst others, Kuhn and Nozick, while probability and Bayesianism gave the inductivist better tools to defend his position.

Apart from summarizing the main historical and philosophical positions in this well-known debate, Glass and Hall also argue on a pragmatic level that scientists would do better to replace top-down hypothesis testing with bottom-up inductive model-building. Framing research by hypotheses adds severe biases. Not only are negatives less valued than positives (confirmation bias), but also researchers are rendered blind to alternative routes, as negatives are not differentiated (categorization bias). Furthermore, not all interesting research (or research proposals) can be framed by a hypothesis. A telling example was the Human Genome Project, of which, when pressed to state a research hypothesis, J. C. Venter, a major player in the project, stated that "it is our hypothesis that this approach will be successful" (Glass, 2006, p. 18).

Therefore, Glass and Hall suggest that research (and research proposals) should better start by asking an open research question, after which data collection can begin. From this data, which is increasingly abundant and elaborate in this era of Big Data, one can extract a first model via the methods of statistical data analysis, which leads to new questions, further data gathering and model refinement. Nowhere should one, according to this view, introduce unproven premises or hypotheses.

Glass and Hall's argument has the merit that it points out to scientists and funding organizations the danger of bias if research hypotheses are given too much weight. In fact, their suggestion to frame research proposals by open research questions instead of hypotheses (as is sometimes required by funding agencies) is an interesting one, but, philosophically, their suggestion to literally eradicate all hypotheses from scientific practice in favor of model-building cannot be taken seriously. I distinguish three main reasons.

First, at all stages of inductive model-building there are always some often implicit but unavoidable hypotheses present. Even when the research project is framed by a research question, choices will have to be made as to which variables should be tested for in obtaining the first data set, and such choices rely on (hidden) assumptions about which variables are plausible and which are not. For instance, if one is looking for the causal factors and catalysts of a particular disease, the data set will probably contain variables such as water quality, diet or medical history of the test subjects, but not whether these subjects are left- or right-handed or what their favorite ice cream topping is. These decisions as to which variables to include rely on initial hypotheses concerning what might plausibly be factors in the investigated disease.

Further, inductive model-building or statistical data analysis is a discipline crucially dependent on the introduction of assumptions to mold vast data sets into models (of data) that can be manipulated for scientific purposes. The discipline has been described as being "more an art, or even a bag of tricks, than science" (Good, 1983). An often cited and telling example is the curve-fitting problem: given the simplest data set of only two variables, there are already an infinity of fitting mathematical functions. Data analysts constantly have to make decisions (based on assumptions) on how to handle outliers, on the tradeoff between simplicity and data fitting, on how the data is best represented (as this influences model construction), on how the variable is spread in the population (is it normally distributed or not?), and so on.

Finally, Glass and Hall's analysis is very focused on scientific experimentation, and their generalization is based on the old inductive idea that the whole process of scientific discovery can be reduced to inferences from data. It was precisely against this view that Nickles (1980) and other philosophers of scientific discovery have argued: discovery, they hold, is not separate from theoretical considerations and choices. As the examples in the next section will show, many models originate from theoretical considerations. Only later on, when sufficient detail is attained, can they be compared with experimental data or models of data. In fact, precisely because of this, it could be said that Glass and Hall's analysis applies only to the models of data and phenomenological models mentioned above, not to explanatory or constitutive models.

9.3.4 Heuristic View: Hypotheses are Necessary Guidelines in Model Construction

A view opposite to the previous stance is that hypotheses somehow have a heuristic and methodological role in the process of model construction. Although this idea is sometimes mentioned (e.g. Nola and Sankey, 2007, p. 25), it is often just implicitly assumed.

In the remainder of this chapter, I will give an explicit account of this stance. In my view, *heuristic hypotheses* are direct attempts to initially answer a research question, but, precisely because the research still needs to be done, they unavoidably contain vague filler terms or black boxes and can do little more than hint at a particular direction of research. Yet, by this hinting they sketch an outline or rough blueprint, or even maybe just identify the type of model(s) needed to substantiate the initial hypothesis. As such, they reduce the initial research problem to the more specific problem of filling in the black boxes of the model outline, resulting finally in an adequate model, of which a fully specified and interpreted hypothesis (in Giere's sense), if confirmed, can provide an answer to the initial research question.

Before giving a detailed account of this position in Sections 9.5 and 9.6, I will first present in Section 9.4 three case studies that will provide the benchmark for my analysis.

9.4 Three Case Studies from Astrophysics

In this section, I introduce three historical cases to illustrate my analysis of the role of hypotheses in model-based science. I have chosen these cases, related to important research questions in modern astrophysics involving different types of models, to show how generally applicable the analysis is. Due to space restrictions, only the first case will be fully elaborated; for the other two cases, only the key steps in my analysis will be indicated, together with further references to the literature.

9.4.1 The Energy Source of the Stars (1920-1930s)

Around 1920, the source of stellar energy was still a mystery.[18] By that time, Eddington had crafted the basic structural model of a (stable) star, largely confirmed by observations made at the time. His model represented stars as spheres of gas in which, at each internal point, there was an equilibrium between the inward gravitational pressure and the outward gas and radiation pressure, resulting in concentric layers of increasingly lower pressures and temperatures towards the surface.

But he did not know what fueled this radiant energy.[19] Clearly, the solar energy could not be the result of a chemical reaction, such as exothermic oxidation (fire). Even if the sun would be totally composed of carbon, its mass would be barely enough to radiate the sun's current luminosity for a few thousand years. To solve this problem, Von Helmholtz and Kelvin had defended in the 19th century what was later referred to as the *contraction hypothesis*, which was in turn inspired by the *nebular hypothesis* for the origin of our solar system, proposed by Kant and Laplace.[20] This latter hypothesis situates the origin of the solar system in the gravitational collapse of a gaseous nebula. Inspired by this, Von Helmholtz and Kelvin took as the source of solar energy the inward gravitational energy provided, at first, by the accretion of the sun and, after the sun has started to radiate, by the contraction of the sun as it cools down. Using this model, Kelvin estimated the age of the solar system to be of the order of 10 million years – in contradiction to estimates based on the biological and geological record. Darwin (1859/2009) suggested, for instance, in *On the Origin of Species*, based on some geological calculations, that the earth was at least 300 million years old, the time he thought to be needed for the evolution of our current biodiversity. As a matter of fact, this whole situation led to a public controversy between these two leading scientists.

[18] For a thorough and detailed version of this history, see Shiaviv (2010). For a good introduction see Bahcall (2000) or Mazumdar (2005).

[19] This is why I do not call his model a mechanistic model. Clearly, it was his goal to arrive, in the end, at such a mechanistic model.

[20] Although some predecessors had already proposed similar systems, Kant (1755/2012) and Laplace (1796/1830) were the first to (independently) propose a model based on the contraction of a nebulous cloud according to Newton's law of Universal Gravitation.

At the dawn of the 20th century, better geological observations and the discovery of radioactivity quickly discredited the contraction model. The earth (and, hence, the sun) must be older than Kelvin's estimate. Therefore, the contraction model could not supply the requisite energy. Many looked at the new physics that was emerging, hoping it could provide an answer. Rutherford and the young Eddington suggested that radioactive elements might be the source of stellar energy, and Jeans, upon learning of Einstein's $E = mc^2$, suggested that in the extremely hot interior of stars, protons and electrons might annihilate each other, turning their mass into energy.

The experimental breakthrough that prompted Eddington's initial suggestion of nuclear fusion was Ashton's measurements of the mass of He and H nuclei, finding that the mass of a He nucleus was only $99,3\%$ of the combined mass of the four H nuclei it contained. This led Eddington to the hypothesis of nuclear fusion:

> Now mass cannot be annihilated, and the deficit can only represent the mass of the electrical energy set free in the transmutation. [...] If 5 per cent of a star's mass consists initially of hydrogen atoms, which are gradually being combined to form more complex elements, the total heat liberated will more than suffice for our demands, and we need look no further for the source of a star's energy. (Eddington, 1920, p. 353)

This suggestion, although defended fiercely, is clearly just a hypothesis. Apart from Ashton's measurements, he had little or no evidence to back it up, nor did he understand how and when such a fusion process might occur. After all, one should not forget that at the time, neither the neutron nor any nucleus of atomic mass 2 or 3 had yet been discovered. Quantum mechanics had not yet been developed and the amount of hydrogen in the sun was not yet determined. So, Eddington's hypothesis suggested that somehow 4 protons and 2 electrons (which it was thought, at the time, the He nucleus consisted of, see Section 7.2.2) met each other at a single position at a single moment in time, something which Eddington knew was probabilistically nearly impossible, as is illustrated by the following quote:

> Indeed the formation of helium is necessarily so mysterious that we distrust all predictions as to the conditions required. [...] How the necessary materials of 4 mutually repelling protons

> and 2 electrons can be gathered together in one spot, baffles imagination. (Eddington, 1926, p. 301)

Therefore, it is understandable that throughout the 1920s his hypothesis still met with competitors: Jeans kept defending a proton-electron annihilation, while Bohr even thought that in stars the conservation of energy was violated.[21] It was only after numerous contributions of the likes of Gamov, Houterman, Atkinson and Weizsäcker that Bethe (1939) finally put forward a model of stellar energy production which was in satisfactory agreement with the observational record and consisted of two well-described processes that converted hydrogen into helium: the *p-p* chain and the CNO cycle (the latter occurring only in stars more massive than the sun).

Let us review the various characteristics of Eddington's hypothesis of nuclear fusion. Clearly, it fits our characterization: it is a claim about the world with an unknown truth value in answer to a particular research question. In fact, it would be better to state that its truth value is underdetermined. Eddington had no idea how energy could be liberated by combining atoms. There are many possible models – some even totally different from Bethe's model with completely different concepts, elements and forces – that could still be seen as a specification of Eddington's hypothesis.[22]

Still, the credit that Eddington received for this suggestion is justified, as his suggestion was immensely important in redirecting research. In a sense, it simplified the problem of what the source of stellar energy was to the question of how hydrogen nuclei can combine so as to form helium nuclei. This simplification is achieved by providing an initial answer to the question of stellar energy, using a sketchy outline of a stellar model containing a black box process that somehow turns present hydrogen into helium. This is why his idea was so hugely important and why he kept on defending it and urging research in that direction for twenty years, until, finally, Bethe was able to crack open the black box.

[21] Bohr's suggestion (1929/1986) that energy conservation should be renounced must be linked primordially with the problem of the continuous β spectrum (see Chapter 7), but the way in which Bohr combined it with this problem of astrophysics, a field to which he had not contributed at all, shows how pressing the problem of stellar energy still was around 1930.

[22] Consider, for instance, also the history of the briefly mentioned nebular hypothesis. Our current model of the origin of our solar system differs completely from what Kant had in mind (Palmquist, 1987). Still, our current model can be seen as a specification of the severely underdetermined original hypothesis, which is why we still attribute the nebular hypothesis (partly) to Kant.

So what is the nature of the relation here between model and hypothesis? Eddington's model was largely a black box or at most a rough outline, so Giere's characterization of hypotheses does not apply to his hypothesis, as it was heuristic in nature. Only once Bethe's model was available could one say that Eddington's hypothesis, refined by stating that the "combination of hydrogen atoms" has to occur according to Bethe's model, is a hypothesis in Giere's sense: a claim that a fully interpreted model fits a target system.

9.4.2 The Nice Model (2000s)

In 2001, simulations of the model specified by the nebular hypothesis (describing the origin of our solar system), with reasonable assumptions for the initial conditions, confirmed the idea raised a few years earlier that Neptune could not have become such a large planet at such a great distance from the sun (Stewart and Levison, 1998; Levison and Stewart, 2001) – a research problem that triggered, amongst other possible solutions, the hypothesis that Neptune initially formed nearer to the sun and then migrated out (Thommes et al., 1999). Yet this hypothesis was nearly meaningless, as no available model showed how such a migration could have occurred. In 2005, in a series of three papers in *Nature*, the Nice model[23] was presented (Tsiganis et al., 2005; Morbidelli et al., 2005; Gomes et al., 2005). This model postulates that 4 billion years ago there was a period in which Jupiter and Saturn were in 2:1 orbital resonance,[24] which led to a global gravitational instability in our solar system that caused the outer planets to move from orbits much nearer to the sun outwards to their current trajectories. Furthermore, simulations of this model showed that it also explained many other curious features of our solar system, such as the Late Heavy Bombardment (that caused the many lunar craters), the heavy eccentricities of the outer planets' orbits, and the Trojan satellites locked in Jupiter's orbit. In subsequent years, improved simulations and new explanations of

[23] This model, named after the French Mediterranean city where the research was conducted, is generally represented and explored via computer simulations. Philosophically, debates continue concerning how models and simulations relate. On this see, among others, Humphreys (2004); Frigg and Reiss (2009); Winsberg (2010).

[24] This means that, during this period, Jupiter completed two revolutions around the sun in the same time Saturn completed a single revolution. The frequent and regular alignment of these two bodies exerted an extra and periodic gravitational pull on the trajectories of other nearby objects. In the case of Jupiter and Saturn, by far the two most massive bodies circling around the sun, this can lead to serious disruptions in the trajectories of those nearby objects.

further features of the solar system, such as the characteristics of the Kuiper belt, have led to a general acceptance of the Nice Model (Crida, 2009).

The Nice model is clearly a very different type of model than the stellar model discussed in Section 9.4.1. Whereas the stellar model was mainly a very general theoretical model applicable to any star, the Nice model is an applied model tailored to our solar system and established by numerous computer simulations, in which mainly the initial conditions were sought that, given the well-known principles of Newtonian dynamics, could result in the observed specificities of our solar system.

Still, we find here the same type of relation between the model and the heuristic hypothesis that led to its development. The initial suggestion, i.e. that Neptune formed closer to the sun and then migrated out due to gravitational forces in our solar system, provided a first tentative but direct answer to the research question of why Neptune was so massive. Yet, this suggestion was largely vacuous without an exact model or set of initial conditions to specify how such a migration might have occurred. On the other hand, it was precisely the persuasive plausibility of this initial heuristic hypothesis that motivated and coordinated a large research effort to conduct the numerous computer simulations that led to the substantiation of this claim by explicating the unknown mechanism of Neptune's migration. Only now that this model has been built can we reformulate the hypothesis as a fully interpretable hypothesis in Giere's sense: Neptune formed closer to the sun and then migrated out according to the conditions and the mechanism described by the Nice model.

9.4.3 Dark Matter (1930s-present)

Notwithstanding some earlier references to dark stars or matter, the start of the modern search for dark matter is to be found in Zwicky (1933/2009).[25] Having found that the rotation curves of galaxies in the Coma Cluster were much too high to be explained by the mass of the visible stars, he suggested that dynamical models of galaxies should incorporate the presence of non-visible dark matter to explain the observed rotational speeds. In the following decades, the problem was largely cast aside, although a growing number of studies for different galaxies confirmed the high rotational speeds. Gradually, more galactic models incorporating dark matter were advanced, attributing more and more features to it. For instance, Ostriker

[25] Classic histories of dark matter are Trimble (1987); Van den Bergh (1999); Rubin (2003).

and Peebles (1973) calculated that, in contrast with visible matter which is mostly found in the galactic disk, dark matter is mostly present in the galactic halo. The enumeration of the various indications of its existence in a highly-influential review paper of Faber and Gallagher (1979) convinced most astrophysicists of its existence by 1980. In subsequent decades, we have seen an enormous increase in the number of suggestions to characterize dark matter, while some of these possibilities, such as neutrinos or brown dwarfs and other massive dark astronomical bodies (so-called MACHOs), have already been ruled out. At the same time, other hypotheses have been raised to address the initial problem of the galactic rotation curves (e.g. the MOND hypothesis proposed a modification of Newtonian Dynamics), but we also see an increase in the use of the concept 'dark matter' in other models that explain other features of our galaxy, such as gravitational lensing and fluctuations in the cosmic background radiation. Nowadays, the incorporation of the concept in virtually any successful galactic or cosmological model is considered by almost everyone to be sufficient proof of its existence. On the other hand, although some options have been ruled out and some characteristics have been determined, there is still no satisfactory account of the nature of dark matter. The best guess at present is that it is an unknown weakly interacting massive particle (a so-called WIMP).

This final case, about a not yet specified hypothetical entity, might seem different from the other two cases. Yet, also here we can find the same interplay between hypotheses and models, the only difference being that, in this case, most of our present models cannot be fully interpreted and specified (in Giere's sense), as dark matter is not yet fully understood. Zwicky's initial heuristic hypothesis, i.e. that there exists a large amount of dark matter in galaxies, has, despite its neglect at the time it was proposed, redirected much research toward specifying the nature of this unknown type of matter and supplementing this claim with suitable models. But, although galactic and cosmological models including dark matter have been substantially refined over the years and have become the only widely accepted models, and even if these models can be operationalized for some explanatory or predictive purposes, the notion 'dark matter' still remains something of a black box in these models.

9.5 Heuristic and Fully Interpretable Hypotheses

Before turning to the relation between models and hypotheses in model-based scientific practice, let me first define more precisely the distinction I have been hinting at between two types of hypotheses: *heuristic hypotheses* and *fully interpretable hypotheses*. This distinction draws on Craver's 2006 distinction between *mechanism sketches*[26] and *ideally complete descriptions of mechanisms*. How exactly my concepts relate to Craver's concepts will be discussed at the end of this section.

A *fully interpretable hypothesis* is a hypothesis the meaning of which (or any part of which) leaves no room for vagueness or ambiguity. In other words, expressions of such hypotheses do not contain any unexplained *filler terms*, terms such as 'process', 'to interact', or 'entity' that have a broad and generic meaning covering up some uncertainty, imprecision or unknown details. Hence, these hypotheses are fully expressed in terms with a precise meaning, which is provided either by the conceptual framework of the field the researcher is working in or by the researcher himself, by means of suitable models. *Heuristic hypotheses*, on the other hand, do contain such unspecified and generic filler terms.

The main idea is that heuristic hypotheses are both unavoidable and useful in the early stages of scientific discovery, as they sketch an early blueprint or incomplete model without committing one to too much (yet unknown) detail. A heuristic hypothesis suggests that research should proceed in a particular direction, i.e. that it aims to fill gaps in the incomplete model instead of trying to address the general research question directly. Fully interpretable hypotheses, on the other hand, can be put forward only after the construction of a full model that specifies how the hypothesis (which is a claim about a part of reality) should be interpreted precisely and under what conditions it should hold. Therefore, in principle, it is possible to design a conclusive experiment to verify whether a fully interpretable hypothesis holds, while heuristic hypotheses can seldom be tested conclusively due to their vagueness and ambiguity. Experiments in this case mostly aim to refine the model and reduce the vagueness and ambiguity.

Before I add some further remarks and consider some examples, it is useful to explain first how these two types of hypotheses relate. As the main criterion that distinguishes these two types is the level of precision in

[26] The notion of a *mechanism sketch* had already been introduced in the seminal paper on mechanisms by Machamer, Darden and Craver (2000).

the expression of the hypotheses, the two distinguished types are actually the extremes of a continuum. Moreover, as it is an unwieldy (if even possible) task to specify all relevant conditions for a particular hypothesis, it is clear that the idea of a fully interpretable hypothesis is actually an idealization (as Craver could only speak of ideally complete descriptions of mechanisms). Therefore, at first sight, it seems as if there exist only heuristic hypotheses, interpretable to a greater or lesser extent. In scientific practice, however, some hypotheses are clearly considered to be sufficiently unambiguous and interpretable, allowing them to be tested conclusively. Therefore, for our purposes, we can evade this conclusion by allowing for a pragmatic threshold of precision sufficient for full interpretability. A hypothesis can be considered *sufficiently fully interpretable* if it invokes no disagreement in the research community as to which is its meaning. Yet the flip side of adopting this social criterion is that a single researcher cannot himself decide whether a hypothesis is fully interpretable. Also, that a particular hypothesis is considered to be fully interpretable at a certain point in time does not ensure that it will remain so in the indefinite future.

A few further remarks are in order concerning the concept of filler terms, including some examples. First, what counts as a filler term is topic-dependent. For instance, the phrase ‘exerting a force’ has a precise meaning in physics, while in economics this would be a filler term for an unspecified process of influence. Having said this, the fact that so many words in various fields can be considered to have a fully specified meaning is precisely because of the cumulative processes of abstraction and concept formation in these sciences. Therefore, whether a phrase counts as a filler term or whether it has a precise meaning (in a particular reference framework) is dependent on the stage of development in the field. Let me return to the examples presented in Section 9.4. When Eddington in 1920 spoke of “the combination of hydrogen atoms” and somewhat later even used the term “nuclear energy”, these concepts were certainly filler terms. Despite having good arguments for why focusing on a possible transition from nuclear mass to energy could possibly solve the problem of stellar energy, he didn’t have any account of how this energy could be released from the nucleus and why this process occurred in stars. It was only after the acceptance of Bethe’s 1939 nuclear fusion models for the *p-p* chain and CNO cycle that the term ‘nuclear energy’ received a precise meaning in astrophysics.

Also, filler terms generally gain precision only gradually. For instance, while the concept ‘dark matter’ was at first a pure filler term to indicate the possibility of unobserved but present matter, the term has gained some precision and delineation in recent decades. It is now accepted that dark matter mostly resides in galactic halos, that there is at least five times more dark matter than regular matter, that it consists of weakly interacting massive unknown particles (WIMPs), which move at relatively slow speeds (with respect to the speed of light) and are electrically neutral, etc. Yet no astronomer at present would claim that the concept of dark matter is fully understood and precisely defined.

Finally, the given examples might suggest that in the discovery process filler terms themselves always gain a more precise meaning. This happens, such as in the case of 'nuclear energy' or 'dark matter', but more often vague filler terms are replaced with more meaningful descriptions, names or acronyms, such as 'nuclear fusion' or 'WIMP'.

So how do these hypotheses relate to models? For fully interpretable hypotheses, as indicated in Section 9.3.2, I follow Giere in regarding such hypotheses as claims that a fully specified model provided with a physical interpretation fits a target system more or less well. This idea can now be extended to heuristic hypotheses. They are also claims that a particular model fits or might fit a target system, but in this case, the models are just bare model sketches, containing black boxes labeled by filler terms. Still in providing such a model sketch, the initial research question is already partially answered, while at the same time the direction is shown for future research, i.e. to fill in the black boxes.

Before I continue my analysis of the relation between hypotheses and models in scientific practice, let me first explain how my distinction relates to Craver's (2006). The main difference is that I have a different target set of objects in mind. Craver discusses a distinction for models describing mechanisms, while I discuss a distinction for scientific hypotheses. Unlike me, however, Craver does not distinguish between models and hypotheses (as can be seen in his discussion on pp. 360-361). Hence, whether or not his models are linguistic in nature does not seem to matter for him or his purposes. We can therefore conclude that heuristic hypotheses, as described above, would probably, according to Craver, qualify as mechanism sketches if they describe a mechanism, while mechanism sketches qualify, for me, as heuristic hypotheses if they are linguistic in nature. If the latter is not the case, any proposition stating that a particular (non-linguistic) mechanism sketch fits a real system will then qualify as a heuristic hypothesis. At the other end of the spectrum, Craver's ideally complete descriptions do include the claim that the model fits a particular mechanism. Therefore, these can be considered as (fully) interpretable hypotheses in my framework, i.e. claims that the fully specified model fits the target mechanism relatively well.

9.6 The Role of Hypotheses in Model-based Scientific Practice

Let me now spell out the role of these two types of hypotheses in the process of scientific discovery in model-based science. This view will incorporate the two theses I defended above, i.e. that hypotheses are necessary in the process of model construction and that hypotheses that are not fully interpretable are valuable and even necessary in this process.

In general, research aimed at constructing models is triggered by a *research question* or *trigger*. In her monograph on abductive reasoning (the inference from observations to explanatory hypotheses), Aliseda (2006) distinguishes between anomalies and novelties as the two types of observational triggers for abductive reasoning. This classification can be adopted for our current purposes provided we keep in mind the main criticism developed by Nickles (1980) and other scholars of scientific discovery against the idea that abductive reasoning could be the logic of scientific discovery (as suggested by Hanson, 1958), i.e. that abductive reasoning neglects the triggering role of theory in scientific discovery: much research is fueled by theoretical considerations. We can also here distinguish between questions triggered by contradictions and questions triggered by lacunas. Therefore, I conceive of four triggers for research aiming at the construction of models: *experimental* (or observational) *novelties*, *experimental* (or observational) *anomalies*, *theoretical gaps* or *lacunas* and *theoretical contradictions*.

In model-based discovery, these triggers or research questions are answered at the end of the research process by proposing a model and claiming that its similarities with the target system can be exploited to address the research question, or, in other words, by stating a (sufficiently) interpretable hypothesis that is sufficiently verified.

As the model is only linked to the trigger or research question through a hypothesis claiming its fit, such a linking hypothesis, constituting the (partial) answer to the research question, must be present through all stages of model construction; though in the early stages it will be heuristic in nature, not fully interpretable.

Now we have to investigate the role of these heuristic hypotheses in the research process itself. If we take a *constraints-based view of scientific discovery*, which is the view Nickles (1978) developed in the tradition of scientific research as problem solving, we can conceive of a scientific problem (or research question) as a set of constraints. Making progress on a problem consists in manipulating these constraints such that the problem turns into a problem that is somehow easier to solve, such as a less complex or a more familiar problem.

In the case of suggesting a heuristic hypothesis as an initial partial answer to a research problem, one deliberately adds a constraint: however vague a heuristic hypothesis might be, it excludes particular solutions and directs research in a particular direction. As such, one progresses on the problem by reducing it to a simpler problem, though always at the risk that one will not find a solution along these lines (if the heuristic hypothesis turns out to have been a wrong path from the start). After reducing the initial research problem to the simpler problem of finding a suitable model to substantiate the filler terms, the heuristic hypothesis remains important as the link between the reduced problem and the initial research question, for it shows how the latter can be answered by means of the answer to the reduced problem.

Let me illustrate this role of heuristic hypotheses with some of the cases of Section 9.4.

Eddington reduced the open problem of stellar energy (a theoretical gap) to the more restricted problem of how hydrogen could combine so as to form helium. After the problem was reduced to finding a suitable model for this combination, Eddington's hypothesis remained the link that made possible that a solution to this reduced problem, namely Bethe's model of hydrogen fusion, could be used to answer the initial research question of where stellar energy originates. By the time Bethe's model was developed, Eddington's hypothesis could be considered a fully interpretable hypothesis.

Similarly, the research question of the improbable accretion of Neptune (an observational anomaly) was reduced by the initial heuristic hypothesis to the more straightforward problem of constructing a model and determining the initial conditions for an outward-directed gravitational slingshot of a planet within our solar system. Only when such a model – the Nice model – was constructed through numerous computer simulations could the original hypothesis that Neptune initially formed much closer to the sun and migrated outwards be considered as the fully interpretable answer to the initial research question or trigger.

A final thing to address is the fact that many research triggers have the form of an anomaly or a contradiction. Heuristic hypotheses addressing such research questions unavoidably sometimes contradict major parts of the agent's (assumed) background knowledge. Yet as history shows clearly, this does not prevent scientists from developing heuristic hypotheses for such overdetermined problems. In such cases, scientists reason according to what Rescher (1964) has called belief-negating hypothetical reasoning: they accept, within the context of this research, the hypothesis together with all beliefs from their belief set that are compatible with it, while suspending judgment on beliefs that are contradictory to it. For instance, in the case of Neptune, researchers had at first to suspend judgment on the idea that the planets in our solar system were formed at the same distance of the sun where we observe them today. The beliefs still compatible with the heuristic hypothesis, such as all of Newtonian dynamics, could then be used as the basis for solving the reduced problem of constructing a suitable model to interpret that heuristic hypothesis. Only once the model is verified and the research question, hence, answered can the initially incompatible beliefs on which judgment was suspended be revised.

9.7 Conclusion

In this chapter, I have addressed the relation between models and hypotheses in model-based science. After reviewing and pointing out the shortcomings of the various stances in the literature, I have presented my own view on the matter.

First, a distinction has to be made between heuristic hypotheses and fully interpretable hypotheses. Heuristic hypotheses are initial and partial answers to research questions that necessarily contain vague filler terms, yet sketch the outline for the type of model that might be needed to answer the research question. Fully interpretable hypotheses, on the other hand, are claims concerning how a fully constructed model can be used to provide an answer to the research question.

Next, I have shown, by examining three cases from astronomy, how initial heuristic hypotheses fuel the process of model construction and how, once the requisite models are built, they gradually evolve into fully interpretable hypotheses that can, if verified, serve as answers to the initial research questions.

Conclusion

Let me conclude this dissertation by readdressing the questions posed in the introduction, which motivated this piece of research. This will show us how far we have come in our understanding of hypothesis formation in science, yet also reveal some prospects for future research by identifying some of the strands that have been left open.

For centuries, following the example of Bacon, philosophers dreamt of a universal scientific method, an algorithmic procedure that would enable mankind to generate new scientific knowledge. Yet the rapid succession of scientific theories from the 19th century onward showed that the right method could never in itself guarantee the correctness of theories: the scientific method has to consist of alternating phases of idea generation and confirmation.

While the method of confirming new ideas became during the twentieth century increasingly more mathematically stringent, philosophers never raised high hopes of finding an equally stringent counterpart for idea generation – a so-called *logic of discovery* – due to the Romantic ideal of scientific geniuses and the influence of Logical Empiricism. Eventually, this hope was fully crushed in the second half of the twentieth century when, due to the historicist turn in the philosophy of science, it became clear that at most there could exist a multitude of discovery patterns.

At the same time, this historical turn in the philosophy of science reaffirmed an old idea that had lost much of its credibility in the wake of Logical Empiricism, i.e. that the process of discovery is a rational affair and, hence, that discovery was a respectable topic for philosophical investigation. Only, one should focus not on finding one general method or logic but on characterizing a multitude of individual methods, the various patterns of discovery and hypothesis formation. This dissertation aims to contribute to this project by blending formal and philosophical methods and actual case studies from the field of physics.

In part II, I have argued that a full and exhaustive list of such patterns is out of our reach, even if we focus only on patterns that are fully formal and contain, therefore, no discipline-dependent content. After all, such an exhaustive analysis of hypothesis formation would draw on assumptions about how our language and reasoning is structured, and, even had we already obtained such knowledge, I see no reason why no other language and reasoning forms can be developed in the future (with or without the aid of artificial intelligence).

This non-exhaustiveness, however, does not too severely restrain us, as we can always choose one particular approach that is sufficiently general to represent many actual instances of hypothesis formation in science. In Part II, I have shown that if we choose to represent hypothesis formation processes in a Fregean formal language, we can identify some major patterns of hypothesis formation that are used over and over again. At the same time, by representing inferences of hypothesis formation in a Fregean language, I have opened the way for these patterns to be modeled in terms of non-monotonic predicate logics and succeeded in this for the patterns of *singular fact abduction* (Chapter 4) and *abduction of generalizations* (Chapter 6).

This project of formally modelling various patterns of hypothesis formation is certainly not completed, and several ways are open to pursue it further. For instance, once could, using the framework of adaptive logics (which I used throughout Part II), devise a further logic for the pattern of *existential abduction*, according to which we can hypothesize a new object of a known class. Other patterns such as *suggested belief revision* and *inductive generalization* are already (to some extent) modeled via other approaches, but it would be useful to investigate the relation between these various approaches and try to present a unified account. The toughest nut to crack would probably be *conceptual abduction*, in which a new concept is hypothesized, as this seems to require a second-order logic, something of which the framework of adaptive logics is (at present) not capable. Yet, it should be possible to devise in this framework a first-order logic of *concept formation*, i.e. a logic that models the attribution of various characteristics to a new concept once it has been hypothesized. The step before this, however, namely the creative act of pointing out a relevant similarity and labeling it with a new concept, is in my opinion hard to model in a Fregean framework. Aside from the fact that it would need to be a second-order approach, it would also require that all the similarities between various concepts, which are the premises for this kind of reasoning, be fully represented, and this could prove to be an impossible task.

A more fruitful way to understand the formation of new hypothetical concepts is to adopt a richer framework and examine how such newly hypothesized concepts structurally relate to the models used by scientists, a project I pursued in Part IV. This gives us a chance to understand why certain simularities are singled out as relevant and labeled with a new hypothetical concept, even if one has not considered all logical possibilities. In Chapter 8, I showed how gaps in a model often prompt an adaptation of old concepts for new purposes, thus creating a new concept that fits the requirements of the model. In Chapter 9, I then showed how hypothesized concepts are, at first, seldom more than a container concept or filler term denoting not much more than the fact that there is an entity or process. Only in the ongoing interplay between hypotheses and models in scientific practice does a concept receive a more precise meaning, a process that relates to the research on concept formation.

The possibility of building formal models of hypothesis formation processes also opens up the possibility of programming artificial agents to form new hypotheses given a certain background theory, although it is clear that the simulation of such patterns is still in an early stage. In the field of artificial intelligence, this problem is closely connected to (and even dependent on) the problem of *knowledge representation*. One has first to learn how to formally represent scientific knowledge in the way humans understand it, including their models, observed similarities, drawn analogies, rich concepts, and so forth. Therefore, AI approaches to hypothesis formation are, at present, not aimed at simulating or performing actual scientific hypothesis formation, but rather at more specific tasks based on precisely structured sets of background knowledge, such as planning or diagnosis. For such tasks it is sufficient to focus on the pattern of *singular fact abduction*, a project to which I have contributed in Chapter 5 by presenting a reformulation of the logic developed in Chapter 4 that is apt for dealing with rapidly growing knowledge bases.

Yet this dream of programming artificially intelligent agents capable of hypothesis formation suggests, in a way, that the idea of a *logic of discovery* has just managed to sneak in again through the back door. Let me spell this out more clearly. By focusing on the history of science, the idea of a logic of discovery was abandoned in the second half of the twentieth century. Yet in studying historical cases, one has observed not only that there exist various patterns of hypothesis formation but also that hypothesis formation is in fact a rational affair. Hence, we conceive of better and worse ways of forming new hypotheses. But if there are better and worse ways, there should also exist a best way to form hypotheses; in other words, a logic of discovery.

The only way to evade this conclusion is by assuming that most of the patterns we observe in actual practice are, taken independently of context, equally rational. As such, there is no single best way, and, hence, no logic of discovery. This should not imply that rational hypothesis formation is an *anything goes* affair. Hypothesis formation can be rationally assessed by asking which patterns of hypothesis formation are applicable to a certain research problem, and then applying those. But this raises the important question of how scientists choose a particular pattern of hypothesis formation if multiple patterns are applicable and all are equally rational? In other words, why do scientists suggest one hypothesis rather than another, even if they have no evidence favouring the former?

This question has so far not been addressed in the literature. Here, I have attempted to shed some initial light on this matter in Chapter 7 (Part III). By empirically studying how scientists did in fact form hypotheses in the case of the anomalous β spectrum, a problem that elicited a very diverse set of suggested hypotheses, I have found that (for the case under investigation) this process depends mostly on how scientists prioritize and structurally relate the various elements of theory and observations, relations that depend on their experience, metaphysical assumptions, and the goals of science as they perceive them. But we also saw that scientists all too often take too firm a stance towards a particular hypothetical position, even when there is no rational reason to commit so much to one hypothesis. It is clear, however, that further case studies from this perspective will be needed to confirm these initial insights.

The issues encountered in the case studies lead us to the question which attitude one should rationally adopt towards hypotheses. As there is no logic of discovery, it is clear that one should endorse multiple and even mutually exclusive options. In practice, we see that scientists and people in general find it hard to rationally pursue several options, and that they, hence, choose a preferred road, based on certain beliefs formed in their experience and education, on which they further base their actions and pursuits. By investigating the epistemology of hypotheses (Chapters 2 and 3, Part I), I have attempted to characterize the attitude towards hypotheses as an attitude that both allows one to entertain mutually exclusive ideas and yet still provides an ample base for further actions and the pursuit of the various options – an attitude that allows us to live with our uncertainties.

Bibliography

Achinstein, P. (1968). *Concepts of Science. A Philosophical Analysis*. Baltimore: Johns Hopkins Press.

Achinstein, P. (1980). Discovery and Rule-Books. In T. Nickles (ed.), *Scientific Discovery, Logic and Rationality* (pp. 117-132). Dordrecht: Reidel.

Aliseda, A. (2006). *Abductive Reasoning. Logical Investigation into Discovery and Explanation.* Synthese Library (Vol. 330). Dordrecht: Springer.

Anderson, D. (1986). The Evolution of Peirce's Concept of Abduction. *Transactions of the Charles S. Peirce Society*, 22(2), 145-164.

Aristotle. (n.d.). *Posterior Analytics*. Retrieved December 5, 2013, from http://www.logoslibrary.org/aristotle/posterior/index

Bacciagaluppi, G., & Valentini, A. (2009). *Quantum Theory at the Crossroads: Reconsidering the 1927 Solvay Conference.* Cambridge University Press. Retrieved from http://arxiv.org/abs/quant-ph/0609184

Bahcall, J. (2000). How the Sun Shines. *Nobelprize.org*. Nobel Media AB 2013. Retreived December 5, 2013, from www.nobelprize.org/nobel_prizes/themes/physics/fusion/

Bailer-Jones, D. (2002). Scientists' Thoughts on Scientific Models. *Perspectives on Science*, 10(3), 275-301.

Bailer-Jones, D. (2003). When Models Represent. *International Studies in the Philosophy of Science*, 17(1), 59-74.

Balzer, W., Ulises Moulines, C., &, Sneed, J. (1987). *An Architectonic for Science. The Structuralist Program*. Dordrect: Reidel.

Batens, D. (1999). Inconsistency-adaptive Logics. In E. Orlowska (ed.), *Logic at Work. Essays dedicated to the memory of Helena Rasiowa* (pp. 445-472). Dordrecht: Springer.

Batens, D. (2004). The Need for Adaptive Logics in Epistemology. In D. Gabbay, S. Rahman, J. Symons, & J. Van Bendegem (eds.), *Logic Epistemology and the Unity of Science* (pp. 459-485). Dordrecht: Kluwer Academic.

Batens, D. (2007). A Universal Logic Approach to Adaptive Logics. *Logica Universalis*, 1, 221-242.

Batens, D. (2011). Logics for Qualitative Inductive Generalization. *Studia Logica*, 97, 61-80.

Batens, D. (n.d.). *Adaptive Logics and Dynamic Proofs. Mastering the Dynamics of Reasoning with Special Attention to Handling Inconsistency.* Unpublished manuscript. Retrieved December 5, 2013, from http://logica.ugent.be/adlog/book.html

Batens, D., Meheus, J., Provijn, D., & Verhoeven, L. (2003). Some Adaptive Logics for Diagnosis. *Logic and Logical Philosophy*, 11/12, 39-65.

Belbruno, E., & Gott, J. (2005). Where did the Moon come from? *The Astronomical Journal*, 129(3), 1724-1745.

Beller, M. (1996). The Rhetoric of Antirealism and the Copenhagen Spirit. *Foundation of Science*, 63(2), 183-204.

Beller, M. (1999). *Quantum Dialogue. The Making of a Revolution.* University of Chicago Press.

Bethe, H. (1939). Energy Production in Stars. *Physical Review*, 55, 434-456.

Blachowicz, J. (1998). *Of Two Minds. The Nature of Inquiry.* State University of New York Press.

Bogen, J., & Woodward, J. (1988) Saving the Phenomena, *The Philosophical Review*, 97(3), 303-352.

Bohr, N. (1928). The Quantum Postulate and the Recent Development of Atomic Theory. *Nature*, 121, 580-590.

Bohr, N. (1929/1986). Ray Spectra and Energy Conservation. In R. Peierls (ed.), *Niels Bohr Collected Works. Vol. 9. Nuclear Physics (1929-1952)* (pp. 85-89). Amsterdam: North Holland Physics. (Original unpublished manuscript written in 1929)

Bohr, N. (1932). Faraday Lecture: Chemistry and the Quantum Theory of Atomic Constitution. *Journal of the Chemical Society*, 1932, 349-384.

Bohr, N., Kramers, H., & Slater, J. (1924). The quantum theory of radiation. *Philosophical Magazine*, 47, 785-802.

Bokulich, A. (2006). Heisenberg meets Kuhn: Closed Theories and Paradigms. *Philosophy of Science*, 73(1), 90-107.

Bokulich, P., & Bokulich, A. (2005). Niels Bohr's Generalization of Classical Mechanics. *Foundations of Physics*, 35(3), 347-371.

Bratman, M. (1992). Practical Reasoning and Acceptance in a Context. *Mind*, 101(401), 1-16.

Brewka, G. (1991). Cumulative Default Logic. *Artificial Intelligence*, 50(2), 183205.

Bromberg, J. (1971). The Impact of the Neutron: Bohr and Heisenberg. *Historical Studies in the Physical Sciences*, 3, 307-341.

Bromberg, J. (1976). The Concept of Particle Creation before and after Quantum Mechanics. *Historical Studies in the Physical Sciences*, 7, 161-191.

Brown, L. (1978). The Idea of the Neutrino. *Physics Today*, 31(9), 23-28.

Brown, L. (2004). The Electron and the Nucleus. In J. Buchwald, & A. Warwick (eds.), *Histories of the Electron: the Birth of Microphysics* (pp. 307-325). Cambridge, MA: MIT Press.

Cameron, A., & Ward, W. (1976). The Origin of the Moon. In *Abstracts of Papers submitted to the Seventh Lunar Science Conference* (pp. 120-122). Houston, TX: Lunar and Planetary Institute.

Camilleri, K. (2007). Bohr, Heisenberg and the divergent views of complementarity. *Studies in History and Philosophy of Modern Physics*, 38, 514-528.

Campos, D. (2011). On the distinction between Peirce's Abduction and Lipton's Inference to the Best Explanation. *Synthese*, 180, 419-442.

Canup, R., & Ward, W. (2002). Formation of the Galilean Satellites: Conditions of Accretion. *The Astronomical Journal*, 124(6), 3404-3423.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press.

Cartwright, N. (1999). *The Dappled World. A Study of the Boundaries of Science*. Cambridge University Press.

Cartwright, N. (2007). Models: The Blueprints of Laws. *Philosophy of Science*, 64, Supplement, S292-S303.

Cartwright, N., Shomar, T., & Suárez, M. (1995). The Tool Box of Science. Tools for the Building of Models with a Superconductivity Example. In Herfel et al. (1995), p. 137-50.

Cassini, A. (2012). La invención del neutrino: un análisis epistemológico. *Scientiæ Studia*, 10(1), 11-39. doi:10.1590/S1678-31662012000100002

Cassidy, D. (1979). Heisenberg's First Core Model of the Atom: The Formation of a Professional Style. *Historical Studies in the Physical Sciences*, 10, 189-224.

Cassidy, D. (1981). Cosmic Ray Showers, High Energy Physics, and Quantum Field Theories: Programmatic Interactions in the 1930s. *Historical Studies in the Physical Sciences*, 12(1), 1-39.

Chadwick, J. (1932). The Existence of a Neutron. *Proceedings of the Royal Society A*, 136, 692-708.

Chisholm, R. (1989). *Theory of knowledge* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Cohen, J. (1992). *An Essay on Belief and Acceptance*. Oxford University Press.

Contessa, G. (2010). Scientific Models and Fictional Objects. *Synthese*, 172(2), 215-229.

Craver, C. (2006). When mechanistic models explain. *Synthese*, 153, 355-376.

Crida, A. (2009). Solar System formation. arXiv:0903.3008v1[astro-ph.EP]

Czarnocka, M. (1995) Models and Symbolic Nature of Knowledge. In Herfel et al. (1995), p. 27-36.

Darden, L. (1991). *Theory Change in Science: Strategies from Mendelian Genetics*. New York: Oxford University Press.

Darden, L. (1997). Recent Work in Computational Scientific Discovery. In M. Shafto, & P. Langley (eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 161-166). Mahwah, NJ: Erlbaum.

Darden, L., & Craver, C. (2002). Strategies in the Interfield Discovery of the Mechanism of Protein Synthesis. *Studies in the History and Philosophy of Biological and Biomedical Sciences*, 33, 1-28.

Darrigol, O. (1992). *From c-Numbers to q-Numbers: The Classical Analogy in the History of Quantum Theory*. Berkeley: University of California Press. Retrieved December 5, 2013, from http://ark.cdlib.org/ark:/13030/ft4t1nb2gv/

Darwin, C. (1859/2009). *On the Origin of Species by Means of Natural Selection*. Project Gutenberg. Retrieved from http://www.gutenberg.org/ebooks/1228 (Original book published in 1859)

De Langhe, R. (2013) To Specialize or to Innovate? An Internalist Account of Pluralistic Ignorance, *Synthese*, in print.

De Mey, T. (2006). Imagination's Grip on Science. *Metaphilosophy*, 37(2), 222-239.

De Regt, H. (1996). Are Physicists' Philosophies Irrelevant Idiosyncrasies? *Philosophica*, 58(2), 125-151.

De Regt, H. (1999). Pauli versus Heisenberg: A Case Study of the Heuristic Role of Philosophy. *Foundations of Science*, 4, 405-426.

Douven, I. (2002). A New Solution to the Paradoxes of Rational Acceptability. *The British Journal for the Philosophy of Science*, 53(3), 391-410.

Douven, I. (2008). The Lottery Paradox and our Epistemic Goal. *Pacific Philosophical Quarterly*, 89(2), 204-225.

Douven, I. (2011). Abduction. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2011 edition). Retrieved from http://plato.stanford.edu/archives/spr2011/entries/abduction/

Downes, S. (1992). The Importance of Models in Theorizing: A Deflationary Semantic View. In D. Hull, M. Forbes, & K. Okruhlik (eds.), *PSA 1992*, vol. 1 (pp. 142-153). East Lansing, MI: Philosophy of Science Association.

Downes, S. (2011). Scientific Models. *Philosophy Compass*, 6(11), 757-764.

Eco, U. (1983). Horns, Hooves, Insteps: Some Hypotheses on Three Types of Abduction. In U. Eco, & T. Sebeok (eds.), *The Sign of Three: Dupin, Holmes, Peirce*(pp. 198-220). Bloomington, IN: Indiana University Press.

Eddington, A. (1920). Presidential Address to section A of the British Association at Cardiff, 24 Augustus 1920. *Observatory*, 43, 353.

Eddington, A. (1926). *The Internal Constitution of Stars*. Cambridge University Press.

Ellis, C., & Wooster, W. (1927). The Average Energy of Disintegration of Radium E. *Proceedings of the Royal Society A*, 117, 109-123.

Engel, P. (1998). Believing, holding true, and accepting. *Philosophical Explorations*, 1(2), 140-151.

Engel, P. (2012). Trust and the Doxastic Family. *Philosophical Studies*, 161(1), 17-26.

Epstein, S. (2008) Why Model? Unpublished Manuscript. Retrieved December 5, 2013, from http://www.santafe.edu/media/workingpapers/08-09-040.pdf

Faber, S., & Gallagher, J. (1979). Masses and Mass-to-Light Ratios of Galaxies. *Annual Review of Astronomy and Astrophysics*, 17, 135-187.

Faye, J. (2008). Copenhagen Interpretation of Quantum Mechanics. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition). Retrieved from http://plato.stanford.edu/archives/fall2008/entries/qm-copenhagen/

Favrholdt, D. (1994). Niels Bohr and Realism. In J. Faye, & H. Folse (eds.), *Niels Bohr and Contemporary Philosophy* (pp. 77-96). Dordrecht: Kluwer.

Feldman, R. (2003). *Epistemology*. Upper Saddle River, NJ: Prentice Hall.

Feldman, R., & Conee, E. (1985). Evidentialism. *Philosophical Studies*, 48(1), 15-34.

Fernandez, B., & Ripka, G. (2013). *Unravelling the Mystery of the Atomic Nucleus. A Sixty Year Journey 1896-1956*. New York: Springer.

Feynman, R. (1981). *Feynman on Doubt and Uncertainty - BBC Horizon* [Videoclip]. Retrieved December 5, 2013, from http://www.youtube.com/watch?v=I1tKEvN3DF0

Flach, P., & Kakas, A. (2000a). Abductive and Inductive Reasoning: Background and Issues. In Flach and Kakas (2000b, pp. 1-27). Dordrecht: Kluwer Academic Publishers.

Flach, P. & Kakas, A. (eds.) (2000b) *Abduction and Induction. Essays on their Relation and Integration*. Dordrecht: Kluwer Academic Publishers.

Foley, R. (1992). The epistemology of belief and the epistemology of degrees of belief. *American Philosophical Quarterly*, 29(2), 111-124.

Folse, H. (1994). Bohr's Framework of Complementarity and the Realism Debate. In J. Faye & H. Folse (eds.), *Niels Bohr and Contemporary Philosophy* (pp. 119-139). Dordrecht: Kluwer.

Frankish, K. (2004). *Mind and Supermind*. Cambridge University Press.

Frankish, K. (2009). Partial Belief and Flat-out Belief. In F. Huber, & C. Schmidt-Petri (eds.), *Degrees of Belief* (pp. 75-93). Synthese Library (Vol. 342). Dordrecht: Springer.

Franklin, A. (1993). *The Rise and Fall of the Fifth Force. Discovery, Pursuit and Justification in Modern Physics*. New York: American Institute of Physics.

Franklin, A. (1999). *Can that be right? Essays on Experiment, Evidence and Science*. Dordrecht: Kluwer.

Franklin, A. (2001). *Are there really neutrinos? An Evidential History*. Cambridge, MA: Perseus Books.

French, S. (2010). Keeping Quiet on the Ontology of Models. *Synthese*, 172(2), 231-249.

Frigg, R., & Hartmann, S. (2012). Models in Science. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2012 edition). Retrieved from `http://plato.stanford.edu/archives/fall2012/entries/models-science/`

Frigg, R., & Reiss, J. (2009). The Philosophy of Simulations: Hot New Issues or Same Old Stew? *Synthese*, 169, 593-613.

Gabbay, D., & Kruse, R. (eds.) (2000). *Abductive Reasoning and Learning* (Volume 4 of Handbook of Defeasible Reasoning and Uncertainty Management Systems). Dordrecht: Kluwer.

Gabbay, D., & Woods, J. (2006). Advice on Abductive Logic. *Logic Journal of the IGPL*, 14(2), 189-219.

Gamov, G. (1928). The Quantum Theory of Nuclear Disintegration. *Nature*, 122, 805-806.

García, A., &, Simari, G. (2004). Defeasible Logic Programming: An Argumentative Approach. *Theory and Practice of Logic Programming*, 4(1), 95138.

Gauderis, T. (2011). An Adaptive Logic based Approach to Abduction in AI. In S. Sardina and S. Vassos (eds.), *Proceedings of the 9th International Workshop on Nonmonotonic Reasoning, Action and Change (NRAC 2011)* (pp. 1-6). Retrieved from `http://ijcai-11.iiia.csic.es/files/proceedings/W4-%20NRAC11-Proceedings.pdf`

Gauderis, T. (2012). The Problem of Multiple Explanatory Hypotheses. In L. Demey and J. Devuyst (eds.), *Future directions for logic. Proceedings of PhDs in Logic III* (pp. 45-54). London: College Publications.

Gauderis, T. (2013a). Modelling Abduction in Science by means of a Modal Adaptive Logic. *Foundations of Science*, 18(4), 611-624.

Gauderis, T. (2013b). To Envision a New Particle or Change an Existing Law? Hypothesis Formation and Anomaly Resolution for the Curious Spectrum of the β Decay Spectrum. *Studies in History and Philosophy of Modern Physics*, in print.

Gauderis, T. (2013c). Pauli's Idea of the Neutrino: how Models in Physics allow to revive old Ideas for new Purposes. In L. Magnani (ed.), *Model-Based Reasoning in Science and Technology: Theoretical and Cognitive Issues* (pp. 449-461). Dordrecht: Springer.

Gauderis, T. (2014a). On Theoretical and Practical Doxastic Attitudes. *Manuscript submitted for publication*.

Gauderis, T. (2014b). Models and Hypotheses. *Manuscript submitted for publication*.

Gauderis, T., & Van De Putte, F. (2012). Abduction of Generalizations. *Theoria*, 27(3), 345-363.

Glass, D. (2006). *Experimental Design for Biologists*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Glass, D., & Hall, N. (2008). A Brief History of the Hypothesis. *Cell*, 134, 378-81.

Glymour, B. (2000). Data and Phenomena: A Distinction Reconsidered. *Erkenntnis*, 52, 29-37.

Giere, R. (1988). *Explaining Science. A Cognitive Approach*. University of Chicago Press.

Giere, R. (1999a). *Science withouth Laws*. University of Chicago Press.

Giere, R. (1999b). Using Models to Represent Reality. In L. Magnani, N. Nersessian, & P. Thagard (eds.), *Model-Based Reasoning in Scientific Discovery* (pp. 41-57). New York: Kluwer/Plenum.

Giere, R. (2004). How Models are Used to Represent Reality. *Philosophy of Science*, 71(5), 742-752.

Giere, R. (2008). Models, Metaphysics and Methodology. In S. Hartmann, C. Hoeffer, & L. Bovens (eds.), *Nancy Cartwright's Philosophy of Science* (pp. 123-133). New York: Routledge.

Giere, R. (2009). Why Scientific Models should not be regarded as Works of Fiction. In M. Surez (Ed.), *Fictions in Science. Philosophical Essays on Modeling and Idealisation* (pp. 248-258). London: Routledge.

Giere, R. (2010). An Agent-Based Conception of Models and Scientific Representation. *Synthese*, 172(2), 269-281.

Godfrey-Smith, P. (2006). The Strategy of Model-based Science. *Biology and Philosophy*, 21, 725-740.

Godfrey-Smith, P. (2009). Models and Fictions in Science. *Philosophical Studies*, 143, 101-116.

Goldberg, D. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*. Boston: Addison-Wesley.

Goldman, A. (1978a). Epistemics: the Regulative Theory of Cognition. *The Journal of Philosophy*, 75(10), 509-523.

Goldman, A. (1978b). Epistemology and the Psychology of Belief. *The Monist*, 61(4), 525-535.

Goldman, A. (1979). Varieties of Cognitive Appraisal. *Nos*, 13(1), 23-38.

Goldman, A. (2010). Why Social Epistemology is Real Epistemology. In A. Haddock, A. Millar, & D. Pritchard (eds.), *Social Epistemology* (pp. 1-28). Oxford University Press.

Gomes, R., Levison, H., Tsiganis, K., & Morbidelli, A. (2005). Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. *Nature*, 435, 466-469.

Good, I. (1983). The Philosophy of Exploratory Data Analysis. *Philosophy of Science*, 50(2), 283-295.

Guerra, F., Leone, M., & Robotti, N. (2012). When Energy Conservation Seems to Fail: The Prediction of the Neutrino. *Science and Education*, in print. doi:

10.1007/s11191-012-9567-0

Gurney, R. & Condon, E. (1928). Wave Mechanics and Radioactive Disintegration. *Nature*, 122, 439.

Hacking, I. (1983). *Representing and Intervening*. Cambridge University Press.

Hanson N. R. (1958). *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge University Press.

Hanson N. R. (1961). Is there a Logic of Scientific Discovery? In H. Feigl, & G. Maxwell (eds.), *Current Issues in the Philosophy of Science* (pp. 20-35). New York: Holt, Rinehart and Winston.

Hanson N. (1963). *The Concept of the Positron*. Cambridge University Press.

Harman, G. (1965). The inference to the best explanation. *Philosophical Review*, 74(1), 88-95.

Harris, T. (2003). Data Models and the Acquisition and Manipulation of Data. *Philosophy of Science*, 70(5), 1508-1517.

Hartmann, S. (1995) Models as a Tool for Theory Construction: Some Strategies from Preliminary Physics. In Herfel et al. (1995), p. 49-67.

Hartmann, W. (1986). Moon Origin: the Impact-trigger Hypothesis. In W. Hartmann, R. Phillips, & G. Taylor (eds.), *Origin of the Moon. Proceedings of the Conference, Kona, HI, October 13-16, 1984* (pp. 579-608). Houston, TX: Lunar and Planetary Institute.

Hartmann, W., & Davis, D. (1975). Satellite-sized Planetesimals and Lunar Origin. *Icarus*, 24, 504-515.

Hausman, D. (1998). *Causal Asymmetries*. Cambridge University Press.

Heilbron, J. (1985). The Earliest Missionaries of the Copenhagen Spirit. *Revue d'histoire des sciences*, 38(3-4), 195-230.

Heisenberg, W. (1967). Quantum Theory and its Interpretation. In S. Rozental (ed.), *Niels Bohr. His life and work as seen by his friends and colleagues* (pp. 94-108). Amsterdam: North Holland.

Hempel, C. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: Free Press.

Herfel, W., Krajewski, W., Niiniluoto, I., & Wojcicki, R. (eds.) (1995) *Theories and Models in Scientific Processes*. Poznan Studies in the Philosophy of Science and the Humanities 44. Amsterdam: Rodopi.

Hintikka, J. (1962). *Knowledge and Belief. An Introduction to the Logic of the Two Notions*. Ithaca, NY: Cornell University Press.

Hintikka, J. (1998). What is Abduction? The Fundamental Problem of Contemporary Epistemology. *Transactions of the Charles S. Peirce Society*, 34, 503-533.

Hoffmann, M. (2010). "Theoric Transformations" and a New Classification of Abductive Inferences. *Transactions of the Charles S. Peirce Society*, 46(4), 570-590.

Howard, D. (1994). What makes a Classical Concept Classical? Towards a Reconstruction of Niels Bohr's Philosophy of Physics. In J. Faye, & H. Folse (eds.), *Niels Bohr and Contemporary Philosophy* (pp. 201-229). Dordrecht: Kluwer.

Howard, D. (2004). Who Invented the "Copenhagen Interpretation"? A Study in Mythology. *Philosophy of Science*, 71(5), 669-682.

Hoyningen-Huene P. (2006). Context of Discovery versus Context of Justification and Thomas Kuhn. In J. Schickore, & F. Steinle (eds.), *Revisiting Discovery and Justification* (pp. 119-131). Dordrecht: Springer.

Hughes, J. (1993). *The Radioactivists: Community, Controversy and the Rise of Nuclear Physics*, PhD Dissertation, Cambridge University.

Hughes, J. (1998). 'Modernists with a Vengeance': Changing Cultures of Theory in Nuclear Science, 1920-1930. *Studies in History and Philosophy of Modern Physics*, 29(3), 339-367.

Hughes, J. (2000). 1932: The Annus Mirabilis of Nuclear Physics? *Physics World*, 13(7), 43-48.

Hughes, J. (2003). Radioactivity and Nuclear Physics. In M. Nye (ed.), *The Cambridge History of Science (Volume 5): The Modern Physical and Mathematical Sciences* (pp. 350-374). Cambridge University Press.

Hughes, J. (2009). Making Isotopes Matter: Francis Aston and the Mass-spectrograph. *Dynamis*, 29, 131-165.

Hughes, R. (1997). Models and Representation. *Philosophy of Science*, 64, Supplement, S325-S336.

Humphreys, P. (2004). *Extending Ourselves. Computational Science, Empiricism, and Scientific Method*. Oxford University Press.

Jeffrey, R. C. (1970). Dracula meets Wolfman: Acceptance versus Partial Belief. In M. Swain (ed.), *Induction, Acceptance and Rational Belief* (pp. 157-85). Synthese Library (Vol. 26). Dordrecht: Reidel Publishing Company.

Jensen, C. (2000). *Controversy and Consensus: Nuclear Beta Decay 1911-1934*. Basel: Birkhuser Verlag.

Kakas, A., &, Denecker, M. (2002). Abduction in Logic Programming. In A. Kakas and F. Sadri (eds.), *Computational Logic: Logic Programming and Beyond. Part I.* (pp. 402436). Dordrecht: Springer Verlag.

Kant, I. (1755/2012). Universal Natural History and Theory of the Heavens. In E. Watkins (ed.), *Natural Science. The Cambridge Edition of the Works of Immanuel Kant* (pp. 182-308). Cambridge University Press. (Original Article Published in German in 1755)

Kapitan, T. (1986). Deliberation and the Presumption of Open Alternatives. *The Philosophical Quarterly*, 36(143), 230-251.

Kapitan, T. (1992). Peirce and the Autonomy of Abductive Reasoning. *Erkenntnis*, 37(1), 1-29.

Kaplan, M. (1981). A Bayesian Theory of Rational Acceptance. *The Journal of Philosophy*, 78(6), 305-330.

Kaplan, M. (1996). *Decision Theory as Philosophy*. Cambridge University Press.

Khmelev, D. (2000) Disputed Authorship Resolution through Using Relative Empirical Entropy for Markov Chains of Letters in Human Language Texts. *Journal of Quantitative Linguistics*, 7(3), 201-207.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press.

Kitcher, P. (2013). Philosophy of Science. In *Encylopaedia Britannica Online*. Retrieved December 5, 2013, from `http://www.britannica.com/EBchecked/topic/528804/philosophy-of-science`

Knuuttila, T. (2011). Modelling and Representing: An Artefactual Approach to Model-Based Representation. *Studies in the History and Philosophy of Science A*, 42(2), 262-271.

Kojevnikov, A. (2011). Philosophical Rhetoric in Early Quantum Mechanics 1925-

1927: High Principles, Cultural Values and Professional Anxieties. In C. Carson, A. Kojevnikov, & H. Trischler (eds.), *Weimar Culture and Quantum Mechanics*. London: Imperial College Press.

Kosmann-Schwarzbach Y. (2010). *The Noether Theorems: Invariance and Conservation Laws in the Twentieth Century*. New York: Springer.

Kroon, F., & Voltolini, A. (2011). Fiction. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2011 Edition). Retrieved from http://plato.stanford.edu/archives/fall2011/entries/fiction/

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Kyburg, H. (1961). *Probability and the Logic of Rational Belief*. Middletown: Wesleyan University Press.

Lakatos, I. (1970). Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos, & A. Musgrave (eds.), *Criticism and the Growth of Knowledge* (pp. 97-196). Cambridge University Press.

Langley, P., Simon, H., Bradshaw, G., & Zytkow, J. (1987) *Scientific Discovery. Computational Explorations of the Creative Process*. Cambridge, MA: MIT Press.

Laplace, P. S. (1796/1830) *The System of the World*. Dublin University Press. (Original book published in French in 1796)

Laudan, L. (1977). *Progress and its Problems: towards a Theory of Scientific Growth*. Berkeley: University of California Press.

Laudan, L. (1980). Why was the Logic of Discovery Abandoned? In T. Nickles (ed.), *Scientific Discovery, Logic and Rationality* (pp. 173-183). Dordrecht: Reidel.

Lehrer, K. (2000). Acceptance and Belief Revisited. In P. Engel (ed.), *Believing and Accepting* (pp. 209-20). Dordrecht: Kluwer Academic Publishers.

Leuridan, B. (2013). The Structure of Scientific Theories, Explanation, and Unification. A Causal-Structural Account. *The British Journal for the Philosophy of Science*, in print. doi:10.1093/bjps/axt015

Levison, H., & Stewart, G. (2001). Remarks on Modeling the Formation of Uranus and Neptune. *Icarus*, 153, 224-228.

Lipton, P. (1991). *Inference to the Best Explanation*. London, Routledge.

Lipton, P. (2004). *Inference to the Best Explanation* (2nd edition). London: Routledge/Taylor and Francis Group.

Longair, M. (2013). *Quantum Concepts in Physics. An Alternative Approach to the Understanding of Quantum Mechanics*. Cambridge University Press.

Lycke, H. (2009). The Adaptive Logics Approach to Abduction. In E. Weber, T. Libert, P. Marage, & G. Vanpaemel (eds.), *Logic, Philosophy and History of Science in Belgium. Proceedings of the Young Researchers Days 2008* (pp. 35-41). Brussels: KVAB.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about Mechanisms. *Philosophy of Science*, 67(1), 1-25.

MacKinnon, E. (1977). Heisenberg, Models and the Rise of Matrix Mechanics. *Historical Studies in the Physical Sciences*, 8, 137-188.

Magnani, L. (2001). *Abduction, Reason and Science: Processes of Discovery and Explanation*. New York: Kluwer/Plenum.

Magnani, L. (2009). *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*. Berlin Heidelberg: Springer.

Mäki, U. (2011). Models and the Locus of their Truth. *Synthese*, 180, 47-63.

Makinson, D. C. (1965). The Paradox of the Preface. *Analysis*, 25, 205207.

Malley, M. (2011). *Radioactivity: a History of a Mysterious Science*. Oxford University Press.

Massimi, M. (2005). *Pauli's Exclusion Principle: the Origin and Validation of a Scientific Principle*. Cambridge University Press.

Massimi, M. (2007). Saving Unobservable Phenomena. *The British Journal for the Philosophy of Science*, 58(2), 235-262.

Massimi, M. (2011). From Data to Phenomena: a Kantian Stance. *Synthese*, 182(1), 101-116.

Mazumdar, I. (2005). Nucleosynthesis and Energy Production in Stars: Bethe's Crowning Achievement. *Resonance*, 10(10), 67-77.

McKaughan, D. J. (2008). From Ugly Duckling to Swan: C. S. Peirce, Abduction, and the Pursuit of Scientific Theories. *Transactions of the Charles S. Peirce Society*, 44(3), 446-468.

McMullin, E. (1992). *The Inference that Makes Science*. Milwaukee, WI: Marquette University Press.

Meheus, J. (1999). The Positivists' Approach to Scientific Discovery. *Philosophica*, 64(2). Retrieved December 5, 2013, from http://logica.ugent.be/philosophica/fulltexts.php

Meheus, J. (2007). Adaptive Logics for Abduction and the Explication of Explanation-seeking Processes. In O. Pombo, & A. Gerne (eds.), *Abduction and the Process of Scientific Discovery* (pp. 97-119). Lisboa: Centro de Filosofia das Ciencias.

Meheus, J. (2011). A Formal Logic for the Abduction of Singular Hypotheses. In D. Dieks, W. Gonzalez, S. Hartmann, T. Uebel & M. Weber (eds.), *Explanation, Prediction, and Confirmation* (pp. 93-108). Dordrecht: Springer.

Meheus, J., & Batens, D. (2006). A Formal Logic for Abductive Reasoning. *Logic Journal of the IGPL*, 14, 221-236.

Mehra, J. (1975). *The Solvay Conferences on Physics*. Dordrecht: Reidel.

Morbidelli, A., Levison, H., Tsiganis, K., & Gomes, R. (2005). Chaotic capture of Jupiter's Trojan astreroids in the early Solar System. *Nature*, 435, 462-465.

Morgan, M., & Morrison, M. (1999). *Models as Mediators. Perspectives on Natural and Social Sciences*. Cambridge University Press.

Morrison, M., & Morgan, M. (1999) Models as Mediating Instruments. In M. Morgan & M. Morrison (eds.), *Models as Mediators. Perspectives on Natural and Social Science* (pp. 10-37). Cambridge University Press.

Miller, A. (1984/1986). *Imagery in Scientific Thought*. Cambridge, MA: MIT Press. (Original work published in 1984)

Muller, F. A. (1997). The Equivalence Myth of Quantum Mechanics Part I. *Studies in History and Philosophy of Modern Physics*, 28(1), 35-61.

Navarro, J. (2004). New Entities, Old Paradigms: Elementary Particles in the 1930s. *LLULL, Revista de la Sociedad Española de Historia de las Ciencias y de las Técnicas*, 27, 435-464.

Navarro, J. (2008). Planck and de Broglie in the Thomson Family. In C. Joas, C. Lehner, & J. Renn (eds.), *HQ-1: Conference on the History of Quantum Physics* (pp. 233-251). Berlin: Max Planck Institut für Wissenschafstgeschichte. Retrieved December 5, 2013, from http://www.mpiwg-berlin.mpg.de/Preprints/P350.

PDF

Navarro, J. (2010). Electron Diffraction *chez* Thomson: Early Responses to Quantum Physics in Britain. *The British Journal for the History of Science*, 43(2), 245-275.

Nersessian, N. (2008). *Creating Scientific Concepts*. Cambridge, MA: MIT Press.

Nickles, T. (1978). Scientific Problems and Constraints. In P. Asquith and I. Hacking (eds.), PSA 1978, Proceedings of the Biennial Meeting of the Philosophy of Science Association (pp. 134-148). East Lansing, MI: Philosophy of Science Association.

Nickles, T. (1980). Introductory Essay: Scientific Discovery and the Future of Philosophy of Science. In T. Nickles (ed.), *Scientific Discovery, Logic and Rationality* (pp. 1-59). Dordrecht: Reidel.

Nickles, T. (1990). Discovery Logics. *Philosophica*, 45. Retrieved December 5, 2013, from http://logica.ugent.be/philosophica/fulltexts.php

Niiniluoto, I. (1999). Defending Abduction. *Philosophy of Science*, 66, Supplement, S436-S451.

Niiniluoto, I. (2011). Scientific Progress. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2011 Edition). Retrieved from http://plato.stanford.edu/archives/sum2011/entries/scientific-progress/

Nola, R., & Sankey, H. (2007). *Theories of Scientific Method. An Introduction*. Durham: Acumen.

Oppy, G. (1998). Propositional attitudes. In E. Craig (ed.), *Routledge Encyclopedia of Philosophy*. London: Routledge. Retrieved December 5, 2013, from http://www.rep.routledge.com/article/V028SECT1

Ostriker, J., & Peebles, P. (1973). A Numerical Study of the Stability of Flattened Galaxies: Or, can Cold Galaxies Survive? *The Astrophysical Journal*, 186, 467-480.

Pais, A. (1986). *Inward Bound. Of Matter and Forces in the Physical World*. Oxford University Press.

Pais, A. (2000). *The Genius of Science. A Portrait Gallery*. Oxford University Press.

Palmquist, S. (1987). Kant's Cosmogony Re-Evaluated. *Studies in the History and Philosophy of Science A*, 18(3), 255-269.

Paul, P. (2000). AI Approaches to Abduction. In D. Gabbay, &, R. Kruse (eds), *Abductive Reasoning and Uncertainty Management Systems*, Volume 4 of *Handbook of Defeasible Reasoning and Uncertainty Management Systems* (pp. 35-98). Dordrecht: Kluwer Academic Publishers.

Pauli, W. (1957/1964). Zur älteren und neureren Geschichte des Neutrinos. In R. Kronig, & V. Weisskopf (eds.), *Wolfgang Pauli. Collected Scientific Papers*. Vol. 2 (pp. 1313-1337). New York: Interscience Publishers. (Original article published in 1957)

Peierls, R. (1986). Introduction. In R. Peierls (ed.), *Niels Bohr Collected Works*. Vol. 9. Nuclear Physics (1929-1952) (pp. 3-84). Amsterdam: North Holland Physics.

Peirce, C. (1958). *Collected papers*. Cambridge, MA: Belknap Press of Harvard University Press.

Peirce, C. (1998). *The Essential Peirce, Volume 2*. Bloomington: Indiana University Press.

Perini, L. (2005). The Truth in Pictures. *Philosophy of Science*, 72(1), 262-285.

Perovic, S. (2008). Why Were Matrix Mechanics and Wave Mechanics Considered Equivalent? *Studies in History and Philosophy of Modern Physics*, 39, 444-461.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Routledge.

Plutynski, A. (2011) Four Problems of Abduction: a Brief History. *Journal of the International Society for the History of Philosophy of Science*, 1(2), 227-248.

Railton, P. (1981). Probability, Explanation and Information. *Synthese*, 48(2), 233-256.

Ramond, P. (2005). Neutrinos: precursors of new physics. *Comptes Rendus Physiques*, 6, 719-728.

Redhead, M. (1980). Models in Physics. *The British Journal for the Philosophy of Science*, 31(2), 145-163.

Reed, B. (2011) Certainty. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2011 Edition). Retrieved from `http://plato.stanford.edu/archives/win2011/entries/certainty/`

Reichenbach, H. (1938). *Experience and Prediction*. University of Chicago Press.

Reiter, R., &, De Kleer, J. (1987). Foundations of Assumption-based Truth Maintenance Systems: Preliminary Report. In *Proceedings of the Sixth National Conference on Articial Intelligence (AAAI-87)* (pp. 183-188).

Rescher, N. (1964). *Hypothetical Reasoning*. Amsterdam: North-Holland.

Rubin, V. (2003). A Brief History of Dark Matter. In M. Livio (ed.), *The Dark Universe. Matter, Energy and Gravity* (pp. 1-13). Cambridge University Press.

Rueger, A. (1992). Attitudes towards Infinities: Responses to Anomalies in Quantum Electrodynamics, 1927-1947. *Historical Studies in the Physical and Biological Sciences*, 22(2), 309-337.

Rutherford, E. (1911). The Scattering of α and β Particles by Matter and the Structure of the Atom. *Philosophical Magazine*, 6(21), 669-688.

Rutherford, E. (1914). The Structure of the Atom. *Philosophical Magazine*, 27, 488-498.

Rutherford, E. (1919). Collision of a Particles with Light Atoms. IV. An Anomalous Effect in Nitrogen. *Philosophical Magazine*, 37, 581-587.

Rutherford, E. (1920). Bakerian Lecture: Nuclear constitution of Atoms. *Proceedings of the Royal Society A*, 97, 374-400.

Rutherford, E. (1926). Atomic Nuclei and their Transformations. *Proceedings of the Physical Society*, 39, 359-372.

Rutherford, E., & Chadwick, J. (1929). Energy Relations in Artificial Disintegration. *Proceedings of the Cambridge Philosophical Society*, 25, 186-192.

Rutherford, E., Chadwick, J., & Ellis C. (1930). *Radiations from Radioactive Substances*. Cambridge University Press.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Schaffner, K. (1980). Discovery in the Biomedical Sciences. Logic or Irrational Intuition? In T. Nickles (ed.), *Scientific Discovery: Case Studies* (pp. 171-212). Dordrecht: Reidel.

Schickore, J., & Steinle, F. (2006) Introduction: Revisiting the Context Distinction. In J. Schickore, & F. Steinle (eds.), *Revisiting Discovery and Justification* (pp. vii-xix). Dordrecht: Springer.

Schurz, G. (2008a). Patterns of Abduction. *Synthese*, 164, 201-234.

Schurz, G. (2008b). Common Cause Abduction and the Formation of Theoretical Concepts in Science. In C. Dégremont, L. Keiff, & H. Rückert (eds.), *Dialogues Logics and other Strange Things. Essays in honour of Shahid Rahman* (pp. 337-364). London: College Publications.

Searle, J. (1983). *Intentionality*. Cambridge University Press.

Shah, M. (2007). Is it Justifiable to Abandon All Search for a Logic of Discovery? *International Studies in the Philosophy of Science*, 21(3), 253-269.

Shiaviv, G. (2010). *The Life of Stars: The Controversial Inception and Emergence of the Theory of Stellar Structure*. Heidelberg: Springer-Verlag.

Shomar, T. (2008). Bohr as a Phenomenological Realist. *Journal for General Philosophy of Science*, 39, 321-349.

Simon, H. (1973) Does Scientific Discovery have a Logic? *Philosophy of Science*, 40, 471-480.

Simon, H., Langley, P., & Bradshaw, G. (1981). Scientific Discovery as Problem Solving. *Synthese*, 47(1), 1-27.

Smith, G. (2008). Newton's Philosophiae Naturalis Principia Mathematica. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2008 Edition). Retrieved from http://plato.stanford.edu/archives/win2008/entries/newton-principia/

Soddy, F. (1921). The Origins of the Conceptions of Isotopes. In *Nobel Lectures, Chemistry, 1901-1921* (pp. 371-399). Amsterdam: Elsevier. Retrieved December 5, 2013, from http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1921/soddy-lecture.html

Sosa, E. (1991). *Knowledge in Perspective*. Cambridge University Press.

Stachel, J. (2009). Bohr and the Photon. In W. Myrvold, & J. Christian (eds.), *Quantum Reality, Relativistic Causality, and Closing the Epistemic Circle* (pp. 69-83). Dordrecht: Springer.

Steup, M. (1988). The Deontic Conception of Epistemic Justification. *Philosophical Studies*, 53(1), 65-84.

Steup, M. (2008). Doxastic Freedom. *Synthese*, 161(3), 375-92.

Stewart, G., & Levison, H. (1998). On the Formation of Uranus and Neptune. *Proceedings 29th Annual Lunar and Planetary Science Conference,* Abstract no. 1960.

Straßer, C. (2013). *Adaptive Logics for Defeasible Reasoning: Applications in Argumentation, Normative Reasoning and Default Reasoning.* Dordrecht: Springer.

Stuewer, R. (1983). The Nuclear Electron Hypothesis. In W. Shea (ed.), *Otto Hahn and the Rise of Nuclear Physics* (pp. 19-67). Dordrecht: Reidel.

Stuewer, R. (1985). Gamow's Theory of Alpha Decay. In E. Ullmann-Margalit (ed.), *The Kaleidoscope of Science: The Israel Colloquium Studies in History, Philosophy and Sociology of Science* (pp. 147-186). Dordrecht: Reidel.

Stuewer, R. (1986). Rutherford's Satellite Model of the Nucleus. *Historical Studies in the Physical and Biological Science*, 16(2), 321-352.

Suppe, F. (ed.) (1977). *The Structure of Scientific Theories* (2nd ed.). Urbana, IL: University of Illinois Press.

Suppe, F. (1989). *The Semantic Conception of Theories and Scientific Realism*. Urbana, IL: University of Illinois Press.

Suppes, P. (1960). A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences. *Synthese*, 12(2/3), 287-301.

Suppes, P. (1962). Models of Data. In E. Nagel, P. Suppes & A. Tarski (eds.), *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress* (pp. 252-261). Stanford University Press.

Tanona, S. (2004). Idealization and Formalism in Bohr's Approach to Quantum Theory. *Philosophy of Science*, 71(5), 683-695.

Teller, P. (2001). Twilight of the Perfect Model Model. *Erkenntnis*, 55(3), 393-415.

Thagard, P. (1988). *Computational Philosophy of Science*. Cambridge, MA: MIT Press.

Thagard, P. (1992). *Conceptual Revolutions*. Princeton University Press.

Thagard, P. (2012). *The Cognitive Science of Science. Explanation, Discovery, and Conceptual Change*. Cambridge, MA: MIT Press.

Thagard, P., & Shelley, C. (1997). Abductive reasoning: Logic, visual thinking, and coherence. In M.-L. Dalla Chiara et al. (eds.), *Logic and Scientific methods* (pp. 413-427). Dordrecht: Kluwer.

Thommes, E., Duncan, M., & Levison, H. (1999). The Formation of Uranus and Neptune in the Jupiter-Saturn region of the Solar System. *Nature*, 402, 635-638.

Thomson, G. P. (1928) The Disintegration of Radium E from the Point of View of Wave Mechanics. *Nature*, 121, 615-616.

Thomson, G. P. (1929) On the Waves associated with β Rays and the Relation between Free Electrons and their Waves. *Philosophical Magazine*, 42, 405-417.

Trimble, V. (1987). Existence and Nature of Dark Matter in the Universe. *Annual Review of Astronomy and Astrophysics*, 25, 425-472.

Tsiganis, K., Gomes, R., Morbidelli, A., & Levison, H. (2005). Origin of the orbital architecture of the giant planets of the Solar System. *Nature*, 435, 459-461.

Turri, J. (2012). A Puzzle about Withholding. *Philosophical Quarterly*, 62(247), 355-64.

Van De Putte, F. (2012). *Generic formats for prioritized adaptive logics. With applications in deontic logic, abduction and belief revision*. Ph. D. thesis, Ghent University. Retrieved December 5, 2013, from `http://logica.ugent.be/centrum/writings/doctoraten.php`

Van De Putte, F., &, Straßer, C. (2013). A logic for prioritized normative reasoning. *Journal of Logic and Computation*, 23(3), 563-583.

Van den Bergh, S. (1999). The Early History of Dark Matter. *Publications of the Astronomical Society of the Pacific*, 111, 657-660.

Van Fraassen, B. (1980). *The Scientific Image*. Oxford University Press.

Van Fraassen, B. (2008). *Scientific Representation: Paradoxes of Perspective*. Oxford University Press.

Von Helmholtz. H. (1856). On the Interaction of Natural Forces (Königsberg, 7 February 1854). *Philosophical Magazine*, 11(4), 489.

Vorms, M. (2011). Representing with Imaginary Models: Format Matters. *Studies in the History and Philosophy of Science A*, 42(2), 287-295.

Weinberg, S. (1999) What is Quantum Field Theory, and what did we think it was? In T. Cao (ed.), *Conceptual Foundations of Quantum Field Theory* (pp. 241-251). Cambridge University Press.

Weintraub, R. (1990). Decision-Theoretic Epistemology. *Synthese*, 83(1), 159-177.

Weisberg, M. (2007). Who is a Modeler? *The British Journal for the Philosophy of Science*, 58, 207-233.

Wieland, J. W. (2012). Can Pyrrhonists act normally? *Philosophical Explorations*, 15(3), 277-289.

Williams, S. (1989). Belief, Desire and the Praxis of Reasoning. *Proceedings of the Aristotelian Society*, 90, 119-142.

Williamson J. (2005). *Bayesian Nets and Causality*. Oxford University Press.

Wimsatt, W. (1987/2007). False Models as a Means to Truer Theories. In W. Wimsatt (ed.), *Re-engineering Philosophy for Limited Beings* (pp. 94-132). Cambride, MA: Harvard University Press. (Original article published in 1987)

Winsberg, E. (2010). *Science in the Age of Computer Simulation*. University of Chicago Press.

Wood, J. (1986). Moon over Mauna Loa – a Review of Hypotheses of Formation of Earth's Moon. In W. Hartmann, R. Phillips, & G. Taylor (eds.), *Origin of the Moon. Proceedings of the Conference, Kona, HI, October 13-16, 1984* (pp. 17-56). Houston, TX: Lunar and Planetary Institute.

Woodward, J. (1989). Data and Phenomena. *Synthese*, 79(3), 393-472.

Woodward, J. (2011). Data and Phenomena: a Restatement and Defense. *Synthese*, 182(1), 165-179.

Zwicky, F. (1933/2009). Republication of: The Redshift of Extragalactic Nebulae. *General Relativity and Gravitiation*, 41, 207-224. (Original article published in German in 1933)