Erik J. Olsson
Sebastian Enqvist
Editors

LOGIC, EPISTEMOLOGY, AND THE UNITY OF SCIENCE 21

Belief Revision meets Philosophy of Science

Springer

Belief Revision Meets Philosophy of Science

LOGIC, EPISTEMOLOGY, AND THE UNITY OF SCIENCE

VOLUME 21

Logic, Epistemology, and the Unity of Science aims to reconsider the question of the unity of science in light of recent developments in logic. At present, no single logical, semantical or methodological framework dominates the philosophy of science. However, the editors of this series believe that formal techniques like, for example, independence friendly logic, dialogical logics, multimodal logics, game theoretic semantics and linear logics, have the potential to cast new light on basic issues in the discussion of the unity of science.

This series provides a venue where philosophers and logicians can apply specific technical insights to fundamental philosophical problems. While the series is open to a wide variety of perspectives, including the study and analysis of argumentation and the critical discussion of the relationship between logic and the philosophy of science, the aim is to provide an integrated picture of the scientific enterprise in all its diversity.

For further volumes:
http://www.springer.com/series/6936

Erik J. Olsson · Sebastian Enqvist
Editors

Belief Revision Meets Philosophy of Science

Springer

Editors
Prof. Erik J. Olsson
University of Lund
Dept. Philosophy
Kungshuset
222 22 Lund
Sweden
erik_j.olsson@fil.lu.se

Sebastian Enqvist
University of Lund
Dept. Philosophy
Kungshuset
222 22 Lund
Sweden
Sebastian.Enqvist@fil.lu.se

ISBN 978-90-481-9608-1 e-ISBN 978-90-481-9609-8
DOI 10.1007/978-90-481-9609-8
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2010938250

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Editor’s Introduction

Belief revision theory and philosophy of science both aspire to shed light on the dynamics of knowledge – on how our view of the world changes (typically) in the light of new evidence. Yet these two areas of research have long seemed strangely detached from each other, as witnessed by the small number of cross-references and researchers working in both domains. One may speculate as to what has brought about this surprising, and perhaps unfortunate, state of affairs. One factor may be that while belief revision theory has traditionally been pursued in a bottom-up manner, focusing on the endeavors of single inquirers, philosophers of science, inspired by logical empiricism, have tended to be more interested in science as a multi-agent or agent-independent phenomenon.

The aim of this volume is to build bridges between these two areas of study, the basic question being how they can inform each other. The contributors seek their answers by relating the logic of belief revision to such concepts as explanation, coherence, induction, abduction, interrogative logic, conceptual spaces, structuralism, idealization, research agendas, minimal change and informational economy.

Our aim in putting together this volume has been to provide a number of new perspectives that are likely to stir research in new directions, as well as to establish new connections between areas previously assumed unrelated, e.g. between belief revision, conceptual spaces and structuralism. The result is, we believe, a coherent volume of individual papers complementing and shedding light on each other.

We have been very fortunate to be able to attract, as contributors to this volume, some of the best researchers in their respective fields of philosophy, cognitive science and logic, as well as some exceptional scholars in the younger generation. We are extremely proud to have their articles in the volume, and we thank them all for their dedication and fine scholarship.

We hope that this volume will contribute to a greater degree of interaction between the fields of belief revision theory and the philosophy of science. For this reason, we hope that the essays included here will be read by researchers in both fields. However, the fundamental concepts in the philosophy of science are probably known to a significantly wider audience than belief revision theory. For this reason, in order to facilitate for those readers not previously acquainted with belief revision theory, a brief introduction seems in order.

Belief revision theory is a branch of formal epistemology that studies rational changes in states of belief, or *rational theory changes*. The classic framework here, and the best known one, is the socalled AGM theory, named after its creators Carlos Alchourrón, Peter Gärdenfors and David Makinson, which originated in a series of papers in the eighties. Since then, a rather large number of alternative frameworks have emerged, which generalize or deviate from AGM in various ways. But for the purpose of introducing the novice to belief revision theory, a summary of the AGM theory will be enough to give a sense of how belief revision works and what it is about.

The basic way of representing belief states in the AGM theory is to equate an agent's state of belief (at some given time) with a logically closed set of sentences K, i.e. such that K = Cn(K) where Cn denotes the operation of logical closure. Thus, the logical consequences of an agent's beliefs are also counted as beliefs in AGM. Such a set K is sometimes referred to as a "theory", sometimes as a "belief set".

Let α = "Charles is in his office" and let β = "Charles is at home". Let $\mathrm{K}_1 = \mathrm{Cn}(\{\sim\alpha\rightarrow\beta\})$. Then K_1 expresses the state of belief where it is believed that "if Charles is not in his office, then he is at home". Now, say that we learn that Charles is not in his office. Then we need to alter our initial belief state K_1 to include this new belief. How do we do this, rationally? Simply adding the sentence $\sim\alpha$ to K_1 will not do, since the set $K_1 \cup \{\sim\alpha\}$ is not logically closed, i.e. it is not a *bona fide* belief set. What we need to do is to first add $\sim\alpha$ to K_1, and then close the result under logical consequences. This general recipe gives rise to the operation of *expansion*, one of the three basic operations on belief sets in AGM. The expansion of a belief set *K* by a sentence α, denoted $K + \alpha$, is defined by setting

$$K + \alpha =_{\text{df.}} \mathrm{Cn}\,(K \cup \{\alpha\})$$

In our case, the new belief set $K_1 + \sim\alpha$ after learning that Charles is not in his office is $\mathrm{Cn}\,(\{\sim\alpha \rightarrow \beta, \sim\alpha\})$. This new belief set includes the sentence β, i.e. upon learning that Charles is not in his office, we come to believe that Charles is not at home.

Let us denote this new belief set by K_2. Say that we now learn that Charles is not at home. Then we want to add the sentence $\sim\beta$ to K_2. If we expand again at this point, we run into some trouble: since β is in K_2, the expanded set $K_2 + \sim \beta$ contains a contradiction, and so by logical closure (assuming that the underlying logic contains the classical validities) $K_2 + \sim \beta$ will contain *all* sentences in the language – it is a state where we believe everything. This awkward situation has been referred to as "epistemic hell", and is clearly unattractive. We would like to avoid this consequence, and update the initial state K_2 with $\sim\beta$ while preserving consistency. This enters us into the area where belief revision theory becomes interesting.

The operation used for updating a belief state while maintaing consistency in cases like the one above is called *revision*. The revision of a belief set *K* by a sentence α is denoted $K^*\alpha$. A useful way of analysing the problem of revision is by introducing a third operator: *contraction*. Take the belief set K_2 again; we would like to add $\sim\beta$ to this belief set without creating an inconsistency. In order to do this, we have to remove something, in particular we have to remove the negation of $\sim\beta$, or equivalently, β. In order to do this, we have to remove some of our earlier beliefs, $\sim\alpha$ and $\sim \alpha \rightarrow \beta$ since they jointly imply β. Now, we could remove *both* these beliefs of course – this would certainly make room for $\sim\beta$ – but that would not be very *economical*, since removing either of $\sim\alpha$ or $\sim \alpha \rightarrow \beta$ while keeping the other would suffice. Beliefs are valuable things, and a rational agent should not be willing to give up more of his beliefs than what is necessary. This intuition is one of the guiding principles in AGM, and usually goes under the name of *the principle of minimal change*. Contraction is intended as an operation that removes a sentence from a given belief set in accordance with this principle.

The contraction of a belief set *K* by a sentence α is denoted $K \div \alpha$. Given a suitable operator $\div$ of contraction, we can define a revision operator * by setting, for all α:

$$K^{*}\alpha =_{\text{df}} .(K\div \sim \alpha) + \alpha$$

Intuitively: to revise *K* by α is to first remove the negation of α (to "make room" for α) and then expand with α. The definition is commonly known as the *Levi identity*, after Isaac Levi.

With this definition in place, the problem of revision reduces to the problem of contraction. Thus we are left with the task of devising a satisfactory account of contraction. The way AGM handles this problem can be divided into two distinct approaches: we may call one the *axiomatic* approach, and the other the *constructive* approach. The axiomatic approach consists in narrowing down the class of rationally admissible contraction functions by setting up a list of (intuitively plausible) postulates for contraction. The following six postulates are known as the *basic AGM postulates* for contraction:

(closure) $K \div \alpha = \text{Cn}(K \div \alpha)$
(success) $\alpha \notin \text{Cn}(\emptyset)$ implies $\alpha \notin \text{K} \div \alpha$
(inclusion) $K \div \alpha \subseteq K$
(vacuity) $\alpha \notin K$ implies $K \div \alpha = K$
(extensionality) $\text{Cn}(\alpha) = \text{Cn}(\beta)$ implies $K \div \alpha = K \div \beta$
(recovery) $K \subseteq (K \div \underset{\cdot\cdot}{\alpha}) + \ \alpha$

This list is usually extended with the following two supplementary postulates:

(conjunctive inclusion) $\alpha \notin K \div (\alpha \wedge \beta)$ implies $K \div (\alpha \wedge \beta) \subseteq K \div \alpha$
(conjunctive overlap) $K \div \alpha \cap K \div \beta \subseteq K \div (\alpha \wedge \beta)$

The postulates all have some intuitive justification. For instance, the *success* postulate says that an admissible contraction operator $\div$ should do its job properly whenever possible, i.e. if α is not a tautology and so can be removed from *K* while maintaining logical closure, then $\div$ should succesfully remove it. The (highly controversial) *recovery* postulate gives a first formal expresssion to the principle of minimal change: when we contract by α, we should keep so much information that we can regain all our initial beliefs by expanding with α again. A similar set of postulates for revision exists, and these postulates are satisfied by any revision function which is defined from a contraction function satisfying the postulates above (and conversely, if a contraction function gives rise to a revision function that satisfies the AGM postulates for revision, then this function must satisfy the above postulates for contraction).

The *constructive* approach consists in devising explicit constructions of contraction functions. Apart from the socalled *partial meet* contractions, the construction most frequently discussed these days is probably the method of *entrenchment based* contraction. This method uses an auxiliary concept called an *entrenchment relation*, which is a binary relation over the language associated with a given belief set K, satisfying the following postulates:

(transitivity) $\alpha \leq \beta$ and $\beta \leq \chi$ implies $\alpha \leq \chi$
(dominance) $\beta \in \mathrm{Cn}(\alpha)$ implies $\alpha \leq \beta$
(conjunctiveness) either $\alpha \leq (\alpha \wedge \beta)$ or $\beta \leq (\alpha \wedge \beta)$
(minimality) if $\bot \notin K$ then $\alpha \notin K$ iff $\alpha \leq \beta$ for all $\beta \in \mathrm{L}$
(maximality) $\beta \leq \alpha$ for all β only if $\alpha \in \mathrm{Cn}(\emptyset)$

Given a belief set K with an associated entrenchment order $\leq$, we can define a corresponding entrenchment based contraction function $\div$ by setting, for all α:

$$K \div \alpha =_{\text{df.}} \{\beta \in K \mid \alpha < \alpha \vee \beta \text{ or } \alpha \in \mathrm{Cn}(\emptyset)\}$$

The intuition here is that the entrenchment order encodes how entrenched the various beliefs in K are in comparison to each other, or which beliefs the agent would prefer to give up if a choice has to be made. Entrenchment based contractions are then designed so as to remove less entrenched beliefs in favor of the more entrenched ones.

How are the axiomatic approach and the constructive approach related to each other? It turns out that they are very closely related: every entrenchment based contraction satisfies the AGM postulates for contraction (including the supplementary postulates), and vice versa, if a contraction function satisfies the AGM postulates, then there exists an entrenchment order which defines it. In a sense, the AGM postulates are *sound and complete* with respect to entrenchment based contractions. This result is one of the celebrated *representation theorems* of the AGM framework.

Drafts of most of the papers that appear in this volume were originally presented at the first *Science in Flux* conference, organized by Olsson, at Lund University in 2007. (A follow-up conference was organized by Pierre Wagner in 2008 at the CNRS in Paris.) Before they were submitted in their final versions, the papers were revised, often substantially, in order to accommodate various critical points that emerged in the (very lively) discussion at the conference. It has been pointed out to us that there is already a book called "Science in Flux", by J. Agassi, which appears in the Boston Studies in the 1970s. As far as we can see, there is little overlap between the books, and little risk for conflating one with the other, and so we hope we are excused for reusing the title.

We shall briefly present each of the individual contributions:

Raúl Carnota and Ricardo Rodríguez

In their contribution, Raúl Carnota och Ricardo Rodríguez take a closer look at the history behind the influential AGM model of belief revision due to Alchourron, Gärdenfors and Makinson. The AGM theory, as described in the seminal 1985 paper "On The Logic of Theory Change: Partial Meet Contractions and Revision Functions", had a major influence in most subsequent work on belief change. In particular, the constructive approach spelled out in the paper was adopted in AI as a paradigm for how to specify updates of knowledge bases. Throughout the years there has been a steady stream of references to that original AGM paper. Going one step further, Carnota and Rodréguez ask themselves why the AGM theory was so readily accepted within the AI community, their answer being partly that the theory was put forward at a critical time in the history of AI at which the problem of how to update knowledge bases in the face of input possibly inconsistent with the previous corpus was taken to be of utmost importance. The paper also contains a qualitative and quantitative evaluation of the impact of the AGM theory in AI research as well as an account of how the theory has subsequently been developed in different directions.

Sven Ove Hansson

There is a clear connection between belief revision theory and one of the major problems within the philosophy of science, the problem of modelling and understanding the dynamics of empirical theories; both these fields of research deal with the way theories are updated in the light of incoming data. In fact, several of the most influential ideas in 20th century philosophy of science, e.g. Popper's hypothetico-deductive method, Kuhn's theory of paradigm shifts, Lakatos's ideas concerning the "hard core" and the "protective belt" etc. seem to be in essence theories about the dynamics of theories. That is, these theories apparently address the very subject matter of belief revision theory. Given this connection, it is somewhat striking that there has been so little contact between the two fields. Hansson draws the conclusion that belief revision theory as it stands is unsuitable for modelling changes in empirical theories, and sets out to develop a framework which is better suited for this task. He draws some first contours of a model where scientific change is treated as a *partly accumulative* process, through which observational data is added piecewise and theoretical hypotheses are added by a closure operator representing "inference to the best explanation". A set of postulates for this operator is provided, and three versions of a model of theory change are introduced and discussed in the text.

Hans Rott

Scientific change is also the main issue addressed by Hans Rott. In his paper, Rott poses the problem of how, exactly, to explicate the Lakatosian notion of a "progressive problem shift" that plays a crucial role in the understanding of how research programs develop over time, the basic idea being that such a shift is progressive if the transition to a successor theory T′ can somehow explain both the success of its predecessor theory T and the failure of T. That would mean that we would have an account that goes deeper than a plain approximate agreement of the empirical predictions made by the two theories. Rott proposes to accomplish this explication using factual, potential and counterfactual explanations. Thus the successor theory can explain the success of the predecessor theory by implying that the predecessor theory would have been true, had its application conditions been satisfied, but because they are not, the predecessor theory is false. This gives an account of how a single theory can speak, as it were, at the same time in favor of and against another theory. Rott uses the AGM postulates for rational belief changes to spell out these ideas in formal terms, thus connecting a central issue in the philosophy of science with standard theorizing in the logic of belief revision.

Gerhard Schurz

The article by Gerhard Schurz deals with the problem of abduction in the context of belief revision theory. It is noted that belief revision in its usual form lacks an account of the ability of inquiring agents to *learn* in the sense of forming generalized hypotheses on the basis of incoming data, a point which is illustrated with some formal theorems. If belief revision theory is to be applied to problems in the philosophy of science in a fruitful way, then extending the classical models of belief revision to encompass abduction seems to be of great importance: science is not simply the act of collecting and storing data. Arguably, the most important task of scientists is to formulate and test hypotheses on the basis of the empirical data, and any model which claims to capture scientific change must therefore provide an account of this process. After having discussed two attempts in the literature to incorporate such creative elements of belief formation into belief revision theory, Schurz develops an alternative theory based on a theory of abduction developed elsewhere. Various types of abductive revision and expansion are investigated; one of the main observations being that the well known Levi identity breaks down in the context of abductive belief revision.

Sebastian Enqvist

Within the AGM theory, and in belief revision more generally, it is assumed that theories can be represented as sets of sentences or statements. Within the philosophy of science, this seemingly harmless assumption has been challenged by the so-called structuralist theory of science, the seminal exposition of which is Joseph Sneed's 1971 book "The Logical Structure of Mathematical Physics". Structuralism instead reconstructs empirical theories as set theoretical structures, which have no propositional content in themselves, but which can be used to make empirically testable statements about the world. In his paper, Enqvist develops a model of theory change which is founded on the structuralist notion of a "theory net", rather than the classical conception as a set of sentences in a formal language. The notion of theory nets gives an explicit account of the "deep structure" of empirical theories, and Enqvist argues that the fine grained structure of the structuralist's way of representing theories may shed some new light on the problem of theory change. In particular, the specialization relation which forms an essential part of a theory net is investigated in the context of contraction. It is argued that specialization plays an important role in contraction, but that a separate notion of corroboration should also be taken into account. Finally, the possibility of distinguishing novel types of theory changes within the framework is discussed.

Peter Gärdenfors and Frank Zenker

The aim of the contribution by Gärdenfors and Zenker is to apply conceptual spaces, as developed by Gärdenfors in his 2000 book "Conceptual Spaces – The Geometry of Thought", to the dynamics of scientific theories. According to the theory of conceptual spaces, dimensions and their relations provide a topological representation of a concept's constituents and their mode of combination where concepts are seen as n-dimensional geometrical structures and conceptual change, consequently, means the dynamic development of these structures. Gärdenfors and Zenker take the structuralist framework, also addressed by Enqvist, to be a useful contrast to their own thinking on the matter, and in one section they argue that the central notions of structuralism can be expressed in terms of conceptual spaces. As they also observe, however, structuralism is problematic in the context of revolutionary changes. It is here that conceptual spaces may have a distinct worth: as Gärdenfors and Zenker argue, many, or even all, radical changes can be modeled as one of four types of increasingly severe transformations of conceptual spaces.

Bengt Hansson

Bengt Hansson begins his essay by pointing out a certain type of adequacy condition which seems to be present in mature theories like, for example, Newtonian mechanics and which may be demanded of a good theory: a condition which he calls *conceptual closure*. Intuitively, a theory can be said to satisfy conceptual closure if it describes a delineated part of the world which is closed in the sense that everything which influences the factors taken into account in the description is also explicitly part of the description. More formally, Hansson shows that the condition of conceptual closure can be characterized in terms of the notions of homomorphisms and commutative diagrams. While the notion of conceptual closure is derived from examples within the empirical sciences and physics in particular, Hansson argues that the same condition may be required of an adequate theory of belief change or theory change. Some of the well known approaches to modelling epistemic states in belief revision are questioned with regards to whether they satisfy conceptual closure, and the classical principle of minimal change is briefly discussed.

Horacio Arlo-Costa and Arthur Paul Pedersen

Horacio Arlo-Costa and Arthur Paul Pedersen's paper concerns the theory of rational choice and its application to belief revision theory. Correlations between principles of belief change and principles of rational choice have been studied at length by Hans Rott, in particular in his 2001 book "Change, Choice and Inference". This correlation has the nice feature of allowing us to think of principles of belief change in choice theoretic terms, and thus to criticize, assess or justify principles of belief change on choice theoretic grounds. This line of thought is continued in Arlo-Costa and Pedersen's paper, with focus on a particular issue in the theory of rational choice: traditional rational choice theory has been criticized by Amartya Sen for being unable to deal with the role *social norms* play in some cases. After having presented a recent model of rational choice which is intended to take social norms into account, Arlo-Costa and Pedersen develop an alternative model which generalizes the former one. The theory is then applied to belief revision theory. It is argued that the existence of social norms gives rise to counterexamples to some of the classical postulates for belief change, and they attempt to develop a model for "norm-inclusive belief change" which accounts for these examples. Axioms for norm-inclusive belief change are given and related to principles of rational choice by formal correspondence results.

David Westlund

Continuing the choice-theoretic theme, David Westlund's essay concerns the possibility of extending the AGM framework for belief revision to cover the case where collective agents change their beliefs as the result of individual belief changes. He points out the importance of this issue for applications of belief revision to the philosophy of science: scientific theories are not the beliefs of a single person and so changes in the beliefs of the scientific community are not changes in the beliefs of an individual, though what the scientific community may be said to believe is obviously somehow dependent on what individual scientists believe. That is, science is a *social* enterprise, and the dynamics of scientific theories thus contains a social component: to understand how theories change, it is not enough to understand how individual agents change their beliefs. We must also provide an account of how individual belief systems give rise to the beliefs of a scientific community, and how individual belief changes give rise to changes in what the community believes. In order to study this feature of scientific change, Westlund introduces the notion of a *merging function* (borrowed from computer science), which is a function taking a family of belief sets to a "merged" belief set. In terms of these merging functions, some negative results on collective belief change are demonstrated, showing that certain conditions on collective belief change cannot be consistently fulfilled.

Emmanuel Genot

Emmanuel Genot's paper builds on a proposal due to Erik J. Olsson and David Westlund, to extend the representation of epistemic states in the AGM theory to include an account of the research agenda of the inquiring agent. Genot relates this suggestion to the socalled *interrogative model of inquiry* (IMI) due to Jaakko Hintikka. Hintikka's model is a general model of inquiry, which is "Socratic" in the sense that inquiry is treated as a process of asking questions and drawing conclusions from the answers received. Formally, the interrogative model treats inquiry as a game where an agent may make *interrogative moves*, i.e. ask questions in order to retrieve new information, interspersed with *deductive moves*, where information is deduced from given premises together with answers received to previously asked questions. The connection between Hintikka's model and Olsson and Westlund's proposal to extend belief revision with a research agenda is clear: both argue that asking questions somehow plays a central role in inquiry. Apart from this informal similarity, Genot establishes a formal connection between the two approaches, so that results and research problems may be transferred from one to the other. In particular, he applies results from the interrogative model, most notably the so-called “Yes-No theorem”, to attack the problem of updating agendas in the case of contraction – a problem which is largely left open in Olsson and Westlund’s previous work on the agenda. Borrowing some game theoretical terminology, Genot distinguishes between “stategic” and “extensive” update of questions.

Erik J. Olsson

Olsson’s point of departure is a longstanding issue in philosophy of science and belief revision theory concerning what to do when, after thorough investigation, more than one theory or belief set stands out as highly reasonable given all the data. Otto Neurath, the logical empiricist, suggested that it would, in such circumstances, be rationally admissible to decide the matter by coin-flipping. In belief revision theory, this debate has taken the form of a dispute between “functionalists” and “relationalists”. Functionalists hold that the result of revising a cognitive state with some new datum is a unique rationally determined belief state. Relationalists, by means of contrast, insist that there may be several rationally admissible results. In an attempt to contribute to conceptual clarity, Olsson distinguishes between three ways of drawing the functionalist-relationalist distinction. This gives rise to six non-contradictory overall positions. He proceeds to consider arguments in the literature for excluding some of these positions on logical, philosophical or other grounds. Finally, Olsson argues that part of what feeds the functionalist relationalist controversy is a false dilemma based on an implausible conception of what it means rationally to suspend judgment. Making this precise requires a formal framework of the kind that includes a representation of the agent’s research agenda. Here a connection emerges with the paper by Emmanuel Genot in which the notion of an agenda also plays a prominent role.

Caroline Semmling and Heinrich Wansing

Caroline Semmling and Heinrich Wansing, in their contribution to the present volume, investigate an extension of the socalled Stit Theory in which an operator of “deliberatively seeing to it that” (*dstit*) is investigated. This operator allows representation of deliberate actions of agents in a modal object language. Semmling and Wansing extend this theory by adding operators for beliefs, intentions and desires, forming what they call *bdi-stit* logic. They present a semantics for bdi-stit logic, as well as a complete tableaux-style proof system. With this extension in place, it is possible to express an interesting special class of actions: actions involving seeing-to-it-that you believe something, desire something or intend something. That is, we can express the formation of beliefs, desires or intentions as deliberate actions. In particular, the “belief” part of the logic becomes interesting in connection with belief change: seeing-to-it-that you believe in α looks a lot like *expanding* with α. Semmling and Wansing attempt to make this connection explicit, and apply their bdi-stit logic to AGM theory. It is shown how the language of *bdi-stit* logic can be used to define operators for expansion, revision and contraction. Since the belief fragment of the bdi-stit logic in itself contains only the means to express beliefs, on the one hand, and actions in terms of seeing-to-it-that something is brought about on the other, the result is an analysis of the concepts of expansion, contraction and revision in terms of deliberate actions. It turns out that, when translated with the help of these defined operators, several of the AGM postulates are provable in *bdi-stit* logic.

Isaac Levi

Isaac Levi's seminal work on belief revision theory is widely acclaimed and rightly so: many of the issues that the AGM theory dealt with can be traced back to their origin in Levi's work in the 1970s. This includes the classification of changes of belief in terms of expansions and contractions, and the famous so-called Levi identity according to which a revision can be reconstructed as a contraction followed by an expansion. Levi is also well known for his radical theory of knowledge according to which knowledge is nothing but true full belief, where belief is understood as "standard of serious possibility". This thesis, which runs counter to a long tradition in epistemology according to which knowledge entails that the knower has good reasons for his belief, is intimately related to his endorsement of Peirce's belief-doubt model stating that full belief is a state which is satisfactory in itself and therefore in no need of justification. For Levi, the issue of justification arises only in connection with *changes* of belief. Thus, while an inquirer can be justified or unjustified in expanding his corpus by a given item of information, once this expansion has been implemented there is no issue of justification anymore. In his contribution to this volume, Levi, among other things, usefully contrasts his view with conflicting accounts of knowledge in the recent epistemological literature, focusing on the theory advocated by Edward Craig, and he defends and further clarifies his taxonomy of different types of belief change and their decision-theoretic justification.

Paul Thagard

Paul Thagard has, in a number of papers and books, argued in favor of a coherence based approach to belief revision and reasoning in general. In his contribution to this volume he continues this line of work, showing how the recent debate about climate change in the scientific and political spheres can be modeled in his theory of explanatory coherence. The theory is based on a number of general coherence principles, such as "Similar hypotheses that explain similar pieces of evidence cohere" and "The acceptance of a proposition in a system of propositions depends on its coherence with them". While these principles do not fully determine coherence-based acceptance, Thagard has developed computer-implemented algorithms that can compute acceptance and rejection of propositions on the basis of coherence relations. In his paper, Thagard argues that his model can give a good account of the three main kinds of belief change: expansion, revision and contraction. For instance, "[e]xpansion takes place when a new proposition is introduced into a belief system, becoming accepted if and only if doing so maximizes coherence". Since the driving force behind Thagard's system is to satisfy the goal of maximizing coherence, coherence-based belief revision does not satisfy the AGM postulates, which are rather aimed at capturing minimal change. Thagard argues that this is as it should be, and that the AGM dictum to seek maximally conservative revisions is ill-motivated.

Thagard also maintains that his model has certain computational advantages over competing Bayesian accounts of belief change.

Jonas Nilsson and Sten Lindström

Belief revision theory is concerned with the rationality of changes of belief. The purpose of Jonas Nilsson and Sten Lindström's contribution is rather to inquire into the rationality of changes of methodology, as when a methodological rule such as "prefer simpler theories to less simple ones" is revised in favor of "a successor theory must retain all the corroborated empirical content of its predecessors". While there is a long-standing debate in philosophy of science concerning the proper account of methodological change and its rationality, few, if any of the resulting theories have reached the same level of precision as, for instance, the AGM theory of belief revision. For that reason, Nilsson and Lindström propose to investigate methodological change by means of formal methods. Special problems arise here because they would like to have a bootstrap theory rather than a static theory. According to the former, but not the latter, all standards are open for criticism and correction, meaning that there is no distinguished set of standards that can be used to correct other standards but which is itself immune to objections. While their paper is only a first step in this direction, they propose a number of general principles governing such change. According to their principle of Prospective Acceptability, for instance, a revised set of standards S_2 must be better than the original set S_1, according to the members of S_1, except for those standards in S_1 that are criticized and revised. Finally, Nilsson and Lindström discuss the prospects of modelling standards as a kind of beliefs, namely as beliefs about what rationality requires one to do. This would make methodological change a species of belief change, and raise the expectation that postulates for belief change should also hold for methodological change.

It strikes us that there is another way to connect the two areas of belief and methodological change, viz., to view beliefs, following Isaac Levi, as a certain kind of standards, i.e., as standards of serious possibility. This might be an interesting alternative for the following reason. As Nilsson and Lindström show, the AGM preservation principle is not plausible when viewed as a principle for methodological change, contrary to what we would expect if methodological standards are species of beliefs (and the AGM postulates are taken to apply to all kinds of belief change, rather than merely to, say, our beliefs about the world). However, if we instead view beliefs as species of standards, no such expectation seems to arise. From that perspective, the AGM axioms are about a special class of standards, namely, of standards of serious possibility, and there is little reason to believe that everything that can plausibly be said about the rationality with respect to that special class should be true of standard change in general. An interesting problem for further work would be how to weaken the AGM axioms (or other alternative sets of postulates) for changing standards of serious possibility, so that they hold for standards in general.

We mention this as but one example of how the papers in this volume, individually or in combination, may raise new and potentially fruitful research questions.

We believe there are a great many different angles from which the papers in this volume can be studied, and we invite the reader to be as creative as the contributors in deepening the various connections that emerge between belief revision theory and philosophy of science, or in thinking of new ways entirely for how these areas of philosophical and logical research can be brought into closer contact.

Lund, Sweden

Erik J. Olsson
Sebastian Enqvist

Contents

Contributors

Horacio Arló-Costa Department of Philosophy, Carnegie Mellon University, Baker Hall 135, Pittsburgh, PA 15213, USA, hcosta@andrew.cmu.edu

Raúl Carnota Universidad Nacional de Tres de Febrero, Av. del Libertador Gral, San Martín 5569 Piso 14 'B', 1426, Buenos Aires, Argentina, carnotaraul@gmail.com

Sebastian Enqvist Department of Philosophy, Lund University, Kungshuset Lundagård, 222 22 Lund, Sebastian.Enqvist@fil.lu.se

Scott Findlay University of Western Ontario, London, Ontario, sfindla@uwo.ca

Peter Gärdenfors Department of Philosophy, University of Lund, Kungshuset, Lundagard, 22 222 Lund, Sweden, peter.gardenfors@lucs.lu.se

Emmanuel Genot Department of Philosophy, University of Lille 3, Lille, France, emmanuel.genot@etu.univ-lille3.fr

Bengt Hansson Lund University, Lund, Sweden, bengt.hansson@fil.lu.se

Sven Ove Hansson Division of Philosophy, Royal Institute of Technology, Tekniktingen 78, 100 44 Stockholm, Sweden, soh@kth.se

Isaac Levi John Dewey Professor Emeritus, Columbia University, New York, NY, USA, levi@columbia.edu

Sten Lindström Department of Historical, Philosophical and Religious Studies, Umeå University, Sweden, sten.lindstrom@philos.umu.se

Jonas Nilsson Department of Historical, Philosophical and Religious Studies, Umeå University, Sweden, jonas.nilsson@philos.umu.se

Erik J. Olsson Lund University, Lund, Sweden, Erik_J.Olsson@fil.lu.se

Arthur Paul Pedersen Department of Philosophy, Carnegie Mellon University, Baker Hall 135, Pittsburgh, PA 15213, USA, apaulpedersen@cmu.edu

Ricardo Rodríguez Departamento de Computación de la Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Intendente Güiraldes 2160 - Ciudad Universitaria - C1428EGA, Buenos Aires, Argentina, ricardo@dc.uba.ar

Hans Rott Department of Philosophy, University of Regensburg, 93040 Regensburg, Germany, hans.rott@psk.uni-r.de

Gerhard Schurz University of Duesseldorf, Geb. 23.21, Universitaetsstrasse 1, D-40225 Duesseldorf, Germany, schurz@phil-fak.uni-duesseldorf.de

Caroline Semmling Institute of Philosophy, Dresden University of Technology, Dresden, Germany, Caroline.Semmling@gmx.de

Paul Thagard Department of Philosophy, University of Waterloo, Waterloo, ON, Canada, N2L 3G1, pthagard@uwaterloo.ca

Heinrich Wansing Institute of Philosophy, Dresden University of Technology, Dresden, Germany, Heinrich.Wansing@tu-dresden.de

David Westlund Fjärdingsvägen 29, 241 36 Eslöv, Sweden, david@westlund.fm

Frank Zenker Department of Philosophy, University of Lund, Kungshuset, Lundagard, 22 222 Lund, Sweden, frank.zenker@fil.lu.se

Chapter 1
AGM Theory and Artificial Intelligence

Raúl Carnota and Ricardo Rodríguez

1.1 Introduction

Belief revision is a young field of research that has been recognized as a subject in its own right since the late 1970s. The new subject grew out of various research traditions. We will focus our attention on two of these traditions, which converged at the end of the 1980s.[1]

One of these emerged in computer science. Since the beginning of computing, programmers have developed procedures by which databases could be updated. The development of Artificial Intelligence (AI), inspired computer scientists to construct more sophisticated models for database updating. Jon Doyle is very well known for his TMS (Truth Maintenance System)[2] and was also a pioneer in understanding the importance of incorporating concepts of belief revision in AI.[3] Another significant theoretical contribution was a 1983 paper by Ronald Fagin, Jeffrey Ullman and Moshe Vardi.[4] We will analyse this tradition in more detail in Section 1.3.

The second of these two research traditions is philosophical. In a wide sense, belief change has been a subject of philosophical reflection since antiquity. In the twentieth century, philosophers have discussed the mechanisms by which scientific theories develop, and they have proposed criteria of rationality for revisions of probability assignments. Early in the 1970s, a more focused discussion took place

R. Carnota (✉)
Universidad Nacional de Tres de Febrero, Av. del Libertador Gral, San Martín 5569 Piso 14 'B',
1426, Buenos Aires, Argentina
e-mail: carnotaraul@gmail.com

In memory of Carlos E. Alchourrón

[1] Another important line of research was developped in theoric foundations of economic sciences. In particular, we can call it "epistemic foundations of equilibria in games". Cristina Bicchieri was the first researcher to introduce an application of AGM to game theory see Bicchieri, Cristina (1988a). She also attended the TARK'88 Conference. We thanks Horacio Arló Costa for this observation.

[2] Jon Doyle (1979).

[3] Jon Doyle and Philip London (1980).

[4] Ronald Fagin et al. (1983).

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_1,

regarding the requirements of rational belief change. Two milestones can be pointed out. The first was a series of studies conducted by Isaac Levi in the 1970s.[5] Levi posed many of the problems that have since become major concerns in this field of research. He also provided much of the basic formal framework. William Harper's work from the same period has also had a lasting influence.[6] The next milestone was the AGM model, so called after its three originators, Carlos Alchourrón, Peter Gärdenfors, and David Makinson. The genesis of this model is the objective of next section. In their seminal paper "On the Logic of Theory Change: Partial Meet Contractions and Revision Functions".[7] Alchourrón, Gärdenfors, and Makinson investigate the properties that such a functions should have in order to be intuitively appealing. The result was a set of postulates (named AGM postulates) that every belief change operator should satisfy. The "AGM Theory" had a major influence on most subsequent works on belief change, being the dominating paradigm in the field ever since. The publication of the AGM postulates was followed by a series of works by several authors, studying the postulates' effects or providing equivalent formulations. In particular, the constructive approach given there was adopted by an important part of AI practitioners as the (almost) universal model for the specification of the updates of Knowledge Bases as it usually involves a new belief that may be inconsistent with the old ones. From that moment, the references to the original paper within the field of AI have increased drastically.

Why did AGM receive such swift acceptance from the AI community?

In this chapter we will show that the AGM theory came out at a critical time for AI. On the one hand, the 1980s represented a high point in the history of IA, in which this field took advantage of the potential of the nascent information revolution and was turned into an magnet for practitioners of more diverse disciplines, particularly philosophers attracted by the field's promises. On the other hand, this same growth drove many investigators to search for solid foundations in logic and formal systems, tools that had been questioned in earlier years, as was pointed out by Newell.[8] A significant example of this was produced when IBM researchers in the San Jose Research Laboratory called an interdisciplinary conference on the theoretical aspects of reasoning on knowledge that "*could increase the knowledge of the workers of one field about the work developed in other fields*".[9] The first TARK Conference took place in March of 1986. This conference, and the following ones, can be seen as representing this dual aspect: the multidisciplinary convocation outlined the problems and the open search for solutions on the part of the AI community. J. Halpern opened the event with an Overview on Reasoning about Knowledge, in which the question of how to actualize the knowledge bases in the face of inputs that are possibly inconsistent with the previous corpus, was postulated

[5] Levi, I. (1977, 1980).

[6] Harper, W. (1977).

[7] C. Alchourrón et al. (1985).

[8] Alan Newell (1981).

[9] M. Vardi (1988).

as an unresolved issue, a "can of worms" in his own terms.[10] This explains in part why, at the second TARK, in 1988, when AGM debuted in the world of AI, it's impact was immediate, because it proposed a formal path to manipulate the update in the presence of inconsistencies, and, at the same time, it married very well with various attempts already under development, as is the case with the pioneering work we will be discussing in the Section 1.4. One shouldn't believe, as a consequence, that there was a previous lack of formal propositions, but the appearance of an abstract model, "in the knowledge level", in accordance with the Newell formulation, produced a strong intellectual attraction. The high level of abstraction of the new approach was what a lot of researchers in AI were seeking.

In our work we validate this general interpretive frame in two ways. On one hand, through the papers of outstanding researchers in the field. On the other hand, reconstructing, through articles and interviews, the particular paths through which AGM theory had come to be known and used by the first researchers of AI who included it in their work.

More than 20 years since its formulation, the AGM model continues to be amply cited as a reference in works in different areas and especially in computer science and AI. In Section 1.5 we will make a qualitative and quantitative evaluation of the impact of the AGM theory and an analysis of the lines of investigation which inspired said model.

1.2 The Origins of the Theory of Rational Belief Revision

On the late 1960s, David Makinson visited Argentina and was invited to lecture on deontic logic. After the lecture, he met Carlos Alchourrón. Makinson was surprised to find a conversational partner with very clear ideas and piercing arguments about the problems of deontic logics.[11] Since their first interactions, Alchourrón told Makinson about his interest in the concept of derogation of a regulation contained in a code.[12] In the philosophy of law, this notion had been regarded as unproblematic for a long time: one simply removes a regulation from a corpus to form a new and smaller one. But the result of a derogation can be indeterminate. When Alchourrón mentioned this fact to Makinson "*. . . as requiring some kind of formal analysis, I was at first quite unprepared to respond actively. . . I still remember my initial off-the-cuff response: the plurality is just an unfortunate fact of life, and logic cannot adjudicate between the different possibilities.*"[13] However, the subject began, slowly at first, to receive attention and the belief started to gain acceptance that, although logic could

[10] J. Halpern (1986).

[11] "Carlos made some very penetrating remarks that made me feel quite ashamed – there was I, talking about deontic logic to people who I had assumed knew nothing about it, and this guy in front of me evidently had a clearer picture of what is involved than I did". D. Makinson, Personal communication.

[12] This problem began to get the attention of Alchourrón and his colleague Eugenio Bulygin early in the 1970s.

[13] D. Makinson. Personal communication to the authors.

not really resolve between the different results of a contradiction, some general principles that any derogation should satisfy could be found, and, in consequence, there might exist interesting ways of formally generating the complete family of possible derogations. A first attempt was published in a joint article in 1981 (that we shall henceforth identify as AM81), titled "*Hierarchies of Regulations and their Logic*".[14] Though nowadays it can be considered, in D. Makinson's opinion, as a difficult progress towards the formulation of ideas that later would be known as "maxichoice" and "safe contraction",[15] it is interesting to review it here so as to get an insight of how the main ideas were generalized, starting from a purely juridical motivation and some basic intuitions. The scene laid down in "Hierarchies. . ." was that of a judge or official trying to apply a body of regulations or rules of law, on which a derogation of multiple results had been applied, with the goal to reach a verdict.[16] Given a code A, the formal treatment assumed a partial order between the regulations of A, which was considered a natural assumption in the legal field, and this order was used to induce an order in Pot(A) (the set generated by all the subsets of A). The maximal subsets of A that did not imply the regulation to be derogated were called "*remainders*" and conditions were demonstrated about the hierarchy $(A, \leq)$ with the goal of obtaining a unique "*remainder*" and, in this way, a unique result for the derogation operation. As important as the treatment of derogation was the examination of the case in which a code had contradictions in relation to certain empirical facts considered true.[17] The authors proposed a solution to this problem (which they called "delivery") using, again, a partial order structure between the regulations.[18] In this question, the authors acknowledged to have been inspired by the discussions of David Ross about potentially conflictive conditional obligations, which Ross interpreted as "prima facie" obligations. This treatment is an antecedent of the work on defeasible conditionals that Alchourrón developed later on.[19] The last part of the article suggested extra juridical applications for derogation and "delivery".[20]

[14] C. Alchourrón and D. Makinson (1981).

[15] "Groping painfully towards ideas that would later became formulated as maxichoice contraction and safe contraction". David Makinson (1996).

[16] The term "contraction" does not appear in the text, only "derogation" is used.

[17] Let A be the set of regulations; $B,\ C \subseteq A$, and let F1 and F2 be sets of facts. The stated situation is that of $B \cup \mathrm{F1} \Rightarrow x$; $C \cup \mathrm{F2} \Rightarrow \neg x$.

[18] In the general case of an inconsistent code A, the proposal was to make a derogation of A by $x \wedge \neg x$.

[19] In the last years of his life, Alchourrón published a series of articles on the logic of defeasible conditional. In these papers he proposed a philosophical elucidation of the notion of defeasiblity and applied it to clarify deontic concepts such as that of *prima facie duty*.

[20] One of these examples of possible applications of delivery was the case of a computational information system that, because of some "accident", had included inconsistent information during a process and had to be kept in operation, in a secure way, while the cause of the error was being repaired. In those years the "TMS" systems were being developed in AI, but the authors, far from suspecting their future involvements, only made reference to conventional systems, which is

During the process of publication of "*Hierarchies*...", Alchourrón and Makinson gained a better awareness that, both the problem and the approach were not limited in the least to the case of regulations. The initial set A could be an arbitrary set of sentences and the problem would then become one of deleting an element of a set or an undesirable consequence of the generated theory. The work in the field of philosophy related to contractions and revisions of beliefs were influential for the change towards a more general perspective between the 1981 paper and their following one, "*On the Logic of Theory Change: Contraction functions and their associated Revision functions*" (AM82).[21] This article reflected the change even in the language used: derogation was generalized as contraction. As Alchourrón himself remarked, in AM82 "*...the consequences of several ways of conceiving contraction are analyzed. In the first one, contraction is identified with the intersection of all maximal subtheories which do not entail the proposition to be deleted. We prove a paradoxical result that disqualifies it. The second way identifies contraction by means of a selection function that selects one of the maximal subtheories which do not entail the proposition to be eliminated. In spite of its intuitive hold, the construction has paradoxical consequences, except when the theory selected is itself maximal...*".[22] "Full meet contraction", "maxichoice contraction", and their problems made their first appearance in this work.[23]

While writing "*On the Logic*....", its authors became aware of the work of Gärdenfors[24] and his postulates of rationality for revision, and perceived that they were working on the same formal problems. Much so, that the intended journal for the publication of the paper (Theoria) was determined by the fact that Gärdenfors was, at the time, its editor. In the Introduction, they remarked that, in the logic of theory change, there were two main processes, contraction, that "*in the deontic context of deleting a regulation from* a *code is known as derogation*" (see footnote 21), and revision (or amendment in the deontic case), and that research on these processes had followed up to that moment two main avenues. "*One, explored by Gärdenfors in a series of publications... works, essentially, through the formulation of a certain number of postulates or conditions.... Another approach, underlying the previous work of the present authors, is the search of explicit definitions of functions of*

obvious because the system of the example was conceived as acquiring inconsistent information through an error that later a support team would have to repair.

[21] C. Alchourrón and D. Makinson (1982).

[22] Carlos Alchourrón. Presentation of his research work included in the application for the competition for full professor of Logic in the Faculty of Philosophy and Literature of the University of Buenos Aires, in 1985. The application was kindly given to the authors by Gladys Palau.

[23] In the case of full meet, the revision derived from it by Levi's identity produces as result only the consequences of the "new belief", losing all the previous background. In the case of maxichoice, if a non complete theory is contracted, the generated revision generated by Levi produces a complete theory; the agent becomes omniscient. This result is counterintuitive, except in the case where the set of starting beliefs is already complete. Both problems are resolved with "*partial meet contraction*".

[24] Among them Peter Gardenfors (1979, 1982).

contraction and revision… and later explore how far the functions thus defined… happen to possess the postulated properties".[25] Consistent with this approach, after constructing "full meet" and "maxichoice", the paper included a demonstration that "maxichoice contraction" satisfied the majority of Gärdenfors' postulates.

Gärdenfors was very impressed by the results obtained by Alchourrón and Makinson in "*Hierarchies…*" and quickly got in touch with them. He had begun to work in Belief Revision because he searched a pragmatic model of explanations.[26] At that moment, he thought that explanations were based on various forms of conditional sentences. Gärdenfors' early work was influenced by ideas of William Harper and Isaac Levi in epistemology and philosophy of science.[27] This led him to develop an epistemic semantics of conditionals. His main thesis was that conditional sentences, in its different forms, are about change of belief and, reciprocally, he considered that conditional sentences are a more important tool to describe how different kind of changes can be made. In addition, he proposed an alternative semantics for conditionals based on epistemic notions instead of one formulated in term of possible worlds and similarity between them (such as Levi and Stalnaker had done). The fundamental concepts of his semantics were states of belief and changes of belief. With the aid of these concepts he gave, not truth conditions, but criteria of acceptability, using a suggestion made by Ramsey, which can be summarized as follows: *accept the conditional $A > B$ in a state of belief K if and only if the minimal change of K necessary to accept A also requires accepting B.*[28] In order to make this suggestion precise, in his 1979 paper[29] he presented a formal characterization of changes of belief which was "almost" the AGM postulates, although it was still strongly connected to an analysis of conditionals. Basically the same postulates are presented in a more general setting in "*An epistemic approach to conditionals*".[30] But only after this, did he also turn to the more general problem of characterizing contractions and revision. Thus, the first formulation of revision independent of conditionals appeared in "*Rules for rational changes of belief*"[31] where the conditions 7 and 8 in the (future) AGM paper were formulated correctly, although they were not presented in the simplest way. He arrived at the postulates by "*stripping off*" probabilities from conditions for conditionalization and making them purely

[25] C. Alchourrón and D. Makinson (1982). The major part of the references in this paper are to Gärdenfors' papers, to whom they express their gratitude for having put them at their disposal, even those that were being typed.

[26] Peter Gärdenfors (1980).

[27] In fact, when, in 1977, Gärdenfors presents, in Helsinki, one of his first papers in this area, "Conditional and Change of Belief", he has an extensive and fruitful interchange of ideas with both researchers.

[28] This idea of a connection between Belief revision and the Ramsey Test for conditionals was abandoned later because it was shown to be impossible (Gärdenfors reported this result in Peter Gärdenfors (1986).

[29] Peter Gärdenfors (1979).

[30] Peter Gärdenfors (1981).

[31] Peter Gärdenfors (1982).

about membership in belief sets. They were initially intended to capture revisions of scientific theories, but then became more general rules for belief revision. One curious fact should be mentioned in relation to this last paper: although Gärdenfors knew the work AM81 and, at the same time of its writing, he also got acquainted with the preliminary versions of "*On the Logic...*" there is no constructivist notion in his paper. This absence happens to be all the more curious when, on the end of the paper, the very same author mentions, as a problem, that the rules (axioms) he presents are not sufficient to characterize a unique change function.

Starting with their interaction as a trio, Gärdenfors adopts the constructivist notion and the model for contraction and revision in terms of maximal subsets of the initial set of beliefs that do not entail the sentence to be deleted. After a rich correspondence between Buenos Aires, Beirut, Lundt and Paris (e-mail had not yet entered into everyday life), they arrived to the foundational paper of what later would be known as AGM (see footnote 7). Let us turn, once again, to Alchourrón to identify the advances brought by this paper: "*The mentioned difficulties... (with reference to the ones that appeared in the previous work)... pushed us to a generalization of our conceptual framework. The contraction of a theory is identified there with the intersection of a non empty selection of maximal subtheories that do not entail the proposition to be deleted. For this construction we demonstrate a representation theorem for the basic postulates of Gärdenfors' theory of rational revision of beliefs. We examine the consequences of introducing relations among the subtheories of a given theory and we prove, amongst other things, the representation theorem for the totality of the axioms of Gärdenfors' theory of rational belief revision. In this way, both approaches coincide, although they have independent intuitive justifications*" (see footnote 22).

Alchourrón and Makinson considered the operation of contraction to be basic (obtaining revision through Levi's identity), but Gärdenfors considered revision a primitive operation (defining contraction through Harper's identity). "*Finally, more or less by a vote of two to one, we ended up by taking contraction as basic and spending a lot of time on its properties.*"[32]

Later, two parallel lines of work were initiated. On one side, Alchourrón and Makinson defined the contraction known as "*safe*", for which the first had predilection, and proved several connections between it and "*partial meet*" on which the work of the trio was based.[33] Alchourrón said about this new operation that "*It happens to be more intuitive and, in a certain way, more realistic, to think that the elements preserved of a theory, when you try to delete one of its consequences, are selected comparing the elements of the theory, rather than its different subtheories. This approach is developed in "On the logic of theory change: safe contraction", where we put forward a definition of contraction of a theory from a relation among its elements. We prove that, under very intuitive conditions of the properties of*

[32] D. Makinson. Personal communication to the authors.

[33] Both subjects are treated in C. Alchourrón and D. Makinson (1985) and C. Alchourrón and D. Makinson (1986) respectively.

the ordering relation, contraction satisfies the conditions of Gärdenfors' axiomatic approach. In the same work and based on the same contraction, we approach the subject of the iteration of rational revisions. Besides, not all the contractions defined by the model of a relation between subtheories is a contraction defined on the base of a relation between the elements, although the reciprocal is always true. However, if the theory has a finite number of non equivalent propositions, both approaches are in correspondence. This is shown in "Maps between some different kinds of contraction function: the finite case" (see footnote 22). Curiously, of the five characterizations of AGM, "*safe contraction*" is the one that has had fewer repercussions in the area of AI, although one of its motivations was to produce an executable model. A possible explanation might reside in the fact that it was published in *Studia Logica*, a journal of very little impact on the AI community[34] and, also, at a time in which AGM theory was still not known to this community.

The other line of work was developed by Gärdenfors and Makinson, and consisted in defining revision functions in terms of an epistemic order of the sentences of a Knowledge Base, which they called "*Epistemic Entrenchment*". This line of work is, as we shall see in the next sections, the one responsible for popularizing AGM theory in the field of AI, starting with its presentation in the conference TARK'88. It should be noted that this approach has had very important derivations in the area of Non Monotonic Reasoning (NMR).

1.3 The State of Artificial Intelligence in the 1980s

Since its birth, AI created great expectations about its revolutionary results. However, after 20 years of patient waiting and great investments, few were the results obtained and very far from fulfilling the initial fantasies. This mismatch between achievements and promises led to several restatements and internal debates within the discipline.

Towards the end of the 1970s and beginning of the 1980s, many renowned researchers in IA, particularly from the logicist or formalist party, disputed the firmness of the bases of the systems they were developing. These doubts reached the Database area, which had recently incorporated a logicomathematical foundation (relational algebra and calculus), and in which a fraction of its workers were also part of AI projects, by way of the extended concept of a Knowledge Base (KB). Questions, such as what exactly did the updating of a data base mean or what could be said that it "knows" an (artificial) "intelligent agent" provided with data from its environment, a program for its manipulation and which also incorporates certain criteria of action, began to be asked. Several lines of research tried to find an

[34] "This approach has never had much echo among computer scientists – perhaps because of the place where it was published – but I have always had a particular affection for it. I think that Carlos quite liked it too, although I would describe myself as the father and he as an uncle". D. Makinson. Personal communication to the authors.

answer to these questions, and their results were published during the 1980s of the past century.

As we shall see, the modeling reached in those same years by Alchourrón, Gärdenfors and Makinson – in an independent fashion and as a response to different motivations –, was in tune with the agenda and expectations of a relevant sector of the AI community. It is this tuning that allows us to explain the wide and fast impact that the Logic of Theory Change had on this community towards the end of that decade.

1.3.1 The Knowledge Level

Alan Newell, one of the founders of AI,[35] in his "*Presidential Address*" at AAAI '80,[36] titled "*The Knowledge Level*" (see footnote 8) set himself to approach a subject, knowledge and its representation, which several indicators led him to conclude that required great research efforts. This unsatisfactory situation was revealed, in his opinion, by three indicators. The first one was the permanent attribution of a cuasimagical role to knowledge representation. It was a cliché of AI, Newell asserted, to consider representation as the real problem to confront, the *locus* of the true intelligence of a system.

A second indicator was the negative residue of the great controversy about theorem proving and the role of logic that occurred between the end of the 1960s and the beginning of the 1970s, in the past century. The first works on theorem proving for quantified logics culminated in 1965 with the development, by Alan Robinson, of a formulation of first order logic geared to its mechanical processing, named resolution. It followed an intense period of exploration in proof systems based on resolution. The basic idea was to have at hand a general purpose reasoning engine, and that to make logic, and to do it right, was the stepping stone of intelligent action.[37] A few years after, people started to perceive that this was not so: the engine was not powerful enough in practice, not even to prove theorems that were difficult for the human brain or to solve tasks like planning in robots. A reaction surged with the slogan "uniform procedures do not work", from which it arose, as a positive result, a whole new generation of programming languages in AI. The negative residue of this reaction was "bad press" for logic as a tool in AI: logic was static, did not admit control mechanisms for inference, the failure of theorem proving using resolution implied the failure of logic in general, etc.

The third indicator was the set of results from a survey made between 1979/80, by Brachman and Smith, among researchers in AI belonging to different areas and projects, and with different approaches or attitudes towards the critical questions

[35] There is consensus in locating the "birth" of AI in the Dartmouth Conference of 1956.

[36] American Association for Artificial Intelligence Conference (Stanford, 8/19/80). Newell was, at that moment, the President of AAAI.

[37] This point of view was framed inside the leibnitzian tradition.

in the area of knowledge representation.[38] The main result of the analysis of the responses was a real jungle of opinions, without consensus on any substantial question. In the words of one of the people surveyed, quoted by Newell, "*The standard practice of knowledge representation is the scandal of AI*" (see footnote 8).

The declared goal of Newell's address was to shed some light on the area, having in mind that research in Knowledge Representation and Reasoning (KRR) should be a priority in the agenda of the discipline.

The focus of his initial search was the question of what was knowledge, how was it related to representation and what is it that a system has when it is said of it that it is based on knowledge.[39]

Here, Newell enunciated the Knowledge Level Hipothesis as follows: "*. . .there exists a distinct computer system level, lying immediately above the symbol level, which is characterized by knowledge as the medium and the principle of rationality as the law of behavior.*"[40]

In the knowledge level there is an agent (the system) that processes its knowledge to determine which actions to perform, within a repertory of possible actions, with the purpose of fulfilling its goals. Its behavioral law is defined by the rationality principle, by which the actions selected have to be the ones that better approach the agent to the fulfillment of its goals, given its present knowledge. Because of the definitional autonomy of every level, the knowledge contents of the agent and its goals are completely specified in the KL, independently of the form in which that knowledge and goals are represented in the system. Newell said that representations exist in the symbolic level (SL), which is the level immediately below KL, and consists of the data structures and processes that embody or realize the knowledge body of the agent, specified in the KL. In the SL, any form of representation (logic, images, plans, models, scenes, texts) may be adequate, so far as efficient processes

[38] Special Issue on Knowledge Representation. SIGART Newsletter. February 1980 (70).

[39] In the work of that time, "knowledge" and "belief" are two terms used indistinctively, even in those cases in which the authors acknowledge their philosophical differences. In the context of AI, "knowledge" would be all that it is assumed to be represented in the data structures of the system.

[40] This "level" refers to a previous notion of level or tier in a computational system. The lowest is the device level, the next one the circuit level and so on up to the level of programs or symbolic systems (SL). Each level is not an "abstraction" of the lower ones, neither a simple "point of view". It has a real existence, even independent of the multiple possible ways in which it may be realized or supported by the lower level. Each level or tier is a specialization of the class of systems capable of being described in the next level. Thus, it is a priori an open question which is the physical realization of a level in the lower structure in the agent at the knowledge level; the determination of behavior by a global principle and the failure to determine behavior uniquely, running counter to the common feature at all levels that a system is a determinate machine. . .Yet, radical incompleteness characterizes the knowledge level. As Newell said ". . .Sometimes behavior can be predicted by the knowledge level description; often it cannot. The incompleteness is not just a failure in certain special situations or in some small departures. The term *radical* is used to indicate that entire ranges of behavior may not be describable at the knowledge level, but only in terms systems at a lower level (namely, the symbolic level). However, the necessity of accepting this incompleteness is an essential aspect of this level. . . .". Alan Newell (1981).

that extract knowledge from them exist. However, if an observer wants to predict the behavior of the agent, he would not need to know these physical realizations or representations.[41] The knowledge level, Newell remarks, "*. . . allows to predict and understand the behavior of the agent without having an operational model of the process really made by him. . .*" (see footnote 8).

Some of the main conclusions reached by Newell are the following:

The concept of representation exists in the SL and the abstract characterization of the knowledge that an agent should possess is in the KL. Knowledge serves as the specification of what a symbol structure should be able to do.

Within this framework, logic can be seen in two different ways. One is to consider it as a candidate, among others, for the representation of the knowledge in the SL, both positive elements (its representation structures and extraction processes are well known) and problems (in the first place, the one of inefficiency, referring in this case to the cost for the symbolic process of extracting knowledge from that particular structure). The other way is to consider logic as the appropriate tool for the analysis, in the knowledge level, of what an agent knows. In fact, given a representation in the SL – semantic net, chessboard, graph or any other – if the purpose is to extract what knowledge exists in that representation and to characterize it, the use of logic is a requirement.

In consequence, the restrictions in the use of logic as a means for representation do not affect the role of logic as a fundamental analysis tool of the KL.

When distinguishing strongly KL from SL, we are drawing a similar strong separation between the necessary knowledge to solve a problem and the processing required to extract and exploit that knowledge in real space and time.

From the perspective of KL, whichever the structure S that supports the knowledge K of the agent, an external observer will attribute to it all that the said observer can know from K.[42]

The reflections of Newell[43] gave impulse to the search, by an important group of AI researchers, of appropriate formalisms in KRR that would allow to understand

[41] If a system has a data structure from which it can be said that it represents something (object, procedure, or whatever) and it can use it, by means of certain components that interpret the structure, then it is said about the same system that it has knowledge, the knowledge is embodied in that representation of the thing. When we say that "the program knows K", what we want to say is that there is a certain structure in the program which we "see" as supporting K, and that, also, it selects actions exactly as we would expect that an agent that "knows K" would do, following the rationality principle, that is, the most adequate to reach its goals.

[42] A priori, it could be any consequence from K.

[43] To reinforce the idea that the problem was "on the table" at that time, we note that some similar (but not identical) considerations had been proposed previously. The main example is McCarthy, in his work McCarthy (1977) where he proposed to divide any problem in AI in two parts: the epistemological ("what information is available to an observer and what conclusions can be drawn from information" and "what rules permit legitimate conclusions to be drawn. . . ") and the heuristic ("how to search spaces of possibilities and how to match patterns"), although he left in the epistemological part the question about how the information was represented in the memory of a computer. In fact, this proposal dates back to J. McCarthy and P. Hayes (1969).

which operations were made by the programs constructed to simulate "intelligent behavior".[44]

At the same time, the 1980s experienced a revival of AI and a renewed participation in its controversial debates of philosophers, psychologists and linguists of diverging orientations who found here an appropriate environment to elucidate their own objects of study.[45]

1.3.2 A Presentation of Belief Revision in AI

The same year of the "Presidential Address" by Newell, a paper by Jon Doyle and Philip London "*. . .presents an overview of research in an area loosely called belief revision*" (see footnote 3). The authors remarked that one of their goals was to introduce the literature of BR to AI researchers, a literature that exceeds the area of AI itself and extends to logic, philosophy, epistemology and psychology. This work had a brief introductory text about BR and its relevance for the study of AI problems, and its main contribution consists in a thematic classification of 236 works with complete references, from several areas but linked to BR. The authors justified the compilation effort in these terms: "*. . . Intelligence is often viewed as the ability to reason about and adapt to a changing environment. For this reason most computer programs constructed by artificial intelligence researchers maintain a model of their external environment. The model is updated to reflect changes in the environment resulting from the program's actions or indicated by its perception of external changes. AI programs frequently explore assumptions or hypotheses about these environments; this may lead to further model updating if new information conflicts with old, indicating that some of the currently held assumptions or hypotheses should be abandoned. . .*" (see footnote 3). But the philosophical literature in BR had very little to do with "mundane changes" and, in consequence, "*. . .It remained to Artificial Intelligence researchers to uncover a major problem virtually undiscussed in earlier work. This is the so-called frame problem of McCarthy and Hayes, the problem of how to update models to account for changes induced by actions. The basis of the problem is that even if one can succinctly specify the ways in which a system's environment might change in terms of the effects of actions, it still remains*

[44] Brachman, Levesque, Moore, Halpern, Moses, Vardi, Fagin and others. For example, H. Levesque in "Logic and the complexity of Reasoning", confronting objections of the type "a realistic cognitive activity is much more complex that any type of neat mathematical a priori analysis", notes that a model is interesting only if it serves to explain the behavior one wishes to model, and also that if that behavior is disorderly or mixed up, the model does not have to be so. The model can be more or less idealized or realistic, but this does not alter "the hard fact that a model that has a mistaken behavior or the correct behavior for mysterious reasons does not have any explanatory power". H.J. Levesque (1988).

[45] Some of these debates were the ones that divided symbolists from the researchers that tried to simulate neural mechanisms. See, for example, Graubard Stephen (1988)

to specify some way of judging what stays unchanged in the face of these actions..." (see footnote 3).

Doyle had recently developed his Truth Maintenance System (TMS) (see footnote 2), a knowledge representation method for representing both beliefs and their dependencies (the name *truth maintenance* comes from the ability of these systems to restore consistency). On Newell's conception, the TMS operates in the Symbol Level and, in this sense, the authors of this bibliographic revision, classified the solutions proposed to the problems in BR in two categories, "implementational" and "theoretical". Among the first, they included from "manual updates" to the procedures of the "data-dependency" type (among which TMS was included),[46] whereas in the second they placed, among others, the formal studies of belief systems and non monotonic logics. We can assume that Doyle himself intended, with this review of BR, to promote new research that could contribute foundations as well as operational views towards the resolution of the notorious weaknesses of AI systems.

1.3.3 The Problem of Database Updating

Although towards the end of the 1970s Doyle and London, among others, had already pointed out the problems that, like the "frame problem", required a belief revision in AI systems, the field of Databases was the one where most of the advances occurred in that direction, under the very concrete pressure derived from the development of big information systems in industry.

In 1983, Fagin, Ullman and Vardi published a paper (that we shall henceforth identify as FUV'83) that opened with the warning: "*The ability of the database user to modify the content of the database, the so-called update operation, is fundamental to all database management systems. Since many users do not deal with the entire conceptual database but only with a view of it, the problem of view updating i.e., translating updates on a user view into updates of the actual database, is of paramount importance, and has been addressed by several works ... An assumption that underlies all of these works is that only the view update issue is problematic, because of the ambiguity in translating view updates into database updates, and that the issue of updating the database directly is quite clear* (see footnote 4)."

[46] "Data-dependencies are explicit records of inferences or computations. These records are examined to *determine* the set of valid derivations, and hence the current set of beliefs (that is, those statements with valid arguments). In some cases, they are erased along with the beliefs they support when changes lead to removing a belief and its consequences from the database... In other systems, the dependencies are kept permanently. In this latter case, dependency-based revision techniques can use a uniform procedure, sometimes called truth maintenance (Doyle 79a), to mark each database statement as believed or not believed, depending on whether the recorded derivations currently provide a valid argument for the statement. One might view this sort of dependency analysis as analogous to the *mark~sweep garbage* collection procedures of list-processing systems..." Jon Doyle and Philip London (1980) page 9.

The authors put in question that naïve point of view with several examples. One of them was a database that originally contained the set of propositions {A, B, C} with the integrity constraint A&B → C, and from which C was to be deleted.[47] The examples led to multiple possible results and, for the selection of only one, to the question of minimal change. For the authors, these difficulties had the same underlying problem "*The common denominator to both examples is that the database is not viewed merely as a collection of atomic facts, but rather as a collection of facts from which other facts can be derived. It is the interaction between the updated facts and the derived facts that is the source of the problem* (see footnote 4)."

The FUV83's model was developed for Belief Bases at the syntactic level but, since it employs a model semantics, all the consequences are included and, therefore, at the semantic level, it generates a behavior for Belief Sets. They treat a database as a (not necessary finite) consistent set of statements in first-order logic, which is a description of the world, but not necessarily a complete description (i.e. a maximal consistent set) of it. They point out that when one tries to update a database by inserting, deleting or replacing some first-order statement, several new databases can *accomplish* the update. Moreover, they state that only databases that change the existing database as little as possible should be considered. That is, some partial order that reflects the divergence of the new databases from the old one should be defined and, then, only the databases minimal with respect to this order should be considered.

In order to formalize this notion of *smallest change* when going from a database T to a database S by inserting, deleting or replacing, they consider the set T-S of facts that are deleted from the original database and the set S-T of facts that are added to the database. They want to obtain the smallest change by minimizing both the set of inserted facts and deleted facts. Thus, they say that T1 accomplishes an update u of T with a smaller change than T2 if either T1 has fewer deletions than T2 or T1 has the same deletions as T2 but T1 has fewer insertions than T2. They say that S accomplishes an update u of T minimally, if there is no database S^* such that S^* accomplishes an update u of T with a smaller change than S. By using this definition, they are in condition to formulate the following representation theorem:

1) S accomplishes the deletion of u from T minimally if and only if S is a maximal subset of T that is consistent with $\neg u$.
2) $S \cup \{u\}$ accomplishes the insertion of $\neg u$ into T minimally if and only if S is a maximal subset of T that is consistent with $\neg u$.

They note that there is an interesting duality between deletion and insertion:

S accomplishes the deletion of u from T minimally if and if and only if $S \cup \{\neg u\}$ accomplishes the insertion of $\neg u$ into T minimally

[47] As we remarked in the previous section, this is the same type of problem that Alchourrón and Bulygin had considered a few years before, with respect to the derogation in a legal corpus.

They address also the problem of what should be done in the case that several databases accomplish the update minimally. Their first proposal is to define the update as the intersection of all these databases. In FUV'83 they prove that the insertion defined in this way is not satisfactory because it throws away all the old knowledge each time an inconsistent insertion is attempted. Then, they suggest that the problem may be circumvented if a notion of "*database priorities*" is adopted. They note that not all the elements in a database are equally viable for deletion. To handle this distinction, they introduce the concept of tagged sentences $< i,\varphi >$, where i is a natural number and φ is a sentence. The intention is that the lower the tag of a sentence, the higher its priority. Now, a database is a finite set of tagged sentences. When comparing two databases to see which one of them accomplishes an update with smaller change, this comparison is based on the priorities given to the sentences. Intuitively, each database S that accomplishes the deletion of u from T minimally is now constructed by selecting a maximal subset consistent with $\neg$u from the lowest degree (starting at zero), then a maximal subset of the following degree is added such that it is consistent with $\neg$u, and so on. In the finite case, this process eventually finishes. The result of this process is a selected subset of all the possible databases that accomplish the update of u from T minimally (in the sense of the first definition). The proposed result for this second approach was to define as update the intersection of all these selected databases. This construction is the stepping stone for the models known as prioritized belief base.[48]

From a historical point of view it is relevant to notice the tight relation between these models and the ones proposed, independently and almost contemporaneously, by Alchourrón and Makinson in their first joint works, the ones we denoted in the previous section as AM81 and AM82. In the first place, the set T that accomplishes the deletion of u from T minimally coincides with the notion of remainder set, presented in AM81, when T is a theory (a logically closed set of sentences). Therefore, the first solution proposed in FUV83 for the problem of multiple databases that accomplish an insertion minimally, i.e. the intersection of all of them, has a direct correspondence to the construction that, in AM82, was called meet contraction (later full meet). Alchourrón and Makinson (motivated by the problem of derogation) as well as Fagin et al. noticed that this construction was not satisfactory.

What turns even more interesting the historical analysis of the present chapter is the second solution, which involves a hierarchization of the sentences of a database, which, in the words of Gärdenfors and Rott, "*is somewhat similar to the idea of epistemic entrenchment*", although "*the priorities need not respect the logical relationship between the sentences in the database*".[49] In the facts, FUV83 anticipates, in this second variant, the construction named "partial meet contraction", which would appear for the first time in AGM85. It is also significant the demonstration in FUV83 of the duality between deletion and insertion.

[48] Bernhard Nebel (1992).

[49] Peter Gärdenfors and Hans Rott (1992).

Finally, Fagin et al., who had formulated the problem from the point of view of databases, recognized that it was an instance of a more general problem, very critical to AI systems, the problem of belief revision, exactly as Alchourrón and Makinson had considered it, in AM82, generalizing their analysis about legal corpora.[50]

After this pioneer work, and its follow up in 1986,[51] other developments appeared which also studied the problem of database updating in front of possible inconsistencies, among which we can mention the ones by Borgida,[52] Weber[53] and Winslett.[54] Nonetheless, and in spite of the efforts made, the feeling that this very researchers transmitted was that of a problem devoid of convincing proposals. In 1986, three years after the first work by Fagin et al., Winslett remarked that: "*... What is less well recognized is that the first stage of belief revision, incorporating new extensional belief into the pre-existing set of extensional beliefs, is itself quite difficult if either the new or old beliefs involve incomplete information...*".[55] As we shall see in the next section, all these proposals were afterwards, when AGM theory began to be considered by many AI researchers a reference model, related to it.

1.3.4 The Presentation of AGM in AI

In 1984, what was at first supposed to be a small meeting of interdisciplinary research about the theoretical aspects of reasoning about knowledge organized in IBM San Jose Research Laboratory, was overflown with meetings of an average of forty attendants and an e-mail list of 250 names. Moshe Vardi, one of the organizers, remarked that the attendants "*included computer scientists, mathematicians, philosophers and linguists*" and that "*given the evident interest in the area by groups so diverse, it seemed appropriate a conference, particularly one that could increase the knowledge of the workers of one field about the work developed in other fields* (see footnote 9)." The First Conference on Theoretical Aspects of Reasoning about Knowledge (TARK) took place in march 1986, with a restricted attendance and the intention of stimulating the continuous interaction between the participants. Vardi remarked that "*the general feeling at the end of the meeting was that the interdisciplinary format of the conference had proven to be very successful*" (see footnote 9).

[50] "While the application that we have in mind here is updating databases, we believe that the framework developed here is also relevant to any kind of knowledge base management system. *From the point of view of Artificial Intelligence, what we have here is a logic for belief revision, that is, a logic for revising a system of beliefs to reflect perceived changes in the environment or acquisition of new information*. The reader who is interested in that aspect is referred to Doyle & London". Ronald Fagin et al. (1983).

[51] Fagin, R. et al. (1986).

[52] Borgida A. (1985).

[53] Weber, A. (1986).

[54] Winslett M. (1988).

[55] Winslett M. (1986).

The introductory article by Halpern in the Proceedings of this first TARK was a review of the area of Reasoning about Knowledge (see footnote 10).

His purpose was "*to make a review of the central questions in the research of reasoning about knowledge, common to philosophy, economy, linguistics as well as artificial intelligence and computing theory*" (see footnote 10). He presented here what he denominated the "classical model" of knowledge and belief, towards which researchers from AI had been approximating since the start of the discipline: the model of possible worlds (the agent "knows" what seems to be true in every world that he "thinks is possible").[56] Usually, they were proposals supported by modal logics (A3, A4 and A5, depending on the case) and that was the reason for the many references in the works of the time to Hintikka and Kripke, and to a lesser degree to D. Lewis, Anderson and Belnap.

Halpern noted some limitations of these models, such as the calculation difficulties and logic omniscience. In particular, he focused on what he considered "*the most interesting application that motivates the study of knowledge in AI: how to understand which is the necessary knowledge for an action and how that knowledge can be acquired through processes of communication*" (see footnote 10). And, later, he remarked that "*The greater part of the works… assume, implicitly or explicitly, that the received messages (by the agent) are consistent. The situation gets much more complicated if the messages can be inconsistent. This takes us very quickly to the whole complex set of questions involved in belief revision and reasoning in the presence of inconsistency. I wont́ attempt to open this can of worms here, these are subjects that must be considered at some time or another when you design a knowledge base, for example, because the possibility for the user for acquiring inconsistent information is always present*".[57]

The success of the first TARK led to the call for a second, TARK′ 88, for which 108 papers were submitted and 22 selected, on the basis of their original contributions as well as for their interest for an interdisciplinary audience. One of the papers selected was "*Revision of Knowledge Systems using Epistemic Entrenchment*", by Gärdenfors and Makinson.[58] This was the "official presentation" of AGM and its authors to the AI community.[59] In fact, for Gärdenfors it was "*… the first computer science conference where I had a paper*".[60] The impact of this presentation was immediate and wide. In words of Gärdenfors, "*… this was a big surprise to myself.*

[56] For example, in the case of a distributed system, you could assign knowledge to it externally in this way: a processor X "knows" A if in all the global state in which X may be in, A is true.

[57] J. Halpern (1986) The underlining is ours. Curiously, in this overview, Halpern makes no reference to the work by Ronald Fagin et al. (1983) in spite of the fact that they are coworkers in San Jose and co-organizers of TARK.

[58] P. Gärdenfors and D. Makinson (1988).

[59] Of the remaining papers presented at the conference, only one made reference to AGM theory. Its author was Cristina Bicchieri, a researcher on economic subjects and game theory, who, in a personal communication to the present authors, remarked that it was I. Levi who had made her aware of the relevance of AGM for her work. The paper was Bicchieri, Cristina (1988b)

[60] Peter Gärdenfors. Personal Communication to the authors.

I was aware that database update was an important problem within computer science, since I had seen the work of Fagin's group and Foo's was enthusiastic, but I could never imagine that it would have such an impact in the AI society . . .".[61]

After TARK'88, an increasing number of researchers, many of them with great influence in the area of AI, started to get actively involved in the Logic of Rational Change of Theories.[62]

We have considered in this section four elements that, taken *in toto*, help us to explain the fast diffusion of the AGM model in an important sector of the AI community: the concern for the formal foundation of the systems developed in the area, and, in particular, in Knowledge Representation, with a reevaluation of the role of logic in the "Knowledge Level"; the introduction, motivated by that concern, of the literature of Belief Revision; the discovery of the problems involved in the updating of databases, and the attempts to solve them, and finally the interdisciplinary matrix that many activities in AI had at the time, which eased the acceptance of new models from other areas.

All these events allow us to conjecture that the AGM model attached itself and was functional to a previous process of formulation and search of solutions for the problem of Belief Revision (considered key to solve questions such as the frame problem or the updating of databases) by an influential sector of the AI community.

1.4 First Repercussions of AGM in Artificial Intelligence

The first reference to AGM in authors of the field of AI was published in 1986. The paper[63] was published in a journal of great influence and its purpose was to reveal a new approach to the problem of "*Knowledge Representation and Reasoning*", as formulated by Newell. Its author was H. Levesque, an influential researcher in the field, who, within a review of the main open questions in AI research, dedicates a brief paragraph to "truth maintenance".[64] In a generic mention to the works about belief revision in philosophy, there is a reference, without any comment, to the foundational work of AGM in 1985. Although the *paper* by Levesque had much influence, the fact that AGM was a mere bibliographic reference appears to be, on a historical perspective, an isolated fact, without any influence on its later impact. Several of

[61] Peter Gärdenfors. Personal Communication to the authors. Gärdenfors started, a short time before, to get acquainted with the problems of revision in databases and AI, starting with the works by Ronald Fagin et al., 1983 and by the interchange – on occasion of a sojourn in Canberra at the end of 1986 – with the group of Norman Foo and Anand Rao in Sydney.

[62] In a revised version, published in 1995, of the overview which Halpern used to inaugurate the first TARK in 1986 J. Halpern (1986), he does not speak anymore of BR as a "can of worms". When he refers to BR, he mentions AGM theory. This author, in the years that followed TARK'88, worked in BR and proposed an alternative model to AGM.

[63] H.J. Levesque (1986).

[64] Which was identified in AI with systems of the type of the already mentioned "Truth Maintenance Systems" (TMS) by Jon Doyle, and a family of derived systems.

the researchers surveyed for the present work agree to select TARK'88 as the entry point of AGM theory in AI, with the presentation of the work by Gärdenfors and Makinson, and the presence in the Conference of renowned figures of AI, such as Ray Reiter, Jon Doyle, J. Halpern, H. Levesque, R. Moore, M. Vardi and R. Fagin, among others. In fact, there is an inflexion point in the year 1988, since, in that year and the following, the pioneer works that made use or referenced in an active form AGM theory were published. However, this event by itself does not explain the adoption of the formalization model AGM by numerous researchers in AI. We have to add that, as a part of the search we described in the previous section and of the interest provoked by the topics studied by AI, an important participation of researchers in philosophy and logic occurred in the area and the opening of the journals of those fields to researchers of the latter and vice versa. A consequence of this interaction was the very existence of the TARKs. Besides, as it follows from the works already mentioned in the previous section, the question of database updating with inputs inconsistent with its previous contents by means of truth maintenance (or of consistency to be more precise) was at the order of the day, not only in AI, but also in the field of Databases. The difficulty resided in combining a precise and clear semantics (in the "knowledge level", paraphrasing Newell) with a "realistic" scheme from a computational point of view (in the "symbolic level").

The first author to consider with a certain level of detail the AGM formalism was Mukesh Dalal, in a paper presented in AAAI'88.[65] Dalal said: "*At the core of the very many AI applications built in the past decade, is a knowledge base – a system that maintains knowledge about the domain of interest. Knowledge bases need to be revised when new information is obtained. In many instances, this revision contradicts previous knowledge, so some previous beliefs must be abandoned in order to maintain consistency*..." (see footnote 65). Dalal started from the operations that Levesque had defined for the updating of a KB in the KL.[66] In particular, Dalal noted that the function

$$\text{Tell: KB x } L \rightarrow \text{ KB,}$$

where L is a formal language and Tell a function that adds knowledge to the KB, was defined only if the new information was consistent with the KB.

With the aim to surmount this restriction, Dalal defined

$$\text{Revise: KB x L} \rightarrow \text{KB}$$

which was meant to manage any kind of addition to the KB.

To guarantee that the characterization of Revise was in the KL, Dalal defined revision purely in terms of the models in the KB. He also proposed "*...an equivalent symbol level description by presenting a syntactic method for revising knowledge*

[65] Mukesh Dalal (1988).

[66] H.J. Levesque (1984).

bases..." (see footnote 65). Towards the end of the Introduction, he added: "*We show the relation of our work to research in Philosophy on the formal aspects of the logic of belief change (Makinson 1985) which has recently attracted attention in the AI community. For any revision scheme it is desirable that it preserves as much as possible the beliefs held prior to revision...*" (see footnote 65). Dalal, who developed part of this work without previously knowing the AGM model, formulated a series of non formalized principles which he considered intuitive guidelines for constructing a revision. These were: representation adequacy (the revised knowledge has to have the same representation as the old one, so as to be able to iterate); irrelevance of the syntax (the result of the revision does not depend of the syntactical representation of the old or new knowledge), a fundamental principle in the computational case, that was a weak point of several updating methods previously proposed in the field of AI; maintenance of consistency; primacy of the new information; persistence of previous knowledge (which can be assimilated to minimal loss) and *fairness* (non arbitrariness of the selection among the multiple results of a revision). The abstract representation of the contents of a KB was a finite set of formulas in a propositional language. Dalal generated a quantitative measure of the changes over a set of interpretations, and applied it to define minimal change over the models of a KB. He then constructed an algorithm for the calculation of the formula of the language corresponding to each set of interpretations, so as to be able to define syntactically, starting from the representative formula of the original KB, the formula that would arise as a result of the revision. In this way, the method does not require an explicit construction of models. Dalal recognized that every step of the proposed algorithm required a verification of consistency, which made it, in the general case, NP-complete, and for that reason he considered his research to be preliminary. When he compared his method against other proposals, the first paragraph was dedicated to AGM. Because his approach was constructive, he tried to compare his revision to the definition of the AGM postulates. With this goal in mind, Dalal considered a set of beliefs formed by the logical closure of the formulas that composed the KB ("*as it is suggested by the approach of KL*", he remarks[67]). Under this assumption, Dalal's revision satisfied all the AGM postulates. In the same work, Dalal also discussed the proposals presented in AI for the updating of logical databases by Fagin et al., Winslett, Weber and Borgida. As a consequence of his analysis, Dalal suggested that, in general, these proposals did respect neither the irrelevancy of syntax nor the principle of minimal change.

The work by Dalal, which was motivated directly by the necessities of AI and tried to fit into the formal reformulation of the discipline, intended to forge abstract models that could be analyzed in the knowledge level and it is the first one to establish a clear link to the AGM theory.

[67] Mukesh Dalal (1988). The sense of this expression by Dalal is already suggested in section 3, in the last of the conclusions that we attributed to Newell's work. For an external observer, the beliefs of an agent are, in principle, whichever consequences of his KB. Although it may not be realistic to consider that he "knows" them all, it should be expected that he may be able to arrive at any of them through a "goal" driven search and following the rationality principle.

How was this theory incorporated into his paper?

Dalal remembers that it was an anonymous referee who noted the relevance of analyzing AGM, and Alex Borgida, his counselor, stimulated him to explore the connection. "*One of the anonymous reviewers (assigned by AAAI program committee) suggested the connection with AGM's work. Prof. Alex Borgida, who was initially a co-author of my paper, advised me to probe this connection deeper. We both liked the formal foundations (especially of the principles) provided by AGM and wanted to bring that into the AI work on belief revision*".[68]

On his side, Alex Borgida states to have gotten acquainted with AGM through his friend, David Israel, a philosopher working in the AI world.

> "*I seem to recall that the connection to AGM was pointed out to me by a friend at SRI* (Stanford Research Institute)*, Dr. David Israel -- a famous philosopher turned AI researcher. He may have been one of the AAAI reviewers, too. Mukesh's paper was actually about a model theoretic description of propositional belief updates (essentially, minimal model mutilations), which followed my ideas for minimal mutilations for exceptions to integrity constraints. In this context, AGM made perfect sense, but we did not know the philosophy literature, and Israel did... As usual, it was a matter of being in the right place with the right knowledge*".[69]

As for D. Israel, he seems to have gotten acquainted with AGM through the philosophical literature, and he related it immediately to the research on Non monotonic Logics in AI.[70]

Also in 1988, Ken Satoh published the paper "*Non Monotonic Reasoning by Minimal Belief Revision*".[71] Satoh tried to differentiate his proposal from the known formalisms in Non Monotonic Reasoning (NMR), such as Default Logic by Reiter[72] and Circumscription, by McCarthy.[73] For this, he stated that "*...we define a special belief revision strategy called minimal belief revision ...*" (see footnote 71) and aspired to show that the strategy obtained some classes of non monotonic reasoning. Satoh distinguished knowledge from belief. Knowledge was the subset of beliefs that were valid (if the agent knows p, then he believes in p) and "not revisable" (they increased monotonically) and the remaining beliefs were contingent

[68] Mukesh Dalal. Personal Communication to the authors.

[69] Alex Borgida. Personal Communication to the authors.

[70] "I certainly did, quite early on – though never in print – notice a connection between the AGM work on belief revision and the work in AI on nonmonotonic reasoning and simply assumed, a little glibly, that the latter could be subsumed within the former – roughly the case dealt with via entrenchment, where some parts of a theory are protected against revision. In terms of abstract consequence relations, that wasn't a bad guess; but notice that it simply leaves unaddressed the issues of finding the relevant non-monotonic (default) fixed points and that, in the context of various models of extended logic programming, is where much of the interesting action of late has been. So, I'd give myself a B-/C+ for prophecy on this one.". David Israel. Personal Communication to the authors.

[71] Ken Satoh (1988).

[72] Ray Reiter (1980).

[73] J. McCarthy (1980).

(subject to revision to maintain consistency). The new information had a status of knowledge, and, because of this, the corpus of beliefs had to accommodate to it.

In the classical case of

(1) $\forall x\,(\mathrm{bird}(x)) \supset \mathrm{fly}(x))$
(2) bird (A)
when the information that $\neg\mathrm{fly}(A)$ is acquired, the strategy of minimal belief revision of Satoh changes the belief (1) into the following:
(3) $\forall x(x <> A \equiv \mathrm{bird}(x) \supset \mathrm{fly}(x))$

The main idea of Satoh for belief revision was that the order to minimize depended on the models of the previous beliefs as well as on the models of the new ones.[74] Satoh formalized this idea using second order formulas which he considered similar to McCarthy's Circumscription. The section "Related Research" of Satoh's work compared briefly his proposal to the known formalisms of NMR, to Doyle's TMS, to database updating work, which he disputed for not satisfying what Dalal calls "*irrelevancy of syntax*", and dedicated the final and more extended paragraphs to the "Logic of Theory Change" (AGM) and to the recent proposal by Dalal, which he considered similar to his own, although noticing some differences. Because AGM does not distinguish between knowledge and belief, Satoh considered the situation in which the knowledge was the set of tautologies. In this case, both definitions are comparable. Then, he proved that his proposal satisfied the AGM postulates, except for vacuity and the additional postulates,[75] although if the language was propositional, the first was also satisfied. As Dalal, Satoh remembers that it was a referee who mentioned AGM. "*When I wrote the FGCS 88 paper, I had a feedback from an anonymous referee which mentioned the following paper: . . .On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. . .*". [76]

It is interesting to notice that both mentioned authors did not know AGM theory at the moment of conceiving their work and, nonetheless, that they were in an almost perfect tuning with it. This fact reinforces our idea that the AGM formalization appeared in a key moment of restatement and search in AI that boosted its impact.

In 1989, "*Minimal Change and Maximal Coherence: A Basis for Belief Revision and Reasoning about Actions*" was published.[77] Its authors, Anand Rao and Norman Foo, researchers from the University of Sydney, Australia, set themselves to approach a burning question in AI: the reasoning about the outcomes of the actions that an agent performs through time. These actions affect the state of the external

[74] Contrary to the AGM model, which only depends on the original ones.

[75] The Vacuity postulate says that if the new information is not contradictory with the old one, then revision consists simply of including it directly (without any removal). The additional postulates say that to revise by a disjunction has to be the same as revision by one of the disjuncts or the revision by the other, or the intersection of both.

[76] Ken Satoh. Personal communication to the authors.

[77] A. S. Rao and N. Foo (1989).

world, including lateral effects or ramifications of the actions.[78] This reasoning involves determining the new state of the world after performing an action and the corresponding change in the beliefs of the agent when going from one instant in time to the following. In this respect, to reason about actions seemed, in the view of the authors, to be associated to belief revision, understood as a process by which an agent revises his set of beliefs in the present, starting from an input coming from the external world. In both cases, the central question was to determine what beliefs should be changed and which should not.[79] In other words, belief revision, as well as reasoning about actions, involved reasoning about change. The authors remarked that “. . .*Little work has been done in analyzing the principles common to both these areas. Work on belief revision has, for the most part, concentrated on building efficient systems that perform belief revision (Doyle 1979, de Kleer, 1986). Work on reasoning about actions has addressed both the computational and foundational aspects . . .but has been studied independently of belief revision. In this chapter we present a unified picture of both areas by formalizing the underlying principles of belief revision and reasoning about actions. Minimal change and Maximal coherence are two of the most important principles involved in reasoning about change. Minimization has been used in AI and philosophical logic in a variety of ways. In this chapter, by minimizing change we mean minimizing the acquisition of belief or minimizing the loss of belief when an agent moves from one state to another. By maximizing coherence we mean retaining as many coherent beliefs as possible during one state change. . .*” (see footnote 77).

For the task that they were set to accomplish, Rao and Foo recognized to have been inspired by the axiomatization of AGM theory of belief revision. The development of the work used the approach and nomenclature of AGM. They defined three possible states of belief of an agent with respect to a formula A, a time t and a world w, i.e., belief in A, belief in A and indifference with respect to A, and considered that a dynamics of belief consists in the process by which the agent goes from one state to one of the other two, giving place to the expansion, contraction or revision of his beliefs. Formally, they constructed a modal system named CS, for Coherence Modal System, with modal operators EXP, CON and REV that represent the three operations of the belief dynamics. This system was formally defined by a set of axioms for each operator. It seems certain that the paper, which was part of the doctoral dissertation of A. Rao, supervised by N. Foo, underwent the impact

[78] It was common to speak about the “frame problem” and the “ramification problem”, and these topics implied presumptive reasoning. Given that, a priori, all the possible consequences of the actions on a given scene are not known, the agent has to assume that it occurred a minimal change within what is expressed by the previous knowledge about the consequences of that actions and go into a new state of belief about the scene in which he is acting which is not “certain”. It is in this point that non monotonic reasoning enters into the problem.

[79] Later, Katsuno and Mendelzon developed the idea of “updating” as different from “revision” to distinguish between the consequences over a Knowledge Base of a change in the World from a change in the beliefs about the World.

of the acquaintance with AGM work when it was already under development and assimilated part of its concepts within a logical schema already established. This hypothesis may be considered confirmed by the testimony of Norman Foo.

> "*Around 1986 my then PhD student Anand Rao and I were working with agent beliefs and we wondered how an agent could change its beliefs when it discovers that some of them were wrong. It was obvious to us that if it gave up some beliefs it may also have to give up others related to it. I was going to New Zealand to visit an ex-professor of mine (in Electrical Engineering), Jack Woodward, and when I was there in Auckland he told me that Krister Segerberg was the professor of philosophy in the University. Jack was then the Head of Electrical Engineering, and he knew Krister well. He arranged for me to meet Krister in the latter's house. When I met with Krister he told me that Peter Gärdenfors was in fact writing a book on belief revision. So when I returned to Sydney I contacted Peter and he offered to send to Anand and me his draft chapters to read. These chapters were eye-openers for us, and we became aware of the AGM seminal work then*".[80]

From this moment on, a strong tradition in belief revision based in AGM started in Sydney.[81] Later research encompassed several areas not immediately recognizable as derived from AGM theory, but that was inspired in and by it. In the words of Norman Foo: "*These chapters were eye-openers for us, and we became aware of the AGM seminal work then. We never looked backed, and I focussed the research of my group on belief revision for the next decade. My students soon produced cutting-edge work in the area. We were also fortunate in being able to attract international post-docs in the area to work with us.* Anand Rao and I produced the first Prolog implementation of a version of AGM revision. *My students, besides Anand Rao (who went on to co-invent BDI logics with Mike Georgeff), who did PhDs in the AGM style – particularly with finite base revision or applications to reasoning about actions – were Mary-Anne Williams, Simon Dixon, Pavlos Peppas, Yan Zhang, Maurice Pagnucco and Boon Toh Low... Our research has since moved on to using ideas inspired by the AGM paradigm to many areas which may not be recognized as such by people not familiar with the history, but we can confidently say that had it not been for the pioneering paper by Alchourrón, Gärdenfors and Makinson we would not have progressed so quickly and so far*."[82]

To put in scale the weight of this tradition, born from the crossover, in the appropriate moment, between a genuine need of AI and a theory that – without intending it – came to the encounter of the first, we can say that the production by Foo, Rao and their students amount to 7% of the references to the AGM article that we collected, and this number increases to 11% if we consider it with respect to the references coming only from AI.[83]

[80] Norman Foo. Personal Communication to the authors

[81] Peter Gärdenfors in a personal communication to the authors confirms that "...when I was in Canberra in the fall of 1986 I was contacted by Norman Foo and Annand Rao in Sydney who invited me to give a talk on belief revision...".

[82] Norman Foo. Personal communication to the authors.

[83] The tables of references are in Section 5.

The other pioneering works, contrary to the previous ones, were acquainted with AGM from the start.

Bernhard Nebel in "A Knowledge Level Analysis of Belief Revision"[84] remarked that: "*... Revising beliefs is a task any intelligent agent has to perform. For this reason, belief revision has received much interest in Artificial Intelligence. However, there are serious problems when trying to analyze belief revision techniques developed in the field of Artificial Intelligence on the knowledge level. The symbolic representation of beliefs seems to be crucial....*" (see footnote 84). He explained this fact by "*...In Artificial Intelligence, a number of so-called truth-maintenance systems.... were developed which support belief revision. However, the question remains how belief revision can be described on an abstract level, independent of how beliefs are represented and manipulated inside a machine. In particular, it is unclear how to describe belief revision on the knowledge level as introduced by Newell (1981). Levesque and Brachman (1986) demanded that every information system should be describable on the knowledge level without any reference to how information is represented or manipulated by the system. However, this seems to be difficult for belief revision. A large number of authors seem to believe that ...considerations of how beliefs are represented on the symbol level seem inevitable for belief revision. ... Reconsidering Newell's original intentions when he introduced the notion of the knowledge level, we note that the main idea was describing the potential for generating actions by knowledge and not providing a theory of how knowledge or beliefs are manipulated.... Hence, we may conclude that belief revision is a phenomenon not analyzable on the knowledge level....However, the theory of epistemic change and the logic of theory change developed by Alchourrón, Gärdenfors, and Makinson ... show that at least some aspects of belief revision can be subject to a knowledge level analysis...*". [85]

Starting from AGMś rationality postulates as a specification in the KL, Nebel searched to reconstruct revision functions in the symbolic level. About the logic of theory change he said that "*This approach ...recently received a lot of interest in the AI community...*" (see footnote 84) and supplied as bibliographic references the presentation by Gärdenfors and Makinson in TARK'88 and the work by Dalal already mentioned. The statement "...received a lot of interest..." backed by an insufficient number of publications of AI researchers is only explainable if the impact was very recent and it was promoting work which had not yet reached its publication.

Nebel recognized that the logic of theory change used an idealized approach that ignored important characteristics required by a revision in a computational environment. In particular, he quoted Gärdenfors himself, when he stated that "*Belief sets cannot be used to express that some beliefs may be reasons for other beliefs...*".[86]

[84] Bernhard Nebel (1989).

[85] Bernhard Nebel (1989). In our opinion, the relativization made by Nebel ("...at least some aspects of belief revision can be subject to a knowledge level analysis...) derives from the "radical incompleteness" that characterizes the knowledge level in Newell's definition.

[86] P. Gärdenfors (1988).

This characteristic seemed to diminish its usefulness for AI, compared to TMS (that, by the way, were constructions in the symbolic level, without an independent characterization of the representational structures). Having this in mind, Nebel remarked that the maintenance of the reasons and justifications could be obtained as a lateral effect if the contraction operation over finite bases was "correctly" selected.

In the first part of his work, he presents the different variants of contraction functions ("*full meet*", "*partial meet*" and *maxichoice*).[87] Then, he defines operations of contraction over finite bases or "belief bases". To be able to prove that a contraction operation çover finite bases satisfies, in some sense, the AGM postulates, Nebel establishes that, for KB a finite belief base, a theory A and a contraction operation over theories $\downarrow$ can be defined in the following way:

$$A = \text{def Cn (KB)}$$
$$A \downarrow x = \text{def Cn}(\text{KB} \sim x)$$

Considering the propositions in the base as "more important" than the derived ones, Nebel constructed a selection function which he used to prove that contraction over bases could be considered as a partial meet contraction. Then, he proved that it satisfied the postulates of contraction (except conjunctive inclusion[88]). He defined revision using Levi's identity, as

$$\text{KB}^{\circ}x = \text{def}(\text{KB} \sim \neg x) \wedge x$$

This allowed him to affirm that the finite case of contraction (over belief bases), was not qualitatively different from the operation of epistemic change over deductively closed sets of beliefs, and that "*...The finite case can be modeled without any problem by a particular selection function. Viewed from a knowledge-level perspective, the only additional information needed for belief revision is a preference relation on sets of propositions. It should be noted, however, that the construction did not lead to an epistemic change function which satisfies all rationality postulates* " (see footnote 84) and that, for conjunctive inclusion to be valid, the selection function was required to be linked to a transitive relation.

At the time Nebel was writing his paper,[89] he was in touch with Gärdenfors (who gave him drafts of his ongoing work and made comments about the advances of Nebel's) and was also connected to Borgida and Dalal, to whom he expressed his gratitude for their "hints and ideas". With respect to the way in which he got

[87] The bibliographic references he mentions embrace almost all the publications of the authors of the theory, from the paper about contraction by Alchourrón and Makinson in 1982 to the paper by Gärdenfors on that same year of 1989, including the one presented in TARK'88.

[88] When you want to delete a conjunction from a theory, at least one of the conjuncts has to be extracted (if not, given that the result is closed by consequence, you would obtain the conjunction once again). The postulate of conjunctive inclusion expresses that if one of the conjuncts is deleted, you should expect that all the formulas that would have been deleted when making an explicit contraction only of that conjunct are also deleted.

[89] Nebel developed his work in the context of an Esprit Project about knowledge management.

acquainted with AGM, Nebel remarks that the connection came through the relation between his thesis supervisor and the philosopher H. Rott. "*I heard about 'belief revision', when I was writing my Ph.D. thesis, probably in 1987 or 1988. My thesis topic was to attack the problem of changing terminologies – or what is called ontologies these days. My supervisor, Wolfgang Wahlster, told me at that time about the Ph.D. thesis work of Hans Rott, which he knew because Wolfgang Wahlster and Hans Rott are both in the 'Deutsche Studienstiftung. In any case, I believe that we set up a meeting and talked about 'belief revision'"*.[90]

The attractiveness of AGM resided, for Nebel, in that it approached in a formal way central questions for AI, which at that moment remained treated by means of ad hoc mechanisms. "*I later started to look into the existing literature from philosophical logic and was fascinated, in particular because AGM addressed the problem of how logical theories could evolve. At that time, there existed a few approaches to model similar things in Computer Science, but they were all more pragmatic and ad hoc...(Also) it seemed to be an interesting operation missing from databases and seem to be similar to what so-called Truth-Maintenance-System did. Further, belief revision appeared to be dual to reasoning in non-monotonic logics"*.[91]

Of all the pioneering work, the one by Katsuno and Mendelzon ("A Unified View of Propositional Knowledge Base Updates"[92]) was the first that attempted a unifying perspective of all the previous proposals that tried to solve the problem of knowledge base updating, in which AGM model was at the centre.[93] In the Introduction, after presenting the question of revision of a KB (conceived as a finite set of sentences in a language L) and defining – informally – the operations of revision, contraction, deletion ("erase" a sentence and its logical equivalents from the KB) and retraction (to unmake the result of a previous operation), the authors remark that "*...Foundational work on knowledge base revision was done by Gärdenfors and his colleagues...The Gärdenfors postulates do not assume any concrete representation of the KB, in fact, KB's are modeled as deductively closed sets of sentences in some unspecified language. When we consider computer-based KB's, we need to fix a formalism and a finite syntactic representation of a KB...The question now arises of whether the result of an update will depend on the particular set of sentences in the KB, or only on the worlds described...*" (see footnote 92). In consequence, they considered that any method oriented to finite knowledge bases had to satisfy Dalals ́ Principle of Irrelevancy of Syntax, "*...the first one to relate his approach to the Gärdenfors postulates, pointing out that his proposal for the revision operator satisfies them...*" (see footnote 92). Katsuno and Mendelzon intended to advance towards a more general characterization, in model theory, of the revision operators for finite bases which

[90] Bernhard Nebel. Personal Communication to the authors.

[91] Bernhard Nebel. Personal Communication to the authors.

[92] H. Katsuno and A. Mendelzon (1989).

[93] Contrasting with Dalal's comparison, this review of proposals intended to verify which AGM postulates satisfied each one of the analyzed frameworks.

would satisfy the AGM postulates. Their main result was a theorem for the characterization of the revision operations that satisfied the six (eight) AGM postulates based on a partial (total) preorder among models which only depend on the KB.

Using this idea of characterizing revision operations in terms of orders among models of the old and new information, Katsuno and Mendelzon, analyzed, as was already mentioned, the different proposals for knowledge base updating that had been published in the previous years, showing that all of them could be captured by this semantics, just by considering in each case which is the order among models, of the old base as well as the new one, that underlies each construction. This unified view included the proposals of Fagin et al., Borgida, Weber, Winslett, Dalal and Satoh (although for Satoh the comparison had to be restricted to the propositional case, since Satoh worked with a KB in first order sentences).

The acquaintance with AGM by Katsuno and Mendelzon occurred by way of a conversation about knowledge base updating of the authors with Ray Reiter, distinguished AI researcher, known by his "*Default Logic*", who had attended TARK'88.

> "*The story goes back to 1988. At that time, I worked for NTT (NipponTelegraph and Telephone Corporation), and NTT kindly gave me a chance for a kind of sabbatical. Then, I stayed at Toronto for one year as a visiting scientist, and Alberto hosted me. I arrived at Toronto in the end of August 1988. In the beginning of September, Alberto arranged a short meeting with Ray Reiter, because Alberto knew that I was very much interested in a series of Ray's works. Then, I introduced myself to Ray, and I probably said to him that I was interested in update of knowledge bases. At the end of the meeting, Ray gave us a copy of two papers: Peter Gärdenfors and David Makinson, Revisions of Knowledge Systems Using Epistemic Entrenchment, 83-95, Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge, Pacific Grove, CA, March 1988 Morgan Kaufmann 1988 and Dalal M., "Updates in Propositional Databases", Technical Report, DCS-TR222, Department of Computer Science, Rutgers University, February 1988. I knew for the first time the existence of AGM paper through the above Gärdenfors and Makinson paper, and I guess that Alberto also didn't know AGM work until then. At least, Alberto didn't say to me that he knew Alchourrón at that time. I remember that, later, Alberto informed me that Alchourrón was a famous philosopher in Argentina, but I cannot recall when he said so".*[94] Also in 1988, the year of the second TARK, the Second Workshop on Non Monotonic Reasoning was held in Germany.[95] The main participants were distinguished AI researchers, who presented here their proposals, some of which used *ad hoc* mechanisms, whereas others made recourse to the toolbox of modal logics. But what especially distinguished this Workshop was the fact that, as a reflection of the formal interests within AI, David Makinson was invited to speak and that his work was the only presentation in the section of General Aspects of Non Monotonic Reasoning (NMR).[96] His presentation of a conceptual framework that allowed to characterize the properties of the notion of logical consequence underlying the several ad hoc formalisms of NMR (like Reiter's Default Logic) made a big impression, probably, because of the same basic reasons that caused

[94] Hirofumi Katsuno. Personal communication to the authors. Regrettably, the distinguished argentine researcher, resident in Canada, Alberto Mendelzon, died prematurely in 2005.

[95] Second Workshop on NMR, Grassau, Germany, June 1988.

[96] David Makinson (1989).

the impact of AGM.[97] Karl Schlecta, an AI researcher present at the Workshop, says that probably his first acquaintance with AGM was through a personal conversation with D. Makinson in Grassau.[98] The following year, Schlechta published "Some results on Theory Revision",[99] written under the influence of AGM and, in particular, of "epistemic entrenchment". Schlechta remarks that "*...The problem of Theory Revision is to 'add' a formula to a theory, while preserving consistency and making only minimal changes to the original theory. A natural way to uniquely determine the process is by imposing an order of 'epistemic entrenchment' on the formulae as done by Gärdenfors and Makinson...*" (see footnote 99). He recognized as a limitation of the model, the difficulty to approach consecutive (iterated) revisions as these are a very common phenomenon in cognitive systems. The first part of his work aimed at a proposal that would overcome this limitation (to construct an order for all the theories within the same language). The second part, independent from the first, was about contractions of $< K, A >$ systems, where K was a set of formulas deductively closed and A a set of axioms for K. In this case, following Schlechta, a revision for $< K, A >$ consisted, essentially, in selecting an adequate subset of A. In the final section, he proposed methods for selecting reasonable consistent subsets of – partially ordered – sets of conflictive information, "hard" as well as default, in a context of languages partially ordered, augmented by "default" information, and where more specific information was more trustworthy in case of conflict, and the incomparable information was treated in a "fair" way. At the time Schlechta was writing his work, he was a colleague of Nebel, in IBM Germany as well as in the Lilog project of the European Community, and was in touch with Gärdenfors and Makinson, to whom he expressed his gratitude for their reading of his drafts and the suggestions received.

From a symmetrical perspective to the TARKs and in the midst of the growth of the belief change community, a Workshop was organized in Konstanz, Germany, in 1989, named "The Logic of Theory Change", whose focus was AGM and its derivations.[100] This was the only opportunity in which the trio composed of Alchourrón, Gärdenfors and Makinson was physically united. In the preface of the Workshop Proceedings, the editors remarked that "*...the logic of theory change is one of the most fecund research programs in philosophical logic to have emerged in recent years. Apart from throwing new light on old problems and generating interesting problems on its own, it has quickly established important links with research in artificial intelligence.. Much of the work on theory change may be read as contribution towards a unified theory of database updating...*" *and then added* "*...Revision of theories is a nonmonotonic operation: in the process of revising a theory so as to include A some of the original theory may get lost. This observation suggest connections to non monotonic reasoning...*". [101] To this event several AI researchers (Doyle, Martins, Brewka and Schlechta) were invited. Among these presentations, Martins did not refer to the AGM theory. Doyle included AGM among other bibliographical references, only to present database updating, such as was considered by

[97] D. Makinson. Personal communication to the authors.

[98] "I probably first heard about Theory Revision in a talk given by David Makinson at NMR Workshop 6/88 in Grassau, Germany". Karl Schlechta. Personal communication to the authors.

[99] Karl Schlechta (1989).

[100] Workshop on The logic of theory change. Andre Furhmann and Michel Morreau. Konstanz. Germany. October 1989.

[101] A. Furhmann and M. Morreau (1990).

AI in those years, and the revision of belief sets, such as was considered in philosophy, as analogous phenomena, and Brewka developed, as the others, his own system and, in passing referred to AGM, focusing on the divergence between the treatment of belief sets (closed by logical consequence) and the need of AI to operate with finite bases. In this opportunity, only Schlechta made AGM a substantial reference of his presentation, as we have already noted. We can say that Konstanz functioned as an inverted TARK'88 and that its effect was to "close the circle" of the connections between fields of research. In the following years, all the AI researchers that participated in Konstanz quoted or made reference to AGM, and the majority of the people invited, independently of their disciplinary origins, worked jointly or separately, in topics of belief change as well as non monotonicity, under the influence of AGM theory and its derivations.

As it surfaces in the discussions contained in these first works, a fair number of proposals treated KB updating, but with strong limitations, be it because the case of revision by inconsistent information with the already existent was put aside of the formalization, because it was approached by *ad hoc* procedures in the symbolic level, without an appropriate clarification, because the proposals did not satisfy the principle of irrelevancy of syntax, or did not satisfy the intuitive principle of minimal loss. These situations led to the adoption of AGM as a model of how to specify in abstract form the properties of a mechanism for revision. In the years following 1990, AGM appeared referenced by numerous researchers in the area of AI, such as R. Reiter, J. Doyle, M. Winslett, J. Halpern, Brewka, Shoham and J.P. Delgrande, among others. Although many disputed its limitations (such as the difficulty to iterate changes) or its idealized or excessively simplified nature, nevertheless it became an obliged reference, as the following paragraph from a 2000 paper, "Belief Revision: a Critique" shows, one of whose authors is J. Halpern, the main organizer of the first TARK.[102]

"The approaches to belief change typically start with a collection of postulates, argue that they are reasonable, and prove some consequences of these postulates. . .The main message of the paper is that describing postulates and proving a representation theorem is not enough. While it may have been reasonable when research on belief change started in the early 1980s. . .it is our view that it should not longer be acceptable. . .While postulates do provide insight and guidance, it is also important to describe what we call the underlying ontology or scenario for the belief change process. . ." (see footnote 102). And, further along, the authors remark that ". . .Our focus is on approaches that take as starting point the postulates for belief revision proposed by Alchourrón, Gärdenfors and Makinson (AGM for now on). . ." (see footnote 102).

[102] N. Friedman and J. Halpern (1999).

1.5 AGM 20 Years Afterwards

Undoubtedly, the AGM model has become one of the paradigms used to develop systems that have to "*cope*" with inconsistent and/or erroneous information. Both in Artificial Intelligence as well as in Computer Sciences, most works trying to approach – by using a formal support – the problem of operating with inconsistent or wrong information, follow the general ideas of this model. A recent example of this phenomenon is "fail proof" (tolerant) systems, where it has been accepted that the system may suffer or generate errors and the outcome of an erroneous situation is "repaired" by means of a revision function. Curiously enough, this may be seen as a late (and unintended) endorsement of some of the ideas in Alchourrón and Makinson's first paper, "*Hierarchies of regulations*..."(see footnote 14). In the course of time, an appropriation of the model by the above sciences occurred, by adaptation, by rethinking various assumptions, or setting aside some and implementing others. It may be speculated that this level of acceptance is essentially due to the fact that it was first in proposing the effective construction of change operations and that, at the same time, it has overcome both the merely descriptive and ad-hoc, as well as the purely axiomatic.

In this section, we seek to quantitatively and qualitatively measure the degree of impact and acceptance of the AGM model. Before introducing ourselves in the above analysis, it is well worth noting that AGM – like any other model – is an idealization that responds to certain intuitions and motivations translated into assumptions which support these intuitions and motivations (all of them questionable *a priori*). The first is that an agent's corpus of beliefs (or belief state) is represented by a set of formulas (possibly infinite) of classical propositional logic, closed by a consequence operation which includes tautologies, is compact and satisfies the rule of "not introducing changes in the premises". Such theories are stable (that is to say, they do not change spontaneously but rather as a reaction to an outside stimulation) and there is no distinction in them between explicit beliefs and derived beliefs. An agent believes in something if its formulation in the language belongs in the theory representing the agent's beliefs. Formulas represent agent beliefs, rather than what the agent knows or should know (an epistemic rather than ontological interpretation). On the other hand, the information triggering change is represented by a formula showing preference for the pre-existing theory. Besides, there are only three epistemic attitudes regarding belief (acceptance, rejection or indetermination), which in turn is what conditions the existence of an equal number of change operations: expansion, contraction, and revision. These are functions which have theory-formula ordered pairs by way of domain and theory by way of rank, and they must comply with certain rational criteria -such as operating by "minimal change" or "maximal preservation", postulating consistent theories, the irrelevance of syntax (the outcome does not depend on the manner in which beliefs are described), and fairness when selecting the resulting theory (if there are several candidate theories which satisfy the above criteria, choosing between them may not be arbitrary). In the case of the AGM model, such choice arises from a transitive total relation outside the corpus. Specifically, it turns out that Expansion is a particular case of

Revision, and Revision and Contraction are inter-definable. Besides, these operations are one-shot, in the sense that the history of any successive changes generating the next change is not taken into account.

All of above assumptions have been disputed for different reasons by different authors. Most of the papers published in recent years have proposed alternative models, which – in essence – pose the following issues:

(1) Changing the choice of representation language for some richer language such as First-order Logic, Modal Logic, Many-valued Logic, Conditional Logic, etc. In general, this entails proposing new notions regarding "consistency" of information, contradiction, derived information, and connection of elements in the set. And using languages not quite as rich, such as Description Logics or Prolog, where the consequence notion does not satisfy all of the properties stated above. Moreover, once the language has been fixed, non-deductive consequence notions might be proposed.

(2) Whichever the chosen language, the question arises of how the corpus of information will be represented: if by means of a single sentence in the language (the conjunction of independent sentences) or by means of a (perhaps infinite) set of them. In the case of a set of sentences, the set could be a closed set, in accordance with the notion of a logical consequence, or a simple enumeration of "naked" facts. This second option evinces the need to then calculate, somehow, the consequences of these facts as well as making a commitment to differentiate -or not- between explicit and implicit information. In the literature, these alternatives appear under the label of Belief Bases vs. Belief Sets.

(3) The above point refers to a flat structure. However, may the corpus have a more complex structure? Can it be made hierarchical? May any information on how to modify that same corpus be presented? Furthermore, if – as is well known – the *epistemic entrenchment* for a theory *K*, characterizes it fully in terms of its change operations, then why not just postulate a change of order? The affirmative proposal is intended to give a response to iteration problems.

(4) Once the language and the way the corpus is represented have been selected, we need to consider how the information that has triggered the change is represented. Is there to be a difference between them? Even if the language were the same, must both representations be homogeneous? If yes, are they both to be formulas or sets of formulas? The former is upheld by most computer science proposals; the latter gives rise to multiple change functions.

(5) Once the above choices have been addressed, we need to fix the characteristics of change operations. Are they able to trigger "spontaneously" or do they need external "stimulation"? That is, are belief states internally stable? Will the epistemic values assigned to language expressions be only acceptance, rejection, indetermination; or must relevance and acceptability degrees be considered? Should change operations respect relevance notions? Which

and how many different ways are there in which a corpus of information may be modified? Which are these ways? Are they independent or inter-definable?

(6) Must there be a relationship between the notion of corpus-belonging in a sentence and the truth of such sentence? May there be an information corpus with false or locally inconsistent information? Must the expressions of language be interpreted as true or false, or neither true nor false, or must there be degrees of truth, inconsistency, uncertainty, etc.? Is the notion to be preserved that of consistency? What happens if we replace consistency with a weaker notion such as coherence?

(7) If we accept to define operations responding to the notion of minimal change, or maximum preservation of the information corpus, it becomes necessary to have some manner of "calculating the value" of the information to be disregarded. Is there an order of preference representing the credibility, soundness or informational value between language expressions? Is this order included in the information corpus, or is it inherent to the change operation? Must minimal change be quantitative or qualitative? Are there weaker notions than a transitive total relation to guarantee minimal change? Maximizing is the same as Optimizing?

(8) If, conversely, we do not accept minimization of information loss, why then – if new information is consistent with the corpus- is Revision equivalent to a simple aggregate? Why does Expansion have to coincide with the logical consequence operation? Why, if a belief is absent from a corpus, does Contraction not cause the modification of the belief? If minimization is a conservative methodology, why Reliabilism is not considered as an alternative?[103]

(9) The relationship between the original and the updated corpus, is it relational or functional? The function or relation exists only between the original corpus, the new information and the modified corpus, or are additional parameters to be considered? Must change operations take into account the history of changes occurred, or is each new operation independent from the previous ones? Must the corpus updating process maintain the interpretation of language expressions, or would it be thinkable for a change to modify the propositions associated to corpus expressions? Is revision of a corpus by an N sequence of sentences equivalent to N successive revisions? This is known as the iteration problem.

(10) Is the operation triggered by new contingent information to be successful always? May new information be accepted partially?

[103] Reliabilism is the principal concern of formal learning theory. See for example K. Kelly, O. Schulte and V. Hendricks 1997.

1.5.1 *Quantitative Analysis of AGM Impact*

The measuring system used is based on the analysis of bibliographic references. This has afforded us a precise context for search and quantification. In turn, despite the broad spectrum of potential selection – including indirect references to the model-, we have decided to limit ourselves exclusively to checking those articles and books which specifically mention (eliminating self-references) the original 1985 AGM, and the book published by Gärdenfors (see footnote 86) in those years, since, in our opinion, only they indicate beyond any doubt that AGM is conceptually present in the article in question. As a result of the search, a worksheet has been prepared with over 1,400 entries, and for each entry, there is a paper title, authors, place of publication and year. The data thus listed were validated and cross-checked with different information sources. Later, the different works were sorted into four different categories – Philosophy, Artificial Intelligence, Computer Science and Logics. In order to categorize them, author profiles, areas of expertise and place of publication we taken in consideration. Other disciplines, such as Cognitive Sciences or Economy, were classified as Philosophy.[104]

Table 1.1 summarizes search and classification quantitative results. The graph in Fig. 1.1 illustrates how references to AGM evolved in time. For the sake of improving visuals, the categories Logics and Philosophy have been grouped together and the same for Artificial Intelligence and Computer Science. We may see that the original article has been increasingly cited, remarkably as of 1994, when references doubled those of 1993. Then, as of 1996 and up to the present, they have experienced exceptional growth (we have not included 2007 on the tables, as we deem that data incomplete). We have also remarked that the proportion of AI plus Computer Science works has not dropped below 50% since 1992 – and, setting aside 3 years, it has always been over 60% – which has been the stable minimum from 2001 to 2006.[105]

1.5.2 *Qualitative Analysis of AGM Impact*

From the publications dated 2006, 81 articles remained in the AI and CC categories. A qualitative analysis was done as a case study, as we sought to characterize the context and show how we put the theory to use.[106] The following characteristics are the result of our analysis:

[104] It is possible to access the table in http//:www.dc.uba.ar/people/ricardo/referencesAGM.xls

[105] It should be noted that, during the first years of the period under analysis, a number of papers were not electronically available for dissemination, nor were they rendered in that format at a later date; for this reason, the cites quoted for those first years are clearly incomplete; a remarkable case is that of Rao and Foo's work, commented under Section 1.4.

[106] Originally, this research meant to honour the memory of Carlos Alchourrón in commemoration of the 10th anniversary of his passing. Hence, we selected this year due to the fact that, at the time, our well established, consolidated data was that of the previous year.

Table 1.1 Number of publications per year. Absolute and relative

Year	Phil+Log	%Phil+Log	AI+Comp	%AI+Comp
1986	2	66.7	1	33.3
1987	1	100.0	0	0.0
1988	5	62.5	3	37.5
1989	13	76.5	4	23.5
1990	5	55.6	4	44.4
1991	11	47.8	12	52.2
1992	11	32.4	23	67.6
1993	11	29.7	26	70.3
1994	19	27.1	51	72.9
1995	25	39.7	38	60.3
1996	23	32.4	48	67.6
1997	24	31.6	52	68.4
1998	37	41.6	52	58.4
1999	47	44.8	58	55.2
2000	50	46.3	58	53.7
2001	46	39.7	70	60.3
2002	44	42.7	59	57.3
2003	36	31.3	79	68.7
2004	45	31.3	99	68.8
2005	46	36.2	81	63.8
2006	41	48.2	44	51.8
Totals	542	38.6	862	61.4

(a) The most attention has been afforded to those that set out to solve the problem of iterative revisions. In all, there are nine papers which, resorting to different proposals, address the problem directly. Essentially, the question here has to do with determining how preferences change vis à vis the option of changing the beliefs themselves. Several articles have used ad hoc "measures" regarding beliefs (rankings, possibilities, probabilities, Dempter-Shafer, utility, etc.), showing whether operations satisfy the postulates originally proposed by Darwiche and Pearl.[107] Some of these models are also non-prioritized. A detailed reading clearly evinces that there still is a lack of consensus regarding acceptable characterization of this family of operations.

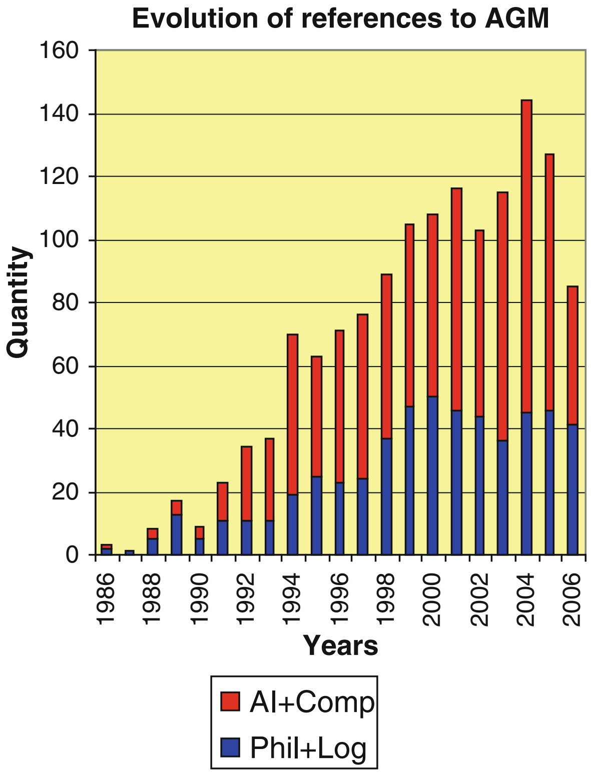
(b) Another outstanding topic is the modeling of abductive reasoning through Aggregate Functions (a special case of Expansions[108]), addressed by six works which follow Pagnucco's line.[109] Originally, Aggregate Functions were postulated by Rott to overcome trivialization problems when conditional sentences are allowed in representation language, and these conditional

[107] Adnan Darwiche and Judea Peral (1997).

[108] Hans Rott (1989).

[109] M. Pagnucco (1996).

Fig. 1.1 Evolution of number of references in the course of time



sentences are then interpreted in terms of a Revision function. Essentially, an Aggregate operation is an Expansion not satisfying the notion of minimality.

Another approach giving up minimality, appears in a pair of papers integrating and combining both learning theory and belief revision. The AGM theory of rational belief revision had said little about whether its proposal assists or prevents the agent's ability to reliably reach the truth as his beliefs change through time. In order to overcome this limitation, these papers analyze the belief revision theory of Alchourr'on, Gardenfors and Makinson from a learning theoretic point of view. They consider a reliability conception instead of a conservative one behind minimality.

(c) A group of five articles considers different proposals to generate *rankings* which represent the epistemic preferences guiding a change operation. Several of them seek to incorporate empirical models which try to determine or capture human preferences determined by various sources such as emotions, frames of mind, etc. Some of these proposals also solve the problem of iterated revisions, since their representation of beliefs is focused on ranking, and change is associated with a change in ranking; thus, a new ranking is obtained based on the previous one and a new sentence.

(d) Several refer directly to the problem of dealing with inconsistent or conflicting information (depending on whether it involves a notion of truth or not). The spectrum is broad but a few cases are worth mentioning to give an overview of this group. One of them proposes an AGM-like axiomatization of the inconsistency level, drawing a parallel with the AGM model. Another one addresses modeling interaction between negotiation and mediation agents by means of an AGM variant. Others bring up inconsistency tolerance and, unknowingly, reinvent the "*delivery*" operation described on "*Hierarchies of regulations and their logics*". Some of the remaining articles illustrate techniques to discover and solve inconsistencies, and – in seeking to show a certain degree of logical formality – make reference to AGM as a pattern for comparison. Some ten articles may be included in this category.

(e) One line of research, also the common denominator for ten other articles, refers to multiple revisions (where revision is done with a set of formulas rather than with a single formula) and, in this context, to merge operations, where new information has no priority *vis à vis* old information. This kind of an operation may be seen as the revision of a set of beliefs by another set of beliefs, and is connected with preference aggregation issues.[110] Along this same line, the use of AGM revisions to model games where certain notions of balance are justified in terms of belief rationality and player interactions must be emphasized.[111]

(f) Common ground for some ten articles is the use of formal languages by way of an alternative to classical logics. Thus, withdrawn areas of natural language appear, "*description logics*", OWL, Prolog, modal languages, Situation Calculus, etc.. The main motivation corresponds to an increased demand for ontology languages, as a result of the applications in "Web Semantics" where – given the dynamics of these systems – change functions are incorporated. The characteristic that these works have in common is having to generalize the AGM model, be it because the language does not have all the logical connectives or because the notion of associated consequence, even if it is Tarskian, does not satisfy any one of the original AGM assumptions.

(g) A class formed by five articles refers to the relationship between the AGM model and the notion of nonmonotonic consequence. This association is quite extensive and significant, and its development is beyond the scope of this work. It is rooted in Alchourron's work[112] on the one hand, and in Gärdenfors and Makinson's on the other.[113]

[110] In the most recent Belief Revision literature from 2005 to the present there is increased attention to issues of belief merging and judgement aggregation (the latter viewed in comparison with older theory on the aggregation of individual preferences, in economics, and the aggregation of votes in political science). We owe this remark to David Makinson.

[111] The work of Cristina Bicchieri (1988b) is one of the precursors for this line of research.

[112] Carlos Alchourrón (1986).

[113] P. Gärdenfors and D. Makinson (1991).

(h) Finally, the remainder of works not included in the above enumeration, refer to standard AGM applications with algorithmic constructions and efficient revision functions and, except in very few cases, the reference is marginal.

(i) It is worth noting that ten of the articles analyzed, mention as a "novelty" the incorporation of the decision theory, used in Economics. In this context, they propose incorporating the notion of "optimization", whose ordering of the belief structure is weaker than the "classical" one, associated to the AGM model.

1.6 Some Conclusions

It is not unusual for a theoretical nucleus developed in a certain field and originally associated with a certain type of intended applications, to be rediscovered in a different field. This phenomenon usually occurs with theories of a relatively formal, abstract or normative, character, for example, the theory of rational decision or game theory. But, in every case, there are particular conditions that make possible the "jump" from one application domain to another, and they are the ones that explain, in a great measure, the episode. In the particular case of the Logic of Theory Change, our research started with a question about those particular conditions:

Why did AGM receive such swift acceptance from the AI community?

As we have shown, TARK'88 was the entry point in AI of AGM theory, with the presentation of the work by Gärdenfors and Makinson, and the presence in the Conference of renowned figures of AI, such as Ray Reiter, Jon Doyle, J. Halpern, H. Levesque, R. Moore, M. Vardi and R. Fagin, among others. In fact, 1988 was an inflexion point, since, in that year and the following, the pioneer works that made use or referenced AGM theory in an active way were published. We also noted that the very existence of the TARK Conferences was a consequence of the revival, during the eighties, of the interest provoked by the topics studied by AI among researchers from others fields like philosophy, logic, linguistics, psychology, etc. The AGM presentation in 1988 was part of this intellectual interchange. However, this event, by itself, does not explain the adoption of AGMs ́formalization model by numerous researchers in AI and does not answer our initial question.

We have considered in this chapter four elements that, taken *in toto*, help us explain the fast diffusion of the AGM model in an important sector of the AI community: the concern for the formal foundation of the systems developed in the area, and, in particular, in Knowledge Representation, with a reevaluation of the role of logic in the "Knowledge Level"; the introduction in AI, motivated by that concern, of the literature of Belief Revision; the discovery of the problems involved in the updating of databases with inputs inconsistent with its previous contents, and the attempts to solve them; and, finally, as we mentioned above, the interdisciplinary matrix that many activities in AI had at the time, which eased the acceptance of new models from other areas.

All these elements allow us to conclude that the AGM model attached itself and was functional to a previous process of formulation and search of solutions for the problem of KB updating by an influential sector of the AI community. A fair number of proposals faced strong limitations due to causes such as the following: revision by information inconsistent with the already existent was put aside of the formalization; revision was approached by *ad hoc* procedures in the symbolic level, without an appropriate clarification; the proposals did not satisfy the principle of irrelevancy of syntax or the intuitive principle of minimal loss. Briefly, the researchers involved in this search faced the difficulty of combining a precise and clear semantics with a “realistic” scheme from a computational point of view. In fact, for many of them, the revision processes seemed to be indissolubly linked to the manipulation of beliefs stored in concrete systems. In the words of Nebel, it was thought that “belief revision is a phenomenon not analyzable on the knowledge level” (see footnote 84).

It was not by chance that the starting point of Alchourrón and Makinson, i.e., the problem of the multiple result of the derogation of a rule in a code, also faced the apparent inevitability of introducing an extralogical and totally *ad hoc* fact. However, the development of their work showed that it was possible to reduce that *ad hoc* component by way of characterizing different forms of ordering subtheories or sentences.[114]

With respect to some of the proposals that were being generated in IA, they were perfectly tuned to the AGM approach. In fact, they resorted to an order among models, although, in general, its properties were not explicitly enunciated. It was an implicit order in the construction of the mechanism, such as Katsuno and Mendelzon showed later on. In this sense, we may say that the AGM formalization showed up in a key moment of restatement and search in AI that boosted its impact. As we have already said, the emergence of an abstract model, “in the knowledge level”, with different kinds of descriptions: axioms, semantics and some constructive methods which all of then coincide exactly, produced a strong intellectual attraction. The high level of abstraction of the new approach was what a lot of researchers in AI were seeking.

In the years following 1990, AGM was referenced by numerous researchers in the area of AI, such as R. Reiter, J. Doyle, M. Winslett, J. Halpern, Brewka, Shoham and J.P. Delgrande, among others. Although many disputed its limitations or its idealized or excessively simplified nature (such as we noted in Section 1.5), nevertheless it became an obliged reference. We believe that its place as a “model pattern” for so many years, and even its extension as such to areas like economic theory, is mainly due to having been the first model to postulate the effective construction of change operations and to overcome, at the same time, the merely descriptive and ad-hoc, and the purely axiomatic.

However, it was not an influence in only one direction. The later evolution of the model owes much to its assimilation by researchers in AI, as it is shown in Section 1.5, as well as in other disciplines. With the passing of time, these areas made the model their own, adopting it, reformulating several of its assumptions, abandoning

[114] The extralogical factor is not totally eliminated (since the behavior of an agent is not wholly definable in the KL), but the arbitrariness can be restricted. The works on safe contraction as well as the ones on epistemic entrenchment go in depth in this sense.

some and enforcing others. This dynamics of evolution also justifies the survival of the original ideas.

Finally, it is worth mentioning the impact that the AGM model has had in two other areas of knowledge that, in our understanding, would deserve an extended analysis like the present one. These areas are Non Monotonic Reasoning and – in more recent years – Game Theory and other problems in theoretical economics. In the first case, its origin is due also to a paper by Gärdenfors and Makinson (see footnote 112). In the second case, the first application of AGM to game theory was the work by Cristina Bicchieri, published in TARK'88, already mentioned in Section 1.3 (see footnote 112).

As David Makinson recently said: "The Basic AGM approach remains a starting point for fresh journeys and a platform for novel constructions. . .".[115]

Acknowledgments We would not be able even to attempt to reconstruct the history of the introduction of the AGM Theory in Artificial Intelligence without invaluable contributions of many of the principle actors of those years' long, extensive and effervescent discussions. All of them approached our inquiry with enthusiasm and made true efforts to honor us with their emotive memories.

Therefore we want to thank Cristina Bicchieri, Horacio Arló Costa, Alex Borgida, Mukesh Dalal, Norman Foo, Peter Gärdenfors, Joseph Halpern, Gilbert Harman, David Israel, Hirofumi Katsumo, Hector Levesque, Bernhard Nebel, Hans Rott, Ken Satoh, Karl Schlechta for accepting to respond our questions and being our partners this time.

We are very grateful to our friend Eduardo Fermé for his sharp observations that helped us to revise the original Spanish draft of our work.

Also, we are indebted to our colleague and friend José Alvarez who helped us to translate to English our first draft. In addition, during this process, he offered us a lot of helpful and interesting discussions on different parts of this chapter.

We own a very special gratitude to Gladys Palau, Eugenio Bulygin, David Makinson and Antonio Martino who committed themselves in a very personal way to the reconstruction of the history, patiently accepting our repeated inquiries and generously sharing their reminiscences of Carlos Alchourrón.

Thanks all of them to accompany us to honor the memory of Carlos Alchourrón whose disciples and friends we consider ourselves to be.

References

Alchourrón, C. 1986. Conditionality and the representation of legal norms. In *Automated analysis of legal texts*, eds. A.A. Martino and F. Socci Natale, 175–186. Amsterdam: North Holland.

Alchourrón, C., and D. Makinson. 1981. Hierarchies of regulations and their logic. In *New studies in deontic logic*, ed. R. Hilpinen, 125–148. Dordrecht: Reidel.

Alchourrón, C., and D. Makinson. 1982. On the logic of theory change: Contraction functions and their associated Revision functions. Theoria. *A Swedish Journal of Philosophy* XLVIII:14–37.

Alchourrón, C., and D. Makinson. 1985. On the logic of theory change: safe contraction. *Studia Logica* 44:405–422.

Alchourrón, C., and D. Makinson. 1986. Maps between some different kinds contraction function: the finite case. *Studia Logica* 45:187–198.

[115] David Makinson and George Kourousias (2006).

Alchourrón, C., P. Gärdenfors, and D. Makinson. 1985. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic* 50(2):118–139.

Bicchieri, Cristina. 1988a. Strategic behavior and counterfactuals. *Synthese* 76:135–169

Bicchieri, Cristina. 1988b. Common knowledge and backward induction: A solution to the paradox. In *Proceedings of the second conference on theoretical aspects of reasoning about knowledge* (2nd TARK), 381–393. California, CA.

Borgida, A. 1985. Language features for flexible handling of exceptions in information systems. *ACM Transactions on Database Systems* 10:565–603.

Dalal, Mukesh. 1988. Investigations into a theory of knowledge base revision. Preliminary report. In *Proceedings of the conference of the American association for artificial intelligence*. August 21–26, 475–479. Sant Paul, MN.

Darwiche, Adnan, and Peral Judea. 1997. On the logic of iterated belief revision. *Artificial Intelligence* 89:1–29.

Doyle, Jon. 1979. A truth maintenance system. *Artificial Intelligence* 12(2):231–272.

Doyle, Jon., and London, Philip. 1980. A selected descriptor-indexed bibliography to the literature on belief revision. *ACM SIGART Bulletin* 71:7–22.

Fagin, Ronald, Ullman, Jeffrey, and Vardi, Moshe. 1983. On the semantics of updates in databases. In *Proceedings of the Second ACM-SIGACT-SIGMOD.SIGART Symposium on Principles of Database Systems*, 352–365. Atlanta, GA.

Fagin, R., G.M. Kuper, J.D. Ullman, and M.Y. Vardi. 1986. Updating logical databases. In *Advances in computing research,* eds. P. Kanellakis and F.P. Preparata, vol. 3, 1–18. London: JAI Press.

Friedman, N., and J. Halpern. 1999. Belief revision: A critique. *Journal of Logic, Language and Information*, 8:401–420.

Furhmann, A., and M. Morreau. 1990. Preface of "The Logic of Theory Change". In *Lecture Notes in Artificial Intelligence*. Berlin: Springer.

Gärdenfors, Peter. 1979. Conditionals and changes of belief. In *The logic and epistemology of scientific change*, eds. I. Niiniluoto and R. Tuomela, 381–404. Amsterdam: North Holland.

Gärdenfors, Peter. 1980. A pragmatic approach to explanations. *Philosophy of Science* 47:404–423.

Gärdenfors, Peter. 1981. An epistemic approach to conditionals. *American Philosophical Quarterly* 18:203–211.

Gärdenfors, Peter. 1982. Rules for rational changes of belief. In *Philosophical essays dedicated to Lennart Aqvist on his fiftieth birthday*, ed. T. Pauli, 88–101. Department of de Philosophy, Universidad de Uppsala.

Gärdenfors, Peter. 1986. Belief revision and the Ramsey test for conditionals. *The Philosophical Review* 95:81–93.

Gärdenfors, P. 1988. *Knowledge in flux*, 67. Cambridge, MA: MIT.

Gärdenfors, P., and D. Makinson. 1988. Revision of knowledge systems using epistemic entrenchment. In *Proceedings of the Second Conference TARK*, 83–95. California, CA.

Gärdenfors, P., and D. Makinson. 1991. Relation between the logic of theory change and non monotonic logic. In *Lecture Notes in Artificial Intelligence vol. 465*, eds. A. Furman and M. Morreau, 185–205. Berlin: Springer.

Graubard Stephen, R. 1988. *The artificial intelligence debate: False starts and real foundations*. Cambridge, MA: MIT Press

Gärdenfors, Peter, and Rott. Hans. 1992. *Belief revision*. 11. Lundt University Cognitive Studies.

Halpern, J. 1986. Reasoning about knowledge: An overview. In *Proceedings of the First Conference TARK*. Monterey, CA.

Harper, W. 1977. Rational conceptual change. In *Philosophy of science association*, 462–494. University of Chicago.

Katsuno, H., and A. Mendelzon. 1989. A unified view of propositional knowledge base updates. In *Proceedings of the 11th IJCAI,* 1413–1419. Detroit, MI: Morgan Kauffman

Kelly, K., O. Schulk, and V. Hendricks. 1997. Reliable belief revision. In *logic and scientific methods*, 179–208. New York, NY: Kluwer.

Levesque, H.J. 1984. Foundations of a functional approach to knowledge representation. *Artificial Intelligence* 23:155–212.

Levesque, H.J. 1986. Knowledge representation and reasoning. *Annual Review of Computer Science* 1:255–287.

Levesque H.J. 1988. Logic and the complexity of reasoning. *Journal of Philosophical Logic* 17:355–389.

Levi, I. 1977. Subjunctives, dispositions and chances. *Synthese* 34:423–455.

Levi, I. 1980. *The enterprise of knowledge*. Cambridge, MA: MIT.

Makinson, David. 1989. General theory of cumulative inference. In *Non-monotonic reasoning*, eds. Michael Reinfrank, Johan de Kleer, Matthew L. Ginsberg, and Erik Sandewall, 2nd International Workshop, Grassau, FRG, June 13–15, 1988, Proceedings. *Lecture Notes in Computer Science* 346. Berlin: Springer.

Makinson, David. 1996. In *Memoriam*, eds. Carlos Eduardo Alchourrón. *Nordic Journal of Philosophical Logic* 1(1):3–10

Makinson, David, and Kourousias George. 2006. Respecting relevance in belief change. *Análisis Filosófico* XXVI(1). Buenos Aires, Mayo, 53–61

McCarthy, J. 1977. Epistemological problems of artificial intelligence. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, 1038–1044. Cambridge, MA: M.I.T

McCarthy, J., 1980. Circumscription. A form of non-monotonic reasoning. *Artificial Intelligence* 13:41–72

McCarthy, J., and P. J. Hayes. 1969. Some philosophical problems from the standpoint of artificial intelligence. In *Machine intelligence*, eds. B. Meltzer, and D. Michie, 4:463–502. Edinburgh: Edinburgh University Press.

Nebel, Bernhard. 1989. A knowledge level analysis of belief revision. In *Principles of knowledge representation and reasoning*, eds. R.J. Brachman, H.J. Levesque, and R. Reiter, 301–311. San Francisco, CA: Morgan Kaufmann.

Nebel, Bernhard. 1992. Syntax-based approaches to belief revision. In *Belief revision*, ed. P. Gärdenfors. Cambridge Tracts in Theoretical Computer Science 29, New York, NY: Cambridge University Press.

Newell, Alan. 1981. The knowledge level. *The AI Magazine* 2(2):1–20.

Rao, A.S., and N. Foo. 1989. Minimal change and maximal coherence – A basis for belief revision and reasoning about actions. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 966–971. San Mateo: Morgan Kaufmann.

Reiter, Ray. 1980. A logic for default reasoning. *Artificial Intelligence* 13:81–132

Rott, Hans. 1989. Conditionals and theory change: Revision, expansions and additions. *Synthese* 81:91–113.

Satoh, Ken. 1988. Non monotonic reasoning by minimal belief revision. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, 455–462, Tokio, Japan.

Schlechta, Karl. 1989. Some results on belief revision. In *Proceedings of the workshop on the logic of theory change*, Konstanz.

Vardi, M. 1988. Preface for the edition of the *Proceedings of the second conference TARK* (Theoretical Aspects of Reasoning about Knowledge), California, USA.

Weber, A. 1986. Updating propositional formulas. In *Proceedings of the Expert Database Systems Conference,* Charleston, SC.

Winslett, M. 1986. Is belief revision harder than you thought? In *AAAI86 proceedings*, 421–427 August 11–15, Philadelphia, PA.

Winslett, M. 1988. A model-based approach to updating databases with incomplete information. *ACM Transactions on Database Systems* 13(2):167–196.

Chapter 2
Changing the Scientific Corpus

Sven Ove Hansson

2.1 Introduction

There is a straightforward connection in terms of subject-matter between belief revision and one of the major issues in the philosophy of science, namely the dynamics of changes in scientific knowledge. But in spite of this connection, there has been relatively little contact between the two disciplines. There is an obvious reason for this lack of contact: The standard framework that is used in the belief change literature is not suitable for analyzing the mechanisms of change in science. The aim of this contribution is to identify the differences and show what modifications are needed to make the format suitable for modelling the development of scientific knowledge.

Belief revision theory is dominated by an input-assimilating approach (Fig. 2.1). The usual models describe how a person or a computer transforms its state of belief upon receipt of an input or an instruction. Between the inputs, the state of belief is assumed to be constant. (Hansson 1999, pp. 3–11) Of course, this is an idealization. Actual subjects change their minds as a result of deliberations that are not induced by new inputs. It is also important to note that in input-assimilating models, no explicit representation of time is included. Instead, the characteristic mathematical constituent is a function that, to each pair of a state and an input, assigns a new state.

As far as I can see, adequate models of scientific change can be input-assimilating, and thus belong to the same general class of models as those that

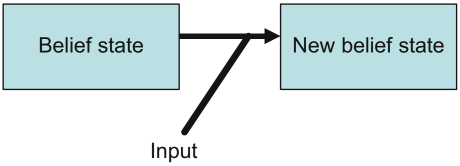


Fig. 2.1 The general structure of input-assimilating models of belief change

S.O. Hansson (✉)
Division of Philosophy, Royal Institute of Technology, Tekniktingen 78,
100 44 Stockholm, Sweden
e-mail: soh@kth.se

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_2,

dominate in belief revision theory. However, some features of the model, including the nature of the inputs, will have to be different.

Belief revision and the philosophy of science need to join forces. For that purpose, mutual adjustments are necessary. I will begin this investigation at the philosophy of science side, and present an informal model of the development of scientific knowledge. This model will be idealized in ways that facilitate contact with belief revision theory. After that I will turn to belief revision theory and propose adjustments on that side that will make fruitful collaborations between the two disciplines possible.

Obviously, the models to be proposed have to be idealized. In other words, they will reflect some aspects of the scientific process, but tone done other aspects. It would not be a realistic aim of this exercise to develop a model that captures all the important features of scientific change. Instead, I propose that we aim at a fit that is at least as good as the fit between AGM theory and changes in individual beliefs. Experience from other fields of research shows that it is often a useful strategy to begin with an oversimplified model to which various complications are then reluctantly added when this is shown to be necessary, rather than trying to capture all the complications already in the initial model.

2.2 The Corpus Model of Scientific Change

To get started we need a simplified model of the development of scientific knowledge that can be used in the adjustment of belief revision theory to this subject-matter. In this section I will propose a fairly simple model of the scientific corpus that can be used for this purpose. It is of course a highly idealized model. (For a somewhat more extensive discussion, that also takes the needs of applied science into account, see Hansson 2007.)

Scientific knowledge begins with data that originate in experiments and other observations.[1] Through a process of critical assessment, these data give rise to the scientific corpus (Fig. 2.2). The corpus consists of those statements about scientific subject-matter that are taken for given by the collective of researchers in their continued research, and thus not questioned unless new data give reason to question them. For practical purposes we can also, roughly, identify the corpus as consisting of those statements that could, at the time being, legitimately be made without reservation in a (sufficiently detailed) textbook (Hansson 1996).

Although the corpus is based on data, it is not a selection of data. Instead its essential components are statements of a more general nature: generalized statements that describe and explain features of the world we live in.[2] These statements

[1] See Section 2.4.2 for a delimitation of data.

[2] As will be clear in Section 2.6, it is a matter of convenience whether the accepted data are treated as elements of the corpus, or whether the corpus consists only of the generalized statements that are based on these data.

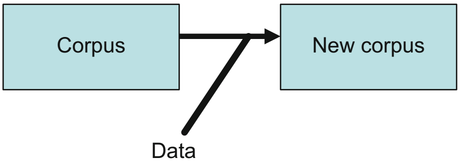


Fig. 2.2 The introduction of new data into the scientific corpus

are of course expressed in a language that mirrors our methods of investigation and the concepts that we have developed. Whereas data refer to what has been observed, the generalized statements in the corpus refer to how things are and to what can be observed.

It follows from this that hypotheses, i.e. “supposition[s] or conjecture[s] put forth to account for known facts” (OED) are not included in the corpus. An hypothesis should only be included when the data provides sufficient evidence for it, and then it ceases to be an hypothesis. It is important to observe that confirmed (or corroborated) hypotheses are not the only generalizations that enter the corpus. Contrary to what is often believed, large parts of modern science are not hypothesis-testing. Scientific investigations always have the purpose of answering some research question, but this does not necessarily mean that they have the purpose of testing an hypothesis. For that to be the case, the research question has to be a yes/no question. (More precisely, the possible outcomes of the research have to be divided beforehand into two mutually exclusive categories, according to how some statement – the hypothesis – fares if the respective outcome is obtained.) Investigations aimed at determining a DNA sequence or the structure of a complex biomolecule are examples of research that is not hypothesis-testing. Such research can instead be called explorative. An empirical study of high-status articles in the natural sciences indicated that a majority of the best research may well be explorative in this sense (Hansson 2006). Therefore, the development of the corpus should not be modelled as driven exclusively by the posing and testing of hypotheses.

The scientific corpus is a highly complex construction. Due to its sheer size, it cannot be mastered by a single person. Different parts of the corpus are maintained by different groups of scientific experts. The areas of expertise are overlapping in complex ways, and the division of the corpus into such areas changes over time. Furthermore, the various parts of the corpus, as defined by the areas of expertise, are all constantly in development. New statements are added, and old ones removed, in each of the many subdisciplines. Often the changes concern more than one area of expertise, and therefore a consolidating process based on contacts and cooperations between interconnected disciplines takes place continuously.

In spite of this the corpus is, at each point in time, reasonably well-defined. In most disciplines it is fairly easy to distinguish those statements that are, for the time being, generally accepted by the relevant experts from those that are contested, under investigation, or rejected. Hence, the vague margins of the corpus are fairly narrow.

The process that leads to modifications of the corpus is based on strict standards of evidence. These standards are an essential part of the ethos of science. The onus of proof falls to those who want to change the corpus – for instance by acknowledging a previously unproven phenomenon, or introducing a new scientific theory. Another way to express this is to say that the corpus has high entry requirements. This is necessary to ensure a reasonably steady progress in science. If we accept new ideas too rashly, then scientific progress can be blocked by mistaken assumptions. On the other hand there are limits to how high the entry requirements can be. Since we cannot leave everything open we will have to take some risks of being wrong.

The entry requirements of the corpus can be described in terms of how we weigh the disadvantages for future research of unnecessarily leaving a question unsettled against those of settling it incorrectly. This is closely related to the values we assign to truth and to avoidance of error. In addition, our decisions on corpus inclusion can be influenced by other epistemic values that concern usefulness in future science, such as the simplicity and the explanatory power of a theory (Hempel 1960, Feleppa 1981).

Whereas epistemic values have an obvious role in determining what we allow into the corpus, influence from non-epistemic values is programmatically excluded. According to the ethos of science, what is included in the corpus should not depend on how we would like things to be but on what we have evidence for. Therefore, it is part of every scientist's training to leave out non-epistemic values from her scientific deliberations as far as possible. This, of course, is not perfectly achieved. As was noted by Ziman, we researchers all have interests and values that we try to promote in our scientific work, "however hard we try to surpass them". But as he also noted, "the essence of the academic ethos is that it defines a culture designed to keep them as far as possible under control" (Ziman 1996, p. 72).

Probably, the largest deviations from this ideal concern non-controversial moral values, i.e. values that are shared by virtually everyone or by everyone who takes part in a particular discourse. A typical example of this is the influence in medical research of the aims of medicine and of some almost universally accepted principles of medical ethics. In other areas, such as economics, the presence of non-epistemic values in the corpus may be more controversial. For our present purposes we can, at least initially, leave open whether the values that determine corpus inclusion are strictly epistemic or whether they also include non-epistemic values. For many purposes, the same formal structures can be used to capture both types of values.

2.3 Alternatives to the Corpus Model

At least two alternative accounts should be considered before we accept this picture of scientific change. One of these approaches is to replace the single corpus by several corpora, perhaps one for each discipline. The other is to replace the sharp limit between elements and non-elements of the corpus by a system with multiple degrees of acceptance, such that most scientific statements have some intermediate degree of acceptance between the highest degree and outright rejection. We can call these alternatives "multiple corpora" respectively "vague corpus".

Beginning with *multiple corpora*, this idea assumes that the different disciplines are largely independent of each other. Arguably this was the case in the early days of science, but today the interdisciplinary interconnections in science are strong and rapidly strengthening. In the last half century or so, integrative disciplines such as astrophysics, evolutionary biology, biochemistry, ecology, quantum chemistry, the neurosciences, social psychology, and game theory have developed at dramatic speed and contributed to tying together previously unconnected disciplines. Through numerous such ties of shared and interdependent knowledge, a large number of disciplines are connected to each other, directly or indirectly. The resulting community of interdependent disciplines includes not only those academic disciplines that are covered by the restrictive English term "science" but also the wider range of disciplines that are covered by the German term "Wissenschaft". The role of natural science in modern archaeology exemplifies this.

Somewhat paradoxically, belief in the coherence of science seems to have been much stronger in the first half of the 20th century than what it is today. Although the reductive account of relations between the disciplines that was popular at that time is not tenable, it is remarkable that interdisciplinary interdependence has increased dramatically in science at the same time as belief in it seems to have receded.

Hence, in actual practice science operates with a common corpus with increasingly strong interconnections. It would of course in principle be possible to claim that nevertheless, in our models of scientific change the corpus should be broken up into parts with less interdependence. However, it is difficult to see how, on any reasonable account of the aims of science, such an approach could be justified. The alternative with multiple corpora does not seem plausible.

We can therefore turn to the other alternative account, that of a *vague corpus*. In the model that I proposed above, scientific statements are subject to a binary classification, according to whether they accepted into the corpus or not.[3] There is an obvious way to replace this classification with many degrees of acceptance: We can apply Bayesian decision theory (Jeffrey 1956). According to the Bayesian ideal of rationality, all statements about the world should have a definite probability value assigned to them. Contingent propositions should never be fully believed, but can be assigned high non-unit probabilities. The resulting belief system is a complex web of interconnected probability statements.

There is a prominent feature of actual human belief systems that Bayesian models do not take into account: the cognitive limitations of human beings. These limitations do in fact severely restrict our probabilistic reasoning. In order to arrive at a manageable belief system we have to "fix" a large amount of our beliefs to (provisional) certainty, and take as true (false) much of that to which we would otherwise only assign a high non-unit (low non-zero) probability.[4] As an example of this,

[3] Of course, both accepted and non-accepted statements are often expressed in probabilistic terms.

[4] On the other hand, we also regard many issues as unsettled or uncertain, but do not assign definite probabilities or degrees of belief to them. Thus, whereas a (hypothetical) Bayesian subject assigns probabilities distinct from 0 and 1 to all contingent factual statements, actual subjects have very

I fully believe in the conjugation pattern for the French verb “être” that I learnt at school. This belief has been confirmed in various encounters with native speakers and writers of the language. If I were a Bayesian I would only assign a high probability to the correctness of this pattern.

The transformation of high probabilities to full belief can be described as a process of uncertainty-reduction, or “fixation of belief” (Peirce 1934). It helps us to achieve a cognitively manageable representation of the world, thus increasing our competence and efficiency as decision-makers. This transformation is just as necessary in the collective processes of science as it is in individual cognitive processes. In science as well, our cognitive limitations make it impossible to keep track of an extensive net of interconnected probabilities. We cannot (individually or collectively) deal with a large body of human beliefs such as the scientific corpus in the massively open-ended manner that an ideal Bayesian subject would be capable of. As one example of this, since all measurement practices are theory-laden, no reasonably simple account of measurement would be available in a Bayesian approach (McLaughlin 1970).

In summary, neither of these two major alternatives to the approach of a distinct, unified corpus seems promising.

2.4 Major Differences That Need to Be Taken into Account

Traditionally, belief revision theory has two major application areas: changes in the beliefs of a single individual and in a computerized database. At least five major differences between these areas and the scientific knowledge process have to be taken into account in the construction of a formal model of scientific change.

2.4.1 The Processes of Change Are Collective

Although some studies have been made on multi-person processes in belief revision, the main focus is on single-agent processes. In contrast, the processes of scientific change are essentially collective in at least two important ways. First, decisions to incorporate a new standpoint in the corpus are not made by a single individual. Instead, this is a fairly complex collective process – an informal decision process that proceeds by consensus or near-consensus among the respected experts. Secondly, there is a division of labour, such that different experts are involved in modifying different parts of the corpus. Nobody masters the whole of science. Instead, we have a complex division of expertise between partly overlapping groups of experts.

few such probabilistic beliefs but instead (i) judgments held to be true or false, and (ii) judgments that are unsettled but to which no exact numerical probability has been assigned.

2.4.2 The Data/Theory Division

In the idealized model of data-driven scientific change presented above, the distinction between data and theory is essential. The inputs into the system are data. However, it is not easy to draw the line between data and theory. As a first approximation, data are observation reports. Observation reports are often made in theory-dependent terms, but they can at least in principle be reformulated in theory-independent terms. In this sense, a report that a measurement of the temperature of a particular liquid yielded 29°C can be treated as data. We can also have another report, saying that according to another measurement its temperature, at the same time, was 30°C. We can treat these two reports as two different data that do not contradict each other. It is a fact that one thermometer measurement yielded 29°C and it is also a fact that another such measurement yielded 30°C. Neither of these cancels the other out. On the other hand, a statement that the temperature of the liquid *was* 29°C goes beyond the observations, and can be invalidated by additional information.

For the present purposes, it will be useful to draw the limit between data and non-data at a sufficiently low level of theory-ladenness to make data mutually non-contradictory. This means that the measurement reports just referred to are data, whereas a statement that the temperature in question was 29°C is taken to be too theory-laden to be treated as data.

2.4.3 A Partly Accumulative Process

The incorporation of new observations in the form of data into the scientific corpus is largely an accumulative process, in the sense that data are added but seldom retracted. The most clear cases of retraction are those in which an observation report turns out to be incorrect. In the construction of a first simplified model of scientific change, we can abstract from such cases. Given this, and given the view of data just proposed, we can model the process of data acquisition as accumulative, i.e. data are added but they are not retracted.

Clearly, accumulativity does not apply to theoretical statements. The acquisition of new data can induce us to give up previous theoretical beliefs. Therefore, we can describe the scientific knowledge process as (only) partly accumulative.

2.4.4 Explanation-Management Rather Than Inconsistency-Management

In a model based on the principles introduced above, no contradictions among data will arise. New data may contradict previously formed theoretical beliefs, but in this idealized approach they will not contradict previous data. The crucial issue, given the accumulated data, is how these data can best be theoretically accounted for, i.e. explained. This should lead us away from the traditional focus in belief revision, which is inconsistency-management, to the broader issue of explanation-management, i.e. search for the best explanation of the available data. Since inconsistencies are bad explanations, the avoidance of inconsistencies can be subsumed under the search for explanations, but the converse subsumption does not hold.

2.4.5 The Irrelevance of Contraction

One of the major types of operation in belief revision theory is that of contraction. Indeed, in much of the formal work, the main focus has been on this operation. Contractions are always contractions by a specified statement. The outcome of a contraction by a sentence p is a new belief state in which this sentence p is not believed. (Unless p is a tautology, in which case it cannot, *per impossibile*, be retracted.) Hence this is an operation in which old beliefs are deleted from the belief state but nothing new is added to it. It is difficult, however, to find examples in real life of such pure contraction, in which no new belief is added. When we give up a belief, this is typically because we have learnt something new that forces the old belief out. (Hansson 1999, pp. 63–65) I once believed that Plato wrote the *Hippias Major*. Then I learnt that the authenticity of this dialogue is contested among specialist scholars. I have therefore ceased to believe that Plato wrote the *Hippias Major* (without starting to believe in the negation of that statement). Strictly speaking, this is not a case of (pure) contraction, since a new belief was acquired to the effect that the authorship of this work is uncertain. In the literature on belief dynamics, examples such as this are often interpreted as referring to (pure) contraction. The new belief that gave rise to contraction is neglected, and is not included in the new belief set. This is an imprecise but convenient convention, that makes it much easier to find examples of contraction.

Some authors prefer to use hypothetical contractions as examples of pure contraction. We sometimes hypothetically give up a belief in order to give a contradictory belief a hearing. Such hypothetical contractions, or contractions for the sake of argument, can unproblematically be constructed as pure contractions (Fuhrmann 1991, Fuhrmann and Hansson 1994, Levi 1991). However, the use of these operations as examples of pure contraction is questionable since they are not seriously undertaken changes in the actual belief state of the agent.

While it is difficult to make sense of pure contraction in the dynamics of human belief, it is much easier to do so in applications to computerized databases. We often have good reasons to instruct a computer to remove an item from its database. (Against the background of this difference it is strange that philosophers contributing to the belief revision literature have put much more emphasis on contraction than computer scientists, who tend to put the focus instead on the operation of revision.)

In the dynamics of scientific knowledge, the role of pure contraction is even more dubious than in the dynamics of individual human belief. We expect changes in science to be driven by new empirical data. When old scientific beliefs are given up – be they empirical or theoretical beliefs – this is not due to an external order as when a computer receives instructions to remove an item from its database. Instead it is (at least on the idealized level on which we need to describe science for the present purposes) a reaction to new data that lead us to exclude old beliefs. Therefore when building a model of scientific change we do not seem to have much use for an operation of (pure) contraction.

2.5 Incorporation or Retrieval

As was indicated in Section 2.1, input-assimilating models can be used to account for scientific change, but these will have to be a different type of input-assimilating models than those that are common in the belief revision literature. We can now delineate the major differences.

In the presence of conflicting information or when new information decreases the plausibility of our explanations, selections are necessary. We then have a choice between (1) making these selections as part of the operations of change when new information is received, and (2) letting operations of change leave conflicts unresolved, and instead make the necessary selections when information is retrieved from the system (Rott 2001).

There is a trade-off in simplicity between retrieval and change. In the AGM model, the retrieval operation is as simple as possible – it is just the identity operation. The change operations of AGM are much more complex. In belief base models we have a somewhat more complex retrieval mechanism, namely a consequence operator. (If the belief base is B, then the set of beliefs to which the agent is committed is $\mathrm{Cn}(B)$.) On the other hand, operations of belief change tend to be somewhat less complex in belief base models than in belief set models. (The expansion of a belief base B by a sentence p is equal to the set-theoretical union $B \cup \{p\}$. The expansion of a belief set K by a sentence p is equal to $\mathrm{Cn}(K \cup \{p\})$.) We can go further than this, and transfer much more from the operations triggered by the receipt of new information to the operations triggered by the retrieval of information. The crucial step is of course to move the selection mechanism from the receipt to the retrieval part of the model. Due to the central role of the accumulation of empirical data in the development of scientific knowledge, this seems to be an option well worth investigating in the search for a model of the scientific knowledge process.

2.6 The Building-Blocks of a Retrieval Model

As already indicated, we need a language $\mathcal{L}$ in which data can be recognized, and distinguished from more theory-laden or outright theoretical statements. A simple way to deal with this in a formal model is to introduce a function $|\,|$ such that for any set A of sentences, $|A|$ is the set consisting of exactly those elements of A that represent data (and nothing but data). Clearly, $|\mathcal{L}|$ is the data-representing fraction of the language, and all inputs consist in the addition of an element of $|\mathcal{L}|$ to the corpus or to the set of data. Given the delimitation of data introduced in Section 2.4.2, $|\mathcal{L}|$ should be logically consistent, so that no combinations of data are (in themselves) inconsistent.

We will need two operations: one for the incorporation of new data and one for developing theory, based on data. (The latter is a retrieval operation in the terminology of Section 2.5.) For the incorporation of data, the standard operation of expansion can be used. It has two variants. As noted above, the expansion of a logically closed set K by a sentence p is equal to the closed set $\mathrm{Cn}(K \cup \{p\})$, whereas the corresponding expansion of a set B that is not logically closed is equal to the set-theoretical union $B \cup \{p\}$. The symbol + is traditionally used for both these operations.

The other operation will be used to draw conclusions, or more precisely: make inferences to the best explanation from a set of data. Given the set B of data we aim at finding the best theoretical structure to account for B. The outcome of these deliberations is a set $\mathrm{C}(B)$ of statements that includes both B and the theoretical statements that have been chosen to account for it. C is an operation of inference to the best explanation. (Alternatively, we could construct $\mathrm{C}(B)$ so that it does not contain B. However, this will give rise to a less straight-forward formal structure, without any real gain in conceptual clarity.)

We will also need to apply the inference relation C to mixed sets, i.e. sets that include both data and theoretical statements (namely previously made inferences). Suppose that after making inferences (to the best explanation) based on the set B of data, obtaining $\mathrm{C}(B)$, we add a new piece of data, p, obtaining $\mathrm{C}(B) + p$. We can then again apply the inference relation, obtaining $\mathrm{C}(\mathrm{C}(B + p))$.

C should satisfy the property

$$|\mathrm{C}(B)| = |B| \text{ (data identity),}$$

i.e. C neither adds nor retracts data.[5] It follows from this that:

$$|B| \subseteq \mathrm{C}(|B|) \text{ (data inclusion)}^{6}$$

[5] What makes this plausible is the restricted delimitation of data that was introduced in Section 2.4.2. Theoretical deliberations can lead us to acquire new beliefs about matters of fact. Hence, given the observation of several thrust nightingales in Saudi Arabia that were ringed in Sweden, we may add the statement "thrust nightingales migrate from Sweden to Saudi Arabia" to our state of belief. However, this statement of fact does not qualify as data on the present account. (Reports about the finding of these ringed birds will, however, qualify as data.)

[6] Proof: $|B| = \||B|\|$
$= |\mathrm{C}(|B|)|$ (data identity)
$\subseteq \mathrm{C}(|B|)$

However, C should *not* satisfy

$$B \subseteq \mathrm{C}(B)\,\text{(inclusion)}$$

This can be seen from a case when $B = \mathrm{C}(A) + p$, and p is some data that gives us reason to revise some of the inferences from A that were drawn in $\mathrm{C}(A)$.

Similarly, it follows from data identity that:

$$\text{If } |A| \subseteq |B| \text{ then } |\mathrm{C}(A)| \subseteq |\mathrm{C}(B)| \text{ (data montonicity)}$$

However,

$$\text{If}\, A \subseteq B \,\text{then}\, \mathrm{C}(A) \subseteq \mathrm{C}(B) \text{ (monotonicity)}$$

does not hold in general, or even if $A = |A|$ and $B = |B|$. The reason for this is that $B \backslash A$ may contain data that give us reason to reject some of the inferences in $\mathrm{C}(A)$.

The following condition:

$$\mathrm{C}(\mathrm{C}(A)) = \mathrm{C}(A)\,\text{(iteration)}$$

says that the C operator is complete in the sense of drawing all the inferences that can be drawn in one single application. This is a reasonable condition.[7] The same applies to

$$\mathrm{C}(A) \nvdash \perp \text{(inferential consistency)}$$

that requires that inference to the best explanation does not lead us to inconsistency.

The following alternative notation for C:

$$A© = \mathrm{C}(A)$$

has the advantage that sequences of expansions and inference operations will be written in the order in which they are performed:

$$\mathrm{C}(\mathrm{C}(B) + p) = B© + p©$$

2.7 Construction of the Model: Three Alternatives

There are at least three ways in which we can combine the operations + and C in a model of the scientific knowledge process.

[7] But see the further comment on it in Section 2.8.

2.7.1 First Construction: Total Independence Between the Two Operations

First and perhaps simplest, we can treat the inference operation (C) as completely independent of the operation of incorporating new data (+). This means that C leaves no trace with any effect after new data have been incorporated, or in formal language:

$$\mathrm{C}(B) + p = B + p$$

A sequence of operations can then look like the following:

$$\begin{array}{l} B \\ B + p_1 \\ B + p_1 + p_2 \\ B + p_1 + p_2 + p_3 \\ B + p_1 + p_2 + p_3 + p_4 \\ \mathrm{C}(B + p_1 + p_2 + p_3 + p_4) \\ B + p_1 + p_2 + p_3 + p_4 + p_5 \\ B + p_1 + p_2 + p_3 + p_4 + p_5 + p_6 \\ \mathrm{C}(B + p_1 + p_2 + p_3 + p_4 + p_5 + p_6) \\ \dots \end{array}$$

Notice here that in the seventh step, after the inference operation has taken us from $B+p_1+p_2+p_3+p_4$ to $\mathrm{C}(B+p_1+p_2+p_3+p_4)$, the next incorporation, in which p_5 is added, takes us to $B+p_1+p_2+p_3+p_4+p_5$, not to $\mathrm{C}(B+p_1+p_2+p_3+p_4)+p_5$. This model satisfies path independence since it makes no difference in what order different data are received.

This approach corresponds to how retrieval works in the usual models in belief revision, (both in belief set models and belief base models). However, in the context of scientific knowledge this is a highly unrealistic structure. In real science, the corpus does not disappear when new data are added so that theoretical deliberations have to start from scratch. To the contrary, the current corpus is the starting-point for any development of new theory. Therefore, this approach is too simplistic.

2.7.2 Second Construction: Inference After Every Incorporation

Another option, that in a sense goes to the other extreme, is to let each acquisition of new data be followed automatically by an adjustment of the corpus that is retained as a starting-point for the next operation on the corpus. This would give rise to series of operations such as the following:

$$
\begin{array}{l}
B\\
\mathrm{C}(B+p_1)\\
\mathrm{C}(\mathrm{C}(B+p_1)+p_2)\\
\mathrm{C}(\mathrm{C}(\mathrm{C}(B+p_1)+p_2)+p_3)\\
\mathrm{C}(\mathrm{C}(\mathrm{C}(\mathrm{C}(B+p_1)+p_2)+p_3)+p_4)\\
\mathrm{C}(\mathrm{C}(\mathrm{C}(\mathrm{C}(\mathrm{C}(B+p_1)+p_2)+p_3)+p_4)+p_5)\\
\mathrm{C}(\mathrm{C}(\mathrm{C}(\mathrm{C}(\mathrm{C}(\mathrm{C}(B+p_1)+p_2)+p_3)+p_4)+p_5)+p_6)\\
\mathrm{C}(\mathrm{C}(\mathrm{C}(\mathrm{C}(\mathrm{C}(\mathrm{C}(\mathrm{C}(B+p_1)+p_2)+p_3)+p_4)+p_5)+p_6)+p_7)\\
\ldots
\end{array}
$$

In this model, we cannot expect *path-independence* to hold, i.e. it can make a difference in the corpus in which order the data is received. However, the assumption that every acquisition of new data is followed by an adjustment of the corpus is highly unrealistic. In this model, expansion (+) and inference (C) have in fact been combined into a single operation of the same type as the previously studied operations of semi-revision or non-prioritized belief revision (Hansson 1997, Hansson et al. 2001). However, this combined operation does not at all correspond to the dynamics of science.

2.7.3 Third Construction: Non-automatic but Retained Inferences

This brings us to the third model, that is intermediate between the two previous ones. Its major features are:

(1) As in the first construction, the inference operation is not performed automatically after each acquisition of new data. Instead the two operations are initiated independently of each other.
(2) As in the second construction, the outcome of an inference (theory development) is retained, and used as a starting-point for further inferences.

A sequence of operations can then develop as follows:

$$
\begin{array}{l}
B\\
B+p_1\\
B+p_1+p_2\\
B+p_1+p_2+p_3\\
B+p_1+p_2+p_3+p_4\\
\mathrm{C}(B+p_1+p_2+p_3+p_4)\\
\mathrm{C}(B+p_1+p_2+p_3+p_4)+p_5\\
\mathrm{C}(B+p_1+p_2+p_3+p_4)+p_5+p_6\\
\ldots
\end{array}
$$

With the alternative notation introduced at the end of Section 2.6, this sequence can also be written as follows:

$$
\begin{array}{l}
B \\
B + p_1 \\
B + p_1 + p_2 \\
B + p_1 + p_2 + p_3 \\
B + p_1 + p_2 + p_3 + p_4 \\
B + p_1 + p_2 + p_3 + p_4 \copyright \\
B + p_1 + p_2 + p_3 + p_4 \copyright\ + p_5 \\
B + p_1 + p_2 + p_3 + p_4 \copyright\ + p_5 + p_6 \\
\ldots
\end{array}
$$

In this model, the outcomes of previous inferences are retained, and they can therefore influence how later inferences are made. Therefore, path independence is not to be expected, i.e. it can make a difference in which order data are received.

It could perhaps be argued that path independence is a desirable property. Why should the order in which different informations are received have an influence on the inferences drawn from them? If the scientific corpus were managed by beings with unlimited cognitive capacity, there would be no good reason to accept such an influence. However, that is not at all the case. Science is a human activity that is deeply affected by the cognitive abilities and inabilities that human beings have (individually and collectively). Contrary to beings with unlimited cognitive capacity, who could afford to start from scratch whenever new information was received, scientists have to economize with their resources for theoretical work. It is for this reason that path independence is an unrealistic and consequently undesirable property of a model of scientific change (cf. Hansson 2010).

The third model is much more realistic than the two previous ones, and it is therefore proposed as a starting-point in the modelling of the scientific knowledge process.

2.8 Further Developments

The construction of an operator C of abduction (inference to the best explanation) that can fill the role in a model of scientific change that has been outlined above should be a major task in the further development of models of scientific change. (For some approaches to abduction, see Pagnucco 1996, Aliseda 2006, Páez 2006, and Schurz 2008.) The highly simplified account introduced here needs to be improved in several ways.

One of the unrealistic features of the models introduced above is that C is assumed to be complete in the sense that it finishes the process of inference to the best explanation, as far as it can be performed on the basis of the available data.[8] In actual practice, theory-development is piecemeal, and affects only parts of the scientific corpus at a time. This feature can be modelled with operations of local

[8] This completeness property is encoded in the iteration property that was introduced in Section 2.6.

change, i.e. operations in which (a restricted version of) the inferential operator C is only applied to a part of the corpus (cf. Hansson and Wassermann 2002).

A major problem for this type of model is how to represent the creative nature of theory development. The function C gives the impression of a deterministic process. In real science, theory development is of course limited by the empirical data but it is not determined by them. The introduction of new theoretical concepts, or new modes of explanation (such as once action at a distance) is particularly difficult to account for in this (or any) belief revision format. The use of an indeterministic inference operator may be a small but important step in the right direction. (See Lindström and Rabinowicz 1991 for an example of an indeterministic operator in a belief revision context.)

One interesting approach that may contribute to the representation of the creative process is to extend the framework so that it contains, in addition to a model of the corpus as explained above, models of the belief systems of individual scientists who contribute to the development of the corpus. The corpus can then be described as the outcome of the interdependent belief changes of these interacting individuals.

The main conclusion from this investigation is that a model of the scientific knowledge process can be an input-assimilating model just like the common models in belief revision, but it must be a different type of input-assimilating model. It has to make a clear difference between data and generalized statements, and its focus should be on search for explanations rather than (mere) avoidance of inconsistencies. It has little or no use for operations of belief contraction. Instead, its major operations should be (i) cumulative addition of new data and (ii) abductive inference, i.e. inference to the best explanation.

References

Aliseda, A. 2006. *Abductive reasoning*. Dordrecht: Springer.

Feleppa, R. 1981. Epistemic utility and theory acceptance: Comments on Hempel. *Synthese* 46:413–420.

Fuhrmann, A. 1991. Theory contraction through base contraction. *Journal of Philosophical Logic* 20:175–203.

Fuhrmann, A., and S.O. Hansson. 1994. A survey of multiple contraction. *Journal of Logic, Language and Information* 3:39–76.

Hansson, S.O. 1996. What is philosophy of risk? *Theoria* 62:169–186.

Hansson, S.O. 1997. Semi-revision. *Journal of Applied Non-Classical Logic* 7:151–175.

Hansson, S.O. 1999. *A textbook of belief dynamics. Theory change and database updating*. Dordrecht: Kluwer.

Hansson, S.O. 2006. Falsificationism falsified. *Foundations of Science* 11:275–286.

Hansson, S.O. 2007. Values in pure and applied science. *Foundations of Science* 12:257–268.

Hansson, S.O. 2010. Multiple and iterated contraction reduced to single-step single-sentence contraction. *Synthese* 173:153–177.

Hansson, S.O., and R. Wassermann. 2002. Local change. *Studia Logica* 70:49–76.

Hansson, S.O., E. Fermé, J. Cantwell, and M. Falappa. 2001. Credibility-limited revision. *Journal of Symbolic Logic* 66:1581–1596.

Hempel, C.G. 1960. Inductive inconsistencies. *Synthese* 12:439–469.

Jeffrey, R.C. 1956. Valuation and acceptance of scientific hypotheses. *Philosophy of Science* 23:237–249.

Levi, I. 1991. *The fixation of belief and its undoing*. Cambridge, MA: Cambridge University Press.

Lindström, S., and W. Rabinowicz. 1991. Epistemic entrenchment with incomparabilities and relational belief revision. In *The logic of theory change*, eds. A. Fuhrmann and M. Morreau, 93–126. Berlin: Springer.

McLaughlin, A. 1970. Science, reason and value. *Theory and Decision* 1:121–137.

Páez, A. 2006. The epistemic value of explanation. http://philsci-archive.pitt.edu/archive/00003081.

Pagnucco, M. 1996. *The role of abductive reasoning within the process of belief revision*, PhD Thesis, Department of Computer Science, University of Sydney.

Peirce, C. 1934. The fixation of belief. In *Collected papers of Charles Sanders Peirce*, eds. C. Hartshorne and P. Weiss, vol. 5, 223–247. Cambridge: Harvard University Press.

Rott, H. 2001. *Change, choice and inference*. Oxford: Oxford University Press.

Schurz, G. 2008. Patterns of abduction. *Synthese* 164:201–234.

Ziman, J. 1996. ‘Postacademic science’: Constructing knowledge with networks and norms. *Science Studies* 9:67–80.

Chapter 3
Idealizations, Intertheory Explanations and Conditionals

Hans Rott

3.1 Lessons from Lakatos

Imre Lakatos, according to Paul Feyerabend (1975, p. 1) “an outstanding thinker and the best philosopher of science of [his] strange and uncomfortable century”, combined a profound knowledge of the history of science with a desire to isolate general structures behind the development of scientific theories. Lakatos considered the links between various versions of a theory, or between different theories about the same domain of phenomena, as being provided by intertheory relations.[1] A toolkit of intertheory relations was thought to be the means that accounts for the dialectics between continuous progress and disruptive changes in scientific research.

Let us briefly recount Lakatos’s (1970) story of falsificationism as a story about what relations should hold between rivalling or successive theories. *Dogmatic falsificationism* is a position that argues against “justificationism”. Assume that a theory T implies an observation sentence A, and that an actual observation tells us that not A ($\sim A$). The advice of a dogmatic falsificationist would be: Give up your theory T, find another, new, maximally “bold” theory T'. No trace of T is left, but $\sim A$ is kept as part of the lore of empirical evidence. It is too easy to criticise this position. First, one can argue that there is no pure observation. Rather, an observational theory T_0 is needed, plus auxiliary hypotheses, initial conditions and ceteris paribus clauses in order to obtain observation sentences. There is no infallible empirical basis, and every candidate observation like $\sim A$ above might be false. Dogmatic falsificationism is not an accurate description of the history of science, it was perhaps not held by anybody in the philosophy of science.

According to *naive (methodological) falsificationism*, falsification involves a number of decisions. They concern the following questions: What are observation sentences? Which of these should be accepted as true? Which sentences describing

H. Rott (✉)
Department of Philosophy, University of Regensburg, 93040 Regensburg, Germany
e-mail: hans.rott@psk.uni-r.de

[1] For this idea, compare Krüger (1980). It has also been very influential in the Sneed-Stegmüller “structuralist” approach to the philosophy of science (see Stegmüller 1979).

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_3,

initial conditions and ceteris paribus clauses should be accepted? Which theories are supported to what degree? The general decision of paramount importance, however is this: Which sentences are to be treated as problematic and which as unproblematic (in Lakatos's metaphorical terms: what is "the nut" and what are "hammer and anvil")? The confrontation is no longer an unmediated one between a theory and "the world", but one between various linguistic entities. Inconsistency rather than falsity is the issue.[2] But naive falsificationism can again be criticized. First (this is not Lakatos's own concern), it is plausible to recognize more fine-grained distinctions of priority than just "problematic" and "unproblematic", distinctions that could be used in a systematic way for direct revisions of theories in the face of conflicting evidence. If scientists didn't have ideas how to make finer distinctions, they would not know how to make the decisions needed for the resolution of logical conflicts. Second, and more importantly, naive falsificationism is not an accurate description of the history of science. In Lakatos's picture, real science is less ad hoc and more creative than envisaged by the naive falsificationist. Conflicts in science are typically not confrontations between (linguistically represented) experience and theory (more accurately: a system of theories). They rather have the form of a triple competition: Experience confronts (at least) two competing theories (systems of theories). There is no instant falsification of any theory independent of the existence of a rival theory. But then, in the presence of rival theories, prima facie confirmation becomes more important than falsification

Let us now turn to *sophisticated falsificationism*, which is quite close to Lakatos's own theory. Scientific progress in a single theory transition from a scientific theory T to another one, T', is captured by the following definition (compare Lakatos 1970, p. 116)

Definition 1 A scientific theory T is *falsified* if and only if another theory T' has been proposed with the following characteristics:

[2] As an aside, it may be worth mentioning that in some sense, it is difficult to have *truth as the goal of inquiry*. The reason is that large-scale revisions by truths are bound to lose other truths (Rott 2000). More exactly, let us suppose that a scientist possesses ideal logical competence in the sense that all her theories are logically closed, but that she is not empirically omniscient (her beliefs do not encompass the whole truth). Then there is no belief-contravening revision by some (true) sentence that strictly increases the set of her truths, even if the sentence is true and the revision leads from a false to a true theory. *Proof* Let W be the set of all truths (a maximal consistent set of sentences), and let T be our prior theory. Assume that A is true. (If A is false, the claim is immediate.) As the transition from T to $T{*}A$ is supposed to be belief-contravening, we know that T implies ~A , so ~A is in T, by logical closure. Apply the hypothesis of empirical non-omniscience to $T{*}A$, i.e. assume that there is a C such that C is in W and C is not in $T{*}A$. Now consider $\sim A \vee C$. The disjunction $\sim A \vee C$ is true, since C is in W, and it is in T, since ~A is in T and T is logically closed. But $\sim A \vee C$ is not in $T{*}A$. This is because A is in $T{*}A$ (the revision by A is supposed to be "successful"), and thus, if $\sim A \vee C$ were in $T{*}A$, this theory $T{*}A$ would include C as well, by closure. But our hypothesis was that C is not in $T{*}A$, so we have found a contradiction, and this completes the proof. – This result has some similarities with Miller's (1974) and Tichý's (1974) trivializations of Popper's early concept of verisimilitude.

(1) T' has excess empirical content over T: that is, it predicts *novel* facts, that is, facts improbable in the light of, or even forbidden, by T;
(2) T' explains the previous success of T, that is, all the unrefuted content of T is included (within the limits of observational error) in the content of T'; and
(3) some of the excess content of T' is corroborated.

It is quite evident that the term "falsification" is not really suitable here any more. In this famous paper of 1970, Lakatos still conveys the impression that his theory is a direct outgrowth of Popperian thinking, but as a matter of fact, many of the central tenets of Kuhn and Feyerabend have crept in (regarding "falsification", for instance, the requirement that there be rival theories). Lakatos goes on to say that a single theory transition cannot be the right unit of scientific inquiry. He suggests that a finite sequence of successive theories should take over this role. A sequence of theories $T_1,\ T_2,\ T_{3,}\ldots$ where each T_{i+1} results from adding auxiliary clauses to (or semantic reinterpretations of) T_i in order to accommodate some (Kuhnian) anomaly, and T_{i+1} has at least as much content as the unrefuted content of T_i is called a *problem shift*. Here is another crucial definition (compare Lakatos 1970, p. 118):

Definition 2 A problem shift is

(1) *theoretically progressive* (*or "constitutes a theoretically progressive problemshift"*) if each new theory has some excess empirical content over its predecessor, that is, if it predicts some novel, hitherto unexpected fact.
A theoretically progressive problem shift is also
(2) *empirically progressive* (*or "constitutes an empirically progressive problemshift"*) if some of this excess empirical content is also corroborated, that is, if each new theory leads us to the actual discovery of some *new fact*.

We saw that the naive falsificationist can decide to establish a rich priority structuring of her theory or theories. What happens to this idea in the model of the sophisticated falsificationist? Lakatos has the following recommendation:

> . . . one had to try to replace first one [theory], then the other, then possibly both, and opt for that new set-up which provides the biggest increase in corroborated content, which provides the most progressive problemshift.[3]

So not anything goes in the sophisticated falsificationist's pluralistic proliferation model. But Lakatos goes one step farther. In order to avoid patched-up patterns of isolated hypotheses and to account for the continuities of normal science, he introduces his *methodology of scientific research programmes:*

Definition 3 The basic unit of appraisal must be not an isolated theory or conjunction of theories but rather a "*research programme*", with a conventionally accepted

[3] Lakatos (1970, p. 130). For further elaboration of this idea, see Lakatos (1970, pp. 121–122, 129, 155).

(and thus by provisional decision "irrefutable") "*hard core*" and with a "*positive heuristic*" which defines problems, outlines the construction of a belt of auxiliary hypotheses, foresees anomalies and turns them victoriously into examples, all according to a preconceived plan. (Lakatos, 1971, p. 99)

Lakatos's idea here is to outlaw *ad hoc* manoeuvres in science. The first element to achieve this is the *hard core* or *negative heuristics* T_{h}. It is that part of a scientific research programme that receives maximum priority. Lakatos's implicit assumption is here that T_{h} is irrefutable by methodological decision, and that scientists can see to it that T_{h} is contained in all revisions of the present instalment of the program. The second element is the *positive heuristics* or *protective belt* T_{p} of auxiliary hypotheses around T_{h} consisting in a "partially articulated set of suggestions or hints on how to change, develop the 'refutable variants' of the research programme, how to modify, sophisticate, the 'refutable' protective belt." (1970, p. 135) This description is reminiscent of the priority structure that the naive falsificationist suggested to use for revisions of scientific theories in response to new "data" contradicting an old ensemble of theories. The suggestion of the present chapter will be to apply such priorities not in direct, "forward-oriented" theory revisions, but rather in "backward-oriented" revisions that are useful in scientific model building and intertheory explanation.

But let us flesh out Lakatos's picture first. Here is an important passage from Lakatos (1970, p. 136):

Definition 4 A "*model*" is a set of initial conditions (possibly together with some of the observational theories) which one knows is *bound* to be replaced during the further development of the programme, and one even knows, more or less, how.

We may call the counterfactual initial conditions mentioned in Definition 4 *idealizing assumptions* or simply *idealizations.* Let us pause for a moment and compare Lakatos' definition with more recent ideas and distinctions concerning the role that models play in scientific idealization. Some terminological conventions have become standard in the literature on which we should at least comment. In their survey article for the *Stanford Encyclopedia of Philosophy*, Frigg and Hartmann (2006) distinguish between *Aristotelian* and *Galilean idealization*:

> Aristotelian idealization amounts to "stripping away", in our imagination, all properties from a concrete object that we believe are not relevant to the problem at hand. This allows us to focus on a limited set of properties in isolation. An example is a classical mechanics model of the planetary system, describing the planets as objects only having shape and mass, disregarding all other properties. Other labels for this kind of idealization include "abstraction" . . . , "negligibility assumptions" . . . and "method of isolation". . . .
>
> Galilean idealizations are ones that involve deliberate distortions. Physicists build models consisting of point masses moving on frictionless planes, economists assume that agents are omniscient, biologists study isolated populations, and so on. It was characteristic of Galileo's approach to science to use simplifications of this sort whenever a situation was too complicated to tackle. For this reason it is common to refer to this sort of idealizations as "Galilean idealizations" . . . ; another common label is "distorted models".

Psillos's (2007, p. 6) dictionary for the philosophy of science offers a related dichotomy under the labels *abstraction* and *idealization*:

> Abstraction: The removal, in thought, of some characteristics or features or properties of an object or a system that are not relevant to the aspects of its behaviour under study. In current philosophy of science, abstraction is distinguished from idealisation in that the latter involves approximation and simplification. Abstraction is an important element in the construction of *models*.

Similarly, Chakravartty (2010, p. 38–39) writes:

> . . . an abstract representation is the result of a process of abstraction; that is, one in which only some of the potentially many factors that are relevant to the behavior of a target system are built into the representation. In such a process other parameters are ignored, either intentionally or unwittingly, so as to permit the construction of a tractable representation. A commonly discussed example of this is the model of the simple pendulum. Here, among other simplifying assumptions made in the construction of the model, one simply omits the factor of frictional resistance due to air. . . .
>
> On the other hand, an idealized representation is the result of a process of idealization; that is, one in which at least one of the parameters of the target system is represented in a way that constitutes a distortion or a simplification of its true nature. In such a process, one is not excluding parameters, as in abstraction, but incorporating them, again either intentionally or unwittingly, in such a manner as to represent them in ways they are not – indeed, as I shall use the term, in ways they could not possibly be. Idealized representations thus furnish strictly false descriptions of their counterparts in the world.

Although I think that the basic intuitions behind these distinctions have some appeal, I am not sure about their separation accuracy. Isn't "stripping away", "removing" or "ignoring" some relevant or irrelevant factor quite the same as pretending, against one's better knowledge, that something which is there isn't there? Doesn't it involve "distortion", "approximation" and "simplification"? To me it seems that *abstracting* from friction, say, is quite the same as expressly setting the friction coefficient to zero which must be an idealization. The approach we shall be advocating later deals with explicit counterfactual assumptions. As far as I can see, it should be suitable for covering both abstractions and idealizations in the sense now widely accepted in the philosophy of science.

Back to Lakatos. For him, the role of models is a key to understanding of the autonomy of science. Scientists according to Lakatos are not much disturbed by the appearance of anomalies because they have preconceived plans how to turn them into positive instances. In typical cases scientists work their way through from heavily idealized models to less and less idealized ones. In the paradigmatic relation between Kepler's laws and Newton's theory of gravitation, the Newtonian program of explaining planetary motion included the following sequence of idealizing assumptions (1970, pp. 135–136):

A_1 If the planets were perfectly spherical, . . .
A_2 If the planets did not attract each other, . . .
A_3 If the planets did not rotate, . . .
A_4 If the planets were point-like bodies, . . .
A_5 If there was only a single planet, . . .
A_6 If the common centre of gravity coincided with the centre of the sun, . . .

Another example, not completely unrelated to the first, is Bohr's program of explaining light emission (1970, p. 146)

A_1 If all atoms were built up as simply as the hydrogen atom, if electromagnetic fields had no effects on them, etc., ...
A_2 If the spin of an electron had no effects, ...
A_3 If the proton-nucleus were fixed and the electron revolved around it in an elliptic orbit, ...
A_4 If the electron revolved around the proton-nucleus in a circular orbit, ...

All these idealizing assumptions can be thought of as revising a complete and true theory of the subject area in question. We shall not engage in a discussion about whether Lakatos's rendering of these examples is historically correct,[4] but rather concentrate on something that might be called the "logic of idealization". I want to suggest that the following picture might be fruitful. Idealizations consist of revisions of a projected theory (the "regulative ideal" of a research programme) by idealizing, counterfactual assumptions. Due to the counterfactual nature of the idealizing assumptions, successive theories are strictly speaking inconsistent with one another, and this explains how they can differ in their treatment of novel facts and the explanation of anomalies. When performing such revisions, the hard core T_h of the research programme is kept constant, while the protective belt T_p (including any "auxiliary theories") is variable. The positive heuristics may often be hard to distinguish from the negative heuristics. This suggests the assignment of degrees of importance or unrevisability. I shall continue to call such degrees *priorities*. Once priorities are admitted, they may become objects of revision, and we have to account for the dynamics of prioritizations. The priorities associated with $T * A_1 * A_2 * \ldots * A_n$ will in general be different from those associated with T, but they may be systematically connected with the latter.[5] If this approach turns out to shed light on the dynamics of more and more complicated scientific research programmes, then revision may be established as an important intertheory relation.[6] Idealization is then essentially different from approximation, and it can be expected to yielding "deeper" intertheory explanations than simple approximations.

It is not quite clear whether Lakatos's story is intended as the description of an actual course of events, or rather as a dialectical unfolding of potential positions. Lakatos started somewhere very close to Popper's view and arrived in the vicinity of Kuhn. Still, in contrast to the latter, Lakatos aimed at something like a *logic of inquiry*,[7] and it is in this respect that we draw inspiration from Lakatos in the present chapter.

[4] I addressed the Kepler-Newton case in some detail in Rott (1989).

[5] Fortunately, 25 years after the seminal paper by Alchourrón et al. (1985), there is a rich variety of suggestions how to change doxastic preferences in iterated belief change. Cf. the survey given in Rott (2009).

[6] That is, a relation *between theories*, not *between research programs*.

[7] This term would be a much better translation of Popper's *Logik der Forschung*.

By admitting mutually inconsistent theories within a single scientific research program (and being aware that this is indeed the normal rather than the exceptional case), Lakatos's methodology can be read as a reconciliation of continuities and incompatibilities in science. This chapter attempts to offer a logical modelling for a similar (though perhaps not quite the same) relationship. I consider pairs of successive theories, T and T', representing scientific progress and argue that a *good* successor theory T' should explain its predecessor T by appeal to the latter's pretheoretically specifiable *applicability conditions*, where the applicability conditions may be either factual or possible or counterfactual in the light of T'. In a truly *superior* successor theory T', not only the success, but also the failure of the predecessor T should be explained, and this by appeal to the *violation of the latter's applicability conditions*. Drawing on models for belief revision from philosophical logic, I shall propose a formal analysis of intertheory relations between successive theories which makes sense of successor relationships even in cases of idealization without approximation.

3.2 Factual, Potential and Counterfactual Explanations

We use a simple deductive notion of explanation to carry out our program of modelling continuities and incompatibilities in scientific change in normal science. Thus *explanation* will be our principal intertheory relation. At the end we want to explicate what a (theoretically) "progressive" problem shift may be, namely, a transition to a successor theory T' that can somehow explain both (the success of) its predecessor theory T and the failure of T. In cases where this model applies, we shall have a kind of explanation that goes deeper than a plain approximate agreement of the empirical predictions made by the two theories.

It is plausible to assume that a theory T is superseded by a theory T' only if the transition from T to T' is in some sense continuous. One idea to make this more concrete is to say that T' explains T. Let us suppose that theories are sets of sentences.[8] Then the simplest concept of intertheoretic explanation is that T' explains T just in case T' deductively entails T. But this does not seem to be a very realistic idea, especially if T' is more general than T. A more adequate concept of intertheoretic explanation acknowledges that T' needs to be supplemented by some sort of additional information that helps establishing the link between earlier and later theory. This link may be conceptual (then we need bridge laws relating the theoretical terms of T to those of T'), but it may also be empirical.[9] After all, once the later theory has been accepted, it may turn out that the earlier theory is valid only within a restricted domain. So some description of this range of application has to be added

[8] We shall presuppose that the theories we are dealing with are consistent.

[9] From now on, I leave aside the problem of incommensurability. If there is conceptual disparity, T is always meant to denote an adequate translation of the original predecessor theory into the language of T', so that no bridge principles must be conjoined to A.

to T' in order to derive T.[10] We shall say that T' explains T if T', taken together with a suitable proposition A, entails T, where A characterizes the *application* or *boundary conditions* of T, as seen from the point of view of T'. We do not need to assume that A is unique, but in the following, A is to stand for some fixed non-theoretical (non-lawlike, empirical) sentence describing the initial or boundary conditions that define T's range of application. If A is already known to be true in T', then the refined definition reduces to the simple one.

The successor theory T' is even better, if, in addition to explaining its predecessor, it can also explain some *anomaly of* T, or in other words, if it can explain *the failure of* T, or "explain away" T (Sklar 1967, p. 112). In this case there is an empirical explanandum E and there are suitably described initial conditions J such that T together with J entails $\sim E$, while T' together with J entails E. It follows that in the domain described by J, the theories T and T' are incompatible.

But how can a theory at the same time explain its predecessor as well as the failure of that very predecessor? There would be no difficulties in this if we could simply decree that J be excluded from T's range of application, i.e., that A and J are logically incompatible. The problem with this is that *strictly speaking*, there is no clear boundary between T's range of application and the domain where the empirical findings get anomalous for T. Even within what could reasonably be called T's range of application, the predictions of T often prove to be *strictly speaking* false, i.e., strictly speaking, they give rise to an anomaly (at least viewed in the light of T'). This is consonant with Duhem's and Feyerabend's challenges who both insisted that successive theories are generally *inconsistent* with each other.[11] Still the question remains: How is continuity, despite incompatibility, possible?

Instead of saying that T' explains (the failure of) T directly, I find it more natural and more accurate to consider A and J as the propositions that explain T and $\sim T$ respectively, relative to (or simply, *in*) T'. I shall therefore change the perspective and use this latter terminology. I would like to propose indeed that applicability conditions play a central role in intertheory relations. The function of A is clear if it non-vacuously defines applicability conditions for T, but what is A good for if T is derivable from T' alone, or if A is plainly incompatible with T'? I suggest that in the former case A may still provide a *factual explanation* of T in T', while in the latter case A may provide a *counterfactual explanation* of T in T'. If A is neither derivable from, nor incompatible with, T', then one can say that A provides a *potential explanation* of T in T'.

The applicability conditions for T need not be true, nor compatible with T', because they may contain simplifying or idealizing, counterfactual assumptions that are necessary for a T'-theorist to derive T *strictly speaking*, and not just an approximate version of it. Any of the three logical relations between T and T' may

[10] This description will normally taken from the non-theoretical part of the language of T'.

[11] And Lakatos concurred, see for instance Lakatos (1970, pp. 157–158). Lakatos even held that some research programmes, like that of Bohr (pp. 140–154), progressed "on inconsistent foundations". Unfortunately our reconstruction will have no place for this phenonmenon.

make the transition from T to T' a rational move. Accordingly, let us distinguish the following cases:

Definition 5 T' is a *good (conservative) successor theory* for T iff there is a range of application A for T such that one of the following three conditions holds:

(a) A explains T in T', or
(b) A can explain T in T', or
(c) A would explain T in T'.

A successor theory T' that is not only good, but really *superior* or *progressive* should be strictly better than its predecessor theory T, and explain not only T (as far as it holds), but also the failure of T (as far as it fails). Depending on whether the explanation of T is factual, potential or counterfactual, the explanation of T's failure is counterfactual, potential, or factual (T fails nowhere, partly or entirely), respectively.

Definition 6 T' is a *superior (progressive) successor theory* for T iff T' is a good successor theory for T with applicability condition A and the corresponding one of the following conditions holds:

(a) $\sim A$ would explain the failure of T in T', or
(b) $\sim A$ can explain the failure of T in T', or
(c) $\sim A$ explains the failure of T in T' respectively.

3.3 Conditionals in Intertheory Explanations

In order to make more sense of the above approach, let us phrase the different types of explanation in natural language. by means of *because* clauses, and their siblings, subjunctive and indicative *if* clauses:

Definition 7

(a) *A explains T* in T' iff “Because A is the case, T holds” is in T';
(b) *A can explain T* in T' iff “If A is the case, then T will hold” is in T';
(c) *A would explain T* in T' iff “If A were the case, then T would hold” is in T'.

The varying formulations presuppose different relations between T' and A. In (a), T' is supposed already to entail that the applicability conditions A of T are satisfied, in (b) that T' neither implies nor excludes the satisfaction of A (or are satisfied under certain circumstances), and in (c) that A is not, or cannot, be satisfied. Case (c) answers the challenge presented by the incompatibility between successive theories.

An easy way of expressing what it means to explain the failure of T is to say it is the explanation of the negation of T. If we follow this line, we have

Definition 8

(a) *J explains the failure of* T in T' (*J explains away* T in T') iff J explains the negation of T in T';

(b) *J can explain the failure of* T in T' (*J can explain away* T in T') iff J can explain the negation of T in T';

(c) *J would explain the failure of* T in T' (*J would explain away* T in T') iff J would explain the negation of T in T'.

It is the violation of applicability conditions, I submit, which explains away T in T'. That is to say, the negation $\sim A$ of A takes the part of J above. In any case, it is not the successor theory T' itself, but rather the non-theoretically specifiable conditions J that do the explaining within T'.

Some of these ideas can already be found in Glymour's (1970, p. 345) discussion of the concept of reduction:

> . . . Galileo's law is an approximation which *would* approach the Newtonian truth as a falling body comes arbitrarily close to the surface of the earth, *if* all forces other than the gravitational attraction of the earth were negligible and if the earth were spherical. Galileo's law fails in fact *because* the earth is not spherical and because forces other than the gravity of the earth are not zero and because the gravitational force is a function of distance. In the explanation of why Galileo's law fails one is not simply committing the fallacy of denying the antecedent. Rather, one is implicitly contrasting a contrary-to-fact situation in which Galileo's law would hold with the real situation, in which Newton's laws entail the denial of Galileo's law – or at least the denial of a formal analogue of that law.

Glymour holds that intertheoretic explanation is "an exercise in the presentation of counterfactuals . . . a theory is explained by showing under what conditions it *would be* true, and by contrasting those conditions with the conditions which actually obtain." (p. 341) I find this diagnosis of Glymour's insightful. Just two things may be noted in a critical vein. First, Glymour wavers regarding the question what exactly is explained by the superior theory, the earlier theory (p. 341) or its failure (p. 345). Second, Glymour runs together questions of approximation ("approach the truth", "comes arbitrarily close") and idealization ("if this and that were the case"), without alerting the reader that approximation and idealization need not go hand in hand.

The meaning of the connectives "because" and "if" appearing in Definition 7 is not self-explanatory. For further analysis, we need a theory of counterfactual reasoning. An suitable model for this is Gärdenfors's (1978, 1988) doxastic semantics for conditionals of the subjunctive as well as the indicative variety. Gärdenfors continues a tradition started by Ramsey (1931, p. 247) and Stalnaker (1968, p. 102) and interprets conditionals in terms of *revisions* (*minimal changes*) of belief states. These changes are required to satisfy certain rationality postulates that have come to be widely known under the name *AGM postulates* (after Alchourrón et al. 1985; also see Gärdenfors 1988, Chapter 3). The most straightforward and well-known idea to implement such a doxastic semantics is the so-called *Ramsey Test*:

(RT) "If A then B" is in a theory T if and only if B is in $T * A$

Here $T{*}A$ is the minimal revision of T needed to accept A. Notice that the Ramsey test does not specify how the revision of T is to be effected, but we may assume that the priority structure mentioned above will play a key role here.[12] Ever since Gärdenfors (1986), the Ramsey test has been beset by various *triviality* or *impossibility results*. Excellent overviews of attacks and defences are given by Lindström and Rabinowicz (1998) and Nute and Cross (2001), a recent defence is made by Bradley (2007).

If we assume that $T{*}A$ is identical with T when A is already contained in T, then the Ramsey test entails that "If A then B" is in T as soon as both A and B are in T. Intuitively, this is not what we want. Rott (1986) takes over Gärdenfors's basic semantic apparatus, but slightly modifies the acceptability conditions for conditionals and suggests a *Strong Ramsey Test* (in place of the usual Ramsey test). The Strong Ramsey Test is intended to introduce an element of relevance of the antecedent for the consequent and to emphasize the affinity between conditionals and sentences containing "because", the latter being regarded as the standard formulations of explanations.[13] Here is the Strong Ramsey Test for the interpretation of conditionals:

(SRT) "If A then B" is in a theory T if and only if B is in $T * A$ but not in $T* \sim A$.

If the Strong Ramsey Test is followed, it is not sufficient that A and B happen to be included in T for "If A then B" to be accepted in T. A has to be positively relevant for B. In Rott (1986), the strong Ramsey test is not applied to T directly, but to a contracted version $T{\div}B$ of T that does not contain B. While this does not change anything for open and counterfactual conditionals, because the consequent B is not accepted in the relevant belief states anyway, it does have effects if the same idea is applied to sentences containing "because".

In the following I suggest to extend the doxastic semantics to cover sentences containing either "if" or "because". Using the AGM postulates for rational belief changes, it can be shown that this proposal reduces to the following conditions.

Indicative or open conditionals

"If A is true, then B is true" is in T if and only if neither of A, $\sim A$, B, $\sim B$ is in T, but the material conditional $A \supset B$ is in T.

Subjunctive or counterfactual conditionals

"If A were true, then B would be true" is in T if and only if $\sim A$ and $\sim B$ are in T and B is in $T * A$.

[12] Lakatos (1970, p. 114) writes, starting with a quote from Popper (2002, p. 78):

> "We need a set of rules to limit the arbitrariness of 'deleting' (or else 'accepting') a protocol sentence. . . ." . . . Popper agrees with Neurath that all propositions are fallible; but he forcefully makes the crucial point that we cannot make progress unless we have a firm rational strategy or method to guide us when they clash.

[13] Much of the intuitive motivation for this amendment of the Ramsey test came from Spohn's (1983, 1988) modelling that uses numbers (ordinals). On the more technical side, Gärdenfors (1987) was quick to show that the Strong Ramsey Test does not protect from triviality results. The term "Strong Ramsey Test" is used in a different sense by Giordano, Gliozzi and Olivetti (2005).

Factual conditionals
"Because A is true, B is true" is in T if and only if A and B are in T and $\sim A$ is in $T * \sim B$.

As intertheoretic explanations always appear to be of the *why-necessary* rather than the *how-possible* type, we use only the *nec*-version of "because" here (Rott 1986, pp. 355, 359). On this reading, the factual "Because A is true, B is true" is accepted if and only if its counterfactual contrapositive "If ~B were true, then ~A would be true" is accepted. This corresponds to an intuition of Goodman's (1954).[14] Formally, it is also equivalent with McCall's (1983, 1984) theory of counterfactuals and "factuals". Our rationale, however, is very different from McCall's. He specifies truth conditions based on branched possible worlds structures, while the words "if" and "because" as analyzed in the present chapter refer to doxastic relations, not to ontological ones.[15]

In the following, the predecessor theory T and its applicability conditions A are being represented as single sentences (think of a conjunction of a finite set of axioms). The successor theory, however, will be thought of as a set of sentences that is closed under some background logic Cn (think of the set of logical consequences of the axioms). This asymmetry is introduced for simplicity's sake only, because the most widely used belief change theories in the AGM tradition support this format. I do not think it introduces any substantial problems. Here is a list of the final consequences of our interpretation of the Definitions 5 and 6.

Observation 1 T' is a good successor theory for T iff
(a) A and T are in T' and ~A is in $T' * \sim T$, or
(b) none of A, ~A, T and ~T is in T' and $A \supset T$ is in T', or
(c) ~A and ~T are in T' and T is in $T' * A$.

The somewhat unexpected positions of A and T in case (a) of Observation 1 are due to the fact that our favoured reading of "Because A, T" is equivalent to its subjunctive contrapositive "If ~T then ~A". The same comment applies *mutatis mutandis* to case (c) of the next observation.

Observation 2 T' is a superior successor theory for T iff it is a good successor theory for T and
(a) $\sim T$ is in $T' * \sim A$, or
(b) $\sim A \supset \sim T$ is in T', or
(c) A is in $T' * T$, respectively.

[14] Goodman (1954, p. 14): "The problem of counterfactuals is equally a problem of factual conditionals, for any counterfactual can be transposed into a conditional with a true antecedent and consequent." Goodman's example is the transformation from "If that piece of butter had been heated to 150°F., it would have melted" to its *contrapositive* (Goodman's term) "Since that butter did not melt, it was not heated to 150°F."

[15] This is not to deny that doxastic structures of any kind may (or should) mirror ontological structures of appropriate kinds, in a way somehow similar to the way in which beliefs (should) mirror facts.

If T' is a good or superior successor theory for T according to one of the cases (b) or (c), then the transition from T to T' is nonmonotonic in the sense that T is not included in T'. Not everything that one thought one knew is kept in the new theory. In case (b), T is not considered wrong from the point of view T', but it is considered valid only in a restricted domain. In case (c), T is considered false (at least, strictly speaking false) from the point of view of T'.

Can we give a more compact characterization of superior successor theories? Yes, it turns out that we can, if we avail ourselves of slightly stronger means. While the first two observations are more or less immediate and require only the basic set of AGM axioms (the first six in the usual numbering), the proof of the next observation is a little more complex and requires an additional condition for belief change. AGM's complete set of rationality postulates for theory revisions validates the "Reciprocity Condition"

(Rec) (B is in $T'{*}C$ and C is in $T' * B$) iff $T'{*}B = T'{*}C$

The condition (Rec) is in fact equivalent to two assumptions that are much weaker than the supplementary AGM postulates (the seventh and eighth postulates in the usual numbering). It characterizes what is known as "cumulative reasoning" in nonmonotonic logic,[16] and this is all we need for the following

Observation 3 T' is a superior successor theory for T iff $T' * A = T' * T$ and $T'{*} \sim A = T'{*} \sim T$.

Observation 3 says that if T' is a superior successor theory for T, then the non-theoretical applicability conditions A for T are, viewed from the standpoint of T', revision-equivalent to the predecessor theory T itself, and this equivalence is independent of whether case (a), (b), or (c) is realized.[17] Observation 3 also says that this condition is not only necessary, but also sufficient for T' being a superior successor theory to T.

Let us prove this. Using (Rec), we can conclude from Observations 1 and 2 that T' is a superior successor theory for T just in case one of the following conditions is satisfied:

(a) A and T are in T', and $T'{*} \sim A = T'{*} \sim T$;
(b) none of A, ~A, T and ~T is in T', and $A{\equiv}T$ is in T';
(c) ~A and ~T are in T', and $T' * A = T' * T$.

[16] The conditions are (Cut) If B is in $T{*}A$ then $T{*}(A\&B)$ is a subset of $T{*}A$ (this is a weakening of AGM's seventh axiom) and its conditional converse, (Cautious monotony) If B is in $T{*}A$ then $T{*}A$ is a subset of $T{*}(A\&B)$ (which is a weakening of AGM's eighth axiom). The full set of AGM postulates corresponds to "rational reasoning" in nonmonotonic logic. Cf. Rott (2001, Chapter 4, especially p. 110).

[17] The concept of revision-equivalence used here is not a concept of full doxastic equivalence. The latter requires not only the identity of one-shot revisions like, e.g., $T' * A = T' * T$, but also the equivalence as a basis for iterated revisions like, e.g., $(T' * A) * B = (T' * T) * B$.

Let T' be a superior successor theory for T, i.e., either (a) or (b) or (c) is true. If (a) holds, then, since A and T are in T', we have $T' * A = T' = T' * T$; if (b) holds, we get, because none of A, ~A, T and ~T is in T' and because $A{\equiv}T$ is in T', both $T' * A = Cn(T' \cup \{A\}) = Cn(T' \cup \{T\}) = T' * T$ and $T' * \sim A = Cn(T' \cup \{\sim A\}) = Cn(T' \cup \{\sim T\}) = T' * \sim T$; if (c) holds, then, since ~A and ~T are in T', we have $T' * \sim A = T' = T' * \sim T$. Thus in all cases $T'{*}A = T'{*}T$ and $T' * \sim A = T' * \sim T$, and we are done with one direction of the claim.

Now suppose for the converse that $T' * A = T' * T$ and $T' * \sim A = T' * \sim T$. From this it follows that A is in T' just in case T is in T', and that ~A is in T' just in case ~T is in T'. Thus the first halves of the cases (a), (b) and (c) cover all possible membership combinations of A and T in T' (T' is assumed consistent). The second halves of cases (a) and (c) follow directly from the supposition. For case (b), we have to take into account that A is in $T' * A = T' * T = Cn(T' \cup \{T\})$, so $T \supset A$ is in T', because T' is supposed to be logically closed; and similarly, we get that $\sim T \supset \sim A$ is in T'. Thus one of the conditions (a), (b) and (c) is satisfied, i.e., T' is a superior successor theory for T. We are done with the second direction of the claim, and this finishes the proof of Observation 3.

While Observation 3 is a neat result, it seems to me that the most realistic or the most frequent relation between successive theories is the one described by Observation 1 (c), according to which $T'{*}A$ implies T. This preserves much the original intentions of the deductive concept of intertheory explanation, and yet it agrees with the doctrine of Duhem, Feyerabend and Lakatos that as a rule, successive theories (strictly speaking) contradict each other. We have found, I believe, an analysis of intertheory explanation that is worth further exploring.

3.4 Conclusion

A good successor theory tells us under what conditions its predecessor is or would be true, a superior theory tells us in addition, under what conditions its predecessor is false. Perhaps the most typical case is that in which the predecessor theory is considered to be wrong. Still the successor theory can say something nice about it: The predecessor theory would be true, *if* its (counterfactual, idealizing) application conditions were satisfied. But *because* they are not, the predecessor theory is false. By having an eye on factual, potential and counterfactual explanations, we were able to give an account of how a single theory can speak, as it were, at the same time in favour of and against another theory. This is also an intertheory relation that helps structuring the development of research programs in the sense of Lakatos.

There is an alternative answer to the problem situation outlined at the beginning of Section 2. It focuses on the concept of approximation and says that typically, T' only approximately explains T. The approach offered in the present chapter does not rule out approximation procedures in counterfactual reasoning. But in addition it seems capable of handling *idealizations* in theories that are most naturally construed as *non-quantitative,* and capable of handling *idealizations without approximate validity*. Laymon (1980) and Nowak (1980, pp. 79–81) discuss examples in point. Newton's idealization in his *experimentum cruces* and the idealizations used by the special theory of relativity in the Michelson-Morley experiment lead to qualitatively false predictions, while the idealizational laws describing Brownian motion or the speed of infusion of a liquid through a small hole prove to be a far cry from being even approximately true.

The consequences of the present analysis have to be tested severely. Moreover, the idea alone is of little value if one does not know *how* revisions are brought about in practice. On the one hand, our picture is committed to the methodological assumption that there are revision-guiding structures, heuristics in Lakatos's sense, which are part of scientific reality, that scientists have doxastic preferences, and they consider them to be essential for their research programs. On the other hand, all these abstract considerations have little worth if they are not found to be applicable to the practice of some serious scientific communities. We need careful studies of actual episodes in the history of science, and see whether they can be understood to conform to the model offered here. We need to identify explicitly the application conditions for the relevant predecessor theories. First steps in this direction were made in Rott (1989, 1991). From the point of view of the Newtonian theory of gravitation, one can justifiably say

> If the planets revolved round the sun like single bodies, then Kepler's laws of planetary motion would hold (in a slightly modified form); but because they don't, Kepler's laws actually fail to hold.

From the point of view of Van der Waals' theory of gases, it makes perfect sense to assert

> If gas molecules were mass-points, that is (roughly), if they were neither spatially extended nor subject to interacting forces, then the ideal gas law would hold; but because they aren't, the ideal gas law actually fails to hold.

But many more and more elaborate case studies have to be conducted in order to find out whether the ideas advanced in this chapter this are not merely a logician's plaything.

References

Alchourrón, Carlos E., Peter Gärdenfors, and David Makinson. 1985. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic* 50:510–530.

Bradley, Richard. 2007. A defence of the Ramsey test. *Mind* 116:1–21.

Chakravartty, Anjan. 2010. Truth and representation in science: Two inspirations from art. In *Beyond mimesis and convention: Representation in art and science*, eds. Roman Frigg and Matthew Hunter, 33–50. Dordrecht: Springer

Feyerabend, Paul K. 1975. Imre Iakatos. *British Journal for the Philosophy of Science* 26:1–18.

Frigg, Roman, and Stephan Hartmann. 2006. Models in science. In *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. http://plato.stanford.edu/entries/models-science

Gärdenfors, Peter. 1978. Conditionals and changes of belief. In *The logic and epistemology of scientific change*, eds. I. Niiniluoto and R. Tuomela, 381–404. *Acta Philosophica Fennica* (1979) 20, Amsterdam.

Gärdenfors, Peter. 1986. Belief revision and the Ramsey test for conditionals. *Philosophical Review* 95:81–93.

Gärdenfors, Peter. 1987. Variations on the Ramsey test: More triviality results. *Studia Logica* 46:319–325.

Gärdenfors, Peter. 1988. *Knowledge in Flux*. Cambridge, MA: MIT.

Giordano, Laura, Valentina Gliozzi, and Nicola Olivetti. 2005. Weak AGM postulates and strong Ramsey test: A logical formalization. *Artificial Intelligence* 168:1–37.

Glymour Clark. 1970. On some patterns of reduction. *Philosophy of Science* 33:340–353.

Goodman, Nelson. 1954. The problem of counterfactual conditionals. In *Fact, fiction, and forecast*, ed. Nelson Goodman, 13–34. London: Athlone.

Krüger, Lorenz. 1980. Intertheoretic relations as a tool for the rational reconstruction of scientific development. *Studies in History and Philosophy of Science Part A* 11:89–101. (Reprinted in Lorenz Krüger, 2005. *Why does history matter to philosophy and the sciences? Selected essays*, eds. Thomas Sturm, Wolfgang Carl, and Lorraine Daston, 79–92. Berlin: de Gruyter.)

Lakatos, Imre. 1970. Falsification and the methodology of scientific research programmes. In *Criticism and the growth of knowledge*, eds. Imre Lakatos and Alan Musgrave, 91–196. Cambridge, MA: Cambridge University Press.

Lakatos, Imre. 1971. History of science and its rational reconstructions. In *PSA 1970 – In memory of Rudolf Carnap*, eds. Roger S. Buck and Robert S. Cohen, 91–136. Dordrecht: Reidel.

Laymon, Ronald. 1980. Idealization, explanation, and confirmation. In *Philosophy of science association*, eds. P.D. Asquith and R.N. Giere, vol. 1, 336–350. East Lansing, MI.

Laymon, Ronald. 1982. Scientific realism and the hierarchical counterfactual path from data to theory. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1:107–121.

Lindström Sten and Wlodek Rabinowicz. 1998. Conditionals and the Ramsey test. In *Handbook of defeasible reasoning and uncertainty management systems*, eds. Dov Gabbay and Philippe Smets, vol. 3 (Belief Change), 147–188. Dordrecht: Kluwer.

McCall, Storrs. 1983. If, since and because: A study in conditional connection. *Logique et Analyse* 26:309–321.

McCall, Storrs. 1984. Counterfactuals based on real possible worlds. *Noûs* 18:463–477.

Miller, David. 1974. Popper's qualitative theory of verisimilitude. *British Journal for the Philosophy of Science* 25:166–177.

Nowak, L. 1980. *The structure of idealization*. Dordrecht: Reidel.

Nute, Donald and Charles, Cross. 2001. Conditional logic. In *Handbook of philosophical logic*, eds. Dov Gabbay and Franz Guenthner, 2nd edn., vol. 4, 1–98. Dordrecht: Kluwer.

Popper, Karl R. 2002. *The logic of scientific discovery: Logik der Forschung*. London: Routledge.

Psillos, Stathis. 2007. *Philosophy of science A–Z*. Edinburgh: Edinburgh University Press.

Ramsey, Frank P. 1931. General propositions and causality. In F.P. Ramsey, *The foundations of mathematics and other logical essays*, ed. J.B. Braithwaite, 237–255. London: Kegan Paul.

Rott, Hans. 1986. Ifs, though, and because. *Erkenntnis* 25:345–370.

Rott, Hans. 1989. Approximation versus idealization: The Kepler-Newton-case. In *Idealization II: Forma and applications*, eds. J. Brzezinski, F. Coniglione, T.A.F. Kuipers, and L. Nowak, 101–124. Poznan Studies in the Philosophy of the Science and the Humanities 17, Amsterdam: Rodopi.

Rott, Hans. 1991. *Reduktion und revision – Aspekte des nichtmonotonen Theorienwandels*. Frankfurt a.M.: Lang.

Rott, Hans. 2000. Two dogmas of belief revision. *Journal of Philosophy* 97:503–522.

Rott, Hans. 2001. *Change, choice and inference*. Oxford: Clarendon.

Rott, Hans. 2009. Shifting priorities: Simple representations for twenty-seven iterated theory change operators. In *Towards mathematical philosophy* (*Trends in logic*, vol. IV), eds. David Makinson, Jacek Malinowski, and Heinrich Wansing, 269–296. Dordrecht: Springer.

Sklar, Lawrence. 1967. Types of inter-theoretic reduction. *British Journal for the Philosophy of Science* 18:109–124.

Spohn, Wolfgang. 1983. Deterministic and probabilistic reasons and causes. *Erkenntnis* 19: 371–396.

Spohn, Wolfgang. 1988. Ordinal conditional functions. A dynamic theory of epistemic states. In *Causation in decision, belief change, and statistics*, eds. William L. Harper and Brian Skyrms, vol. 2, 105–134. Dordrecht: Kluwer.

Stalnaker, Robert. 1968. A theory of conditionals. In *Studies in logical theory*, ed. Nicholas Rescher, 98–112. Oxford: Blackwell.

Stegmüller, Wolfgang. 1979. *The structuralist view of theories: A possible analogue of the Bourbaki programme in physical science*. Berlin: Springer.

Tichý, Pavel. 1974. On Popper's definition of verisimilitude. *British Journal for the Philosophy of Science* 25:155–160.

Chapter 4
Abductive Belief Revision in Science

Gerhard Schurz

4.1 The Problem: Learning Within the Theory of Belief-Revision

4.1.1 AGM Belief Revision and Its Variants (Including Belief Base Revision)

I start this chapter with some major principles of AGM belief revision (so-called after Alchourrón et al. 1985). In what follows, L is an assumed (1st order) language which is identified with the set of its well-formed formulas. Small (Arabic) letters s_i denote arbitrary sentences; h_i hypotheses, e_i evidences; capital letters S_i denote sets of sentences, a_i individual constants, F_i, G_i predicate or relation symbols; $\Vdash$ stands for logical inference (S, $s_1 \Vdash s_2$ abbreviates $S \cup \{s_1\} \Vdash s_2$) and Cn for logical consequence, i.e. $\mathrm{Cn}(S) = \{s \in L\colon\ S \Vdash s\}$. K signifies a logically idealized belief system as represented in the AGM tradition, that is, a set of sentences (believed by an underlying epistemic agent) which is closed under deductive consequence, $K = \mathrm{Cn}(K)$. $|K(L)$ is the set of all belief systems in language L. The AGM-theory describes revisions of K under the influence of a *new* informational *input* s which is accepted by the epistemic agent. If s is consistent with K, the result of the addition of s to K is simply an *expansion* of K which is a function $+\colon |K(L) \times L \rightarrow |\,K(L)$ such that

$$\textit{Expansion}\colon K + s := \mathrm{Cn}(K \cup \{s\}). \qquad (1)$$

If s contradicts $K\,(K \Vdash \neg s)$, then one must first contract K by $\neg s$ before one can expand by s. The so-called contraction of K by s is a function $-\colon |K(L) \times L \rightarrow |\,K(L)$ which satisfies *at least* the following axioms ('$\supset$' for material implication and '$\equiv$' for material equivalence):

G. Schurz (✉)
University of Duesseldorf, Geb. 23.21, Universitaetsstrasse 1, D-40225
Duesseldorf, Germany
e-mail: schurz@phil-fak.uni-duesseldorf.de

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_4,

$$\begin{array}{l} \textit{Contraction}: K - s.\ \textit{Necesary axioms}: \\ \text{Closure: } K - s = \mathrm{Cn}(K - s). \\ \text{Success: if } \Vert\!\not\!\!- s,\ \ \text{then } s \notin \mathrm{Cn}(K - s). \\ \text{Inclusion: } K - s \ \subseteq\ K. \\ \text{Vacuity: if } s \notin K, \text{ then } K - s = K. \\ \text{Extensionality: if } \Vert\!\!-\!\!-s_1 \equiv s_2\ , \text{ then } K - s_1 = K - s_2. \end{array} \tag{2}$$

AGM-contractions obey three further axioms: recovery, conjunctive overlap and conjunctive intersection. These axioms are controversial and I do not require that they are satisfied (cf. Hansson 2006, ch. 2).

A revision of K by s is the result of assimilating an s with contradicts K into one's belief system. According to the so-called 'Levi-identity', revision can be defined as a sequence of a contraction by $\neg s$ and an expansion by s as follows:[1]

$$\textit{Revision}: K^{*}s\ :=\ (K - \neg s) + s \tag{3}$$

Intuitively, $K{-}s$ is intended to be a *minimal* contraction of K (i.e. some *maximal* subset of K) which does not entail s. It is well-known that the minimality intuition does not work well for deductively closed belief sets. For assume $s \in K$ and $K{-}s$ is a *maximal* K-subset not implying s – a so-called *maxichoice* contraction. It is easy to prove that for *every* (arbitrary) sentence $\psi, K - s$ must either contain $s \vee \psi$ or $s \vee \neg\psi$. Hence, the revised system $K^{*}\neg s$ will, for every ψ, contain either ψ or $\neg\psi$; in other words, it will be syntactically complete. This is, of course, a non-sensical result and we will return to that point in Section 4.1.2 on the problem of learning ability of AGM revisions. For the time being, we conclude that although maxichoice contractions (and revisions) are *allowed* by the AGM-axioms, *reasonable* AGM-contractions will have to be logically weaker than maxichoice contractions.

A *variety* of different methods of contractions has been proposed in the literature. Especially important are *partial meet* contraction, defined as the intersection of the subclass of *preferred* maxichoice contractions, and the equivalent *entrenchment-based* contractions (cf. Gärdenfors 1988). The AGM-theory is especially famous for its beautiful *representation theorems*. It has been proved, for example, that an operation '$- : |K(L) \times L \rightarrow| K(L)$' is an AGM (partial meet) contraction iff '$-$' satisfies the above axioms plus recovery. Similar results have been proved for expansion and revision. While expansion is unproblematic, some of the AGM-axioms for contraction and revision – in particular, success, recovery, and the two conjunction axioms – have come under severe criticism.[2] I mention only success and recovery (the other axioms are not important for my purpose):

[1] Based on Levi (1977), Gärdenfors (1981) introduced "Levi-identity". Cf. Levi's "commensuration requirement" (1991, p. 65). Levi speaks of "replacement" instead of "revision".

[2] Important alternatives to AGM-contraction are *severe withdrawal* (Rott and Pagnucco 2000; Levi 2004 called it 'mild contraction') and *belief base contraction* (Hanson 1999). For an overview cf. Hansson (2006).

$$\textit{Success}: \text{for revision}: s \in K^{*}s; \text{ for contraction}: s \notin K - s \tag{4}$$

$$\textit{Recovery}: \text{if } s \in K, \text{then}: K \subseteq (K - s) + s \tag{5}$$

The axiom of recovery is violated in severe withdrawal (cf. footnote 3) and in belief base contraction (see below). Since belief base contraction is important for abductive belief revision, we do not require recovery either. On the other hand, we require the satisfaction of success. This axiom is justified by the fact that we understand belief revision as *input-driven* revision: even if a consistent input information contradicts the other beliefs of the epistemic agent, is accepted 'as true', and the inconsistency of $\mathrm{Cn}(K \cup \{s\})$ is resolved by removing *other* elements than s.

Levi (1991, 117ff) has criticized the axiom of success. He points out that sometimes when s contradicts K it may be more reasonable to reject s than to contract K. For Levi this is possible, because he understands expansion and revision as more or less deliberate operations in which the input s may be any kind of information (see Section 4.1.3). In contrast, I assume here a *weakly* empiricist and *foundation-oriented* view of belief revision, in which the inputs are assumed to be *evidences* which need not be infallible, but whose acceptance (as true) is almost always unproblematic. Under this assumption, the idealization of input-driven belief revision satisfying the axiom of success seems to be appropriate (cf. Schurz 2008b for an explication and defense of a foundation-oriented epistemology).

AGM belief sets are epistemically *non-founded*, insofar the operation of AGM-contraction ignores the *justificational* status of beliefs in K, but treats all beliefs in K *on par*. This favors a *coherentist* interpretation of AGM belief revision, which is in conflict with the foundation-oriented aspect of AGM belief revision as input-driven revision. The following example illustrates why epistemically non-founded belief revision is problematic:

(6) **Example 1:** Assume $h_1, h_2 \in K$, where h_1 is h_2's only justificational support; the new (consistent) evidence is e; $h_1 \Vdash\!\!\!\longrightarrow \neg e$, but $h_2 \nVdash\!\!- \neg e$. Moreover, assume the plausible entrenchment ordering over the belief state $K : \neg e < \neg e \vee h_2$. Then though $h_1 \notin K^{*}e, h_2$ will remain in $K^{*}e$.[3] But since h_2 is no longer justificationally supported in $K^{*}e, h_2$ should have been removed from $K^{*}e$ according to a foundation-oriented viewpoint.

Since belief systems of empirical science are foundation-oriented, a *realistic* modeling of scientific belief revision should reflect the justificational structure of belief systems. Based on this insight Hansson (1999) has developed the alternative model of *belief base revision*. Hansson describes belief systems K as deductive closures of certain belief bases $B, K = \mathrm{Cn}(B)$, where B an arbitrary set of sentences which is *not* deductively closed. He applies the operations of expansion, partial meet contraction and Levi-revision to the bases of belief systems, and lifts the results to

[3] Since entrenchment-based contraction satisfies $s \in K - e$ iff $e < e \vee s$, for every $s, e \in K$ with $e \notin \mathrm{Cn}(\emptyset)$ (cf. Hansson 2006, §2.2).

the full (deductively closed) belief systems as follows: if $K = \mathrm{Cn}(B)$, e is consistent and f is inconsistent with K, then (a) $B+e = B\cup\{e\}$ and $K+e = \mathrm{Cn}(B+e)$, (b) $B-\neg f$ is a partial meet of maxichoice contractions of B w.r.t. $\neg f$, and $K-\neg f = \mathrm{Cn}(B-\neg f)$, and finally (c) B^*f $:= (B - \neg f) \cup \{f\}$, and $K^*s = \mathrm{Cn}(B^*f)$. Hansson (1999, 2006, §5.4) gives a representation theorem for belief base contraction which validates all axioms in (2).

Like Olsson and Westlund (2006) I consider belief base revision as a *variant* of AGM revision (an 'AGM-type' revision), insofar the basic ideas of expansion, contraction and Levi-revision remain untouched; they are just applied to belief bases. From a foundation-oriented viewpoint, Hansson's model does not go far enough insofar it considers only deductive support and ignores non-deductive (e.g. inductive) support. In Hansson's model, the result which is desired for Example (6), $h_2 \notin K^*e$, is only obtained if it is assumed that h_1 but not h_2 is in K's belief base. It is questionable how this assumption can be justified, especially when h_2 is non-deductively supported by h_1.

However that may be, for the major problem of my chapter, the learning ability of belief revisions, the differences between different kinds of AGM-type modelings of contractions and revisions, are not important: this problem arises for all of these models, though in a different way.

4.1.2 The Problem of Learning Ability (or Epistemic Creativity)

Revision of general beliefs in the face of new incoming evidences is usually a *creative* process, in which old hypotheses are not only removed but often improved or renewed. In other words, new hypotheses are *learned*. Here is a first example:

(7) **Example 2:** Assume $h \in K$ is a quantitative hypothesis saying that gas pressure is proportional to gas temperature, and new data e come in and tell us that for low temperatures, the gas pressure is lower than predicted by h. Then scientists will not simply remove h from K, but replace h by a modified hypotheses h^* in which a new non-linear term has been added to the linear relationship predicted by h (for details see (26) in Section 4.3.3).

AGM-type belief revision does not contain any *mechanism* which would provide some sort of *learning ability*. The details of this failure are different, however, for AGM-revision and belief base revision. As we have seen in Section 4.1.1, AGM-maxichoice-revision is *irrationally speculative* – after each revision step it purports to be omniscient. Most of the (often uncountably many) maxichoice revisions K^*e are completely irrational: for instance, in the face of observations of black ravens, some maxichoice revision functions will output, among other things, the complete old and new testament, others will output the anti-inductive conclusion 'all so-far unobserved ravens are white', etc. Partial meet (and epistemic entrenchment) contraction offers the possibility to tame this wild speculation behavior, but it does not tell us *how* we should tame it, because no specification of the preference relation over maxichoice contractions (or of the entrenchment relation over the elements of K) is offered.[4] In conclusion, AGM belief revision does not contain learning mechanism because it allows one to 'learn' whatever is *consistent* with the data – and this is, of course, not learning in the proper sense of well-calculated inductive or abductive learning.

On the other hand, belief base revision (which is our preferred model of revision) does not contain any creative mechanism at all, be it a rational learning or irrational speculation. This holds because it follows from the definitions of belief base revision explained in Section 4.1.1, that for every B and s, $B^{*}s = (B-\neg s)\cup\{s\}$ is a *consistent subset* of $B \cup \{s\}$.

The next three theorems underscore my point about the missing learning-ability. For this purpose I need some further terminology. An evidence e_i is represented by a *singular* sentence of L, i.e. one containing no quantifiers (moreover, e_i is neither L-true nor inconsistent). As in formal learning theory the *stream of evidence* is represented as a potentially infinite sequence $(e) := (e_1, e_2, \ldots)$. The belief revision process is modeled by assuming an initial belief system K_0 which is successively revised by the evidences in (e); thus, $K_n = (K_{n-1}{}^{*}e_n)$. Theorem 1 tells us that whenever a hypothesis h is learned in belief stage K_n, then the implication $e_n \supset h$ must be contained in the e_n-corrected previous belief system $K_{n-1} - \neg e_n$, and iteratively, the implication $\bigwedge_{1\leq \mathrm{i} \leq n} e_i \supset h$ must be contained in the $(e_1, \ldots, e_n)$*-corrected prior* belief system $K_0 - \neg e_{1-n} := (_{\ldots}(K_0 - \neg e_1) - \neg e_2) - \ldots) - \neg e_n)$.

Theorem 1 If $h \in K_n$, then (1.1) $(e_n \rightarrow h) \in K_{n-1} - e_n$, and (1.2) $(\bigwedge_{1\leq \mathrm{i} \leq n} e \rightarrow h \in K_0 - \neg e_{1-n}$.

Proof: *For (1.1):* $K_n = \mathrm{Cn}((K_{n-1} - \neg e_n) \cup \{e_n\})$; so $(K_{n-1} - \neg e_n)$, $e_n \Vdash\!\!\!- h$; hence $(K_{n-1} - \neg e_n) \Vdash\!\!\!- e_n \rightarrow h$. – *For (1.2):* Define $h := h_n$, and $(e_n \rightarrow h_n) = h_{n-1}$. By induction on n, (1.1) implies $K_0 - \neg e_{1-n} \Vdash\!\!\!- (e_1 \rightarrow (e_2 \rightarrow \ldots (\mathrm{e}_{n+1} \rightarrow h) \ldots)$ and hence $K_0 - \neg e_{1-n} \Vdash\!\!\!- \bigwedge_{1\leq \mathrm{i} \leq n} e_i \rightarrow h$. *Q.E.D.*

The non-trivial case of Theorem 1 is given when $h_i \in K_{\mathrm{i}}$ but $\notin K_{i-1}$. I call sentences of the form $e_n \supset h$ or $\bigwedge_{1\leq \mathrm{i} \leq n} e_i \supset h$ *learning sentences*. They describe in an apriori way how the underlying agent *would* generate new hypotheses under the influence of new evidences. Theorem 1 tells us that learning ability can be modeled within AGM-type belief revision (if and) *only* if learning sentences are assumed to be in the belief system from the start and remain there until their antecedent is verified by evidence.

Following from what was said above, there will of course be some maxichoice contraction function which contains and preserves just the right learning sentences, but we don't know this function. One the other hand, belief base revision systems *without* learning sentences will never start learning. This is exemplified in the next two theorems. A sentence h is called an *essentially general* hypotheses if h is not

[4] Apart from that, Rott (2000, pp. 508–512) has shown that entrenchment-based AGM-contraction does not always work in the intended way: given $p, q \in K$ and $p, q \Vdash\!\!\!- r$, it may be happen that $K - r$ retains the less entrenched one of $\{p, q\}$ instead of the more entrenched one.

entailed by any consistent set of singular statements; and h is called a *simple universal* sentence if h is of the form $\forall x M(x)$ where $M(x)$ is built up from monadic predicates and the free variable x. Theorem 2 tells us that if one starts from a prior belief system which contains no general hypotheses at all, then belief base revision will *never* generate any essentially general hypotheses. Theorem 3 asserts that if one starts with a system of *simple* universal hypotheses, then belief base revision may only remove them by falsification, but no new simple universal hypothesis will ever be learned.

Theorem 2 Assume K_0 is the deductive closure of consistent set B_0 of singular statements, and is revised via belief base revision. Then at no time n K_n will contain a essentially general hypotheses.

Proof: Abbreviate $e := \bigwedge_{1 \leq \mathrm{i} \leq n} e_i$. If Theorem 2 were violated for $h_n = h$, then by Theorem 1 $(e \rightarrow h) \in K_0 - \neg e_{1-n}$, hence $B_0 - \neg e_{1-n} \Vdash e \rightarrow h$, and therefore $B_0 - \neg e_{1-n}, e \Vdash h$ would hold. Since $B_0 - \neg e_{1-n} \subseteq B_0$, this means that h would be entailed by a consistent set of singular sentences, which contradicts the essentially general nature of h. *Q.E.D.*

Theorem 3 Assume K_0 is the deductive closure of a consistent set B_0 of *simple* universal hypotheses, and is revised via belief base revision. Then at no time n K_{n} will entail a *new* simple universal hypothesis.

Proof: If Theorem 3 is violated at time n, then by Theorem 1 (and $e := \bigwedge_{1 \leq i \leq n} e_i$) $(e \rightarrow u) \in K_0 - \neg e_{1-n}$ and hence $(B_0 - \neg e_{1-n}), e \Vdash u$ would hold for some simple universal sentence u with $B_0 \nVdash u$. So there exists some model (D_1, I_1) which verifies $B_0 - \neg e_{1-n} \cup \{e\}$ and some *disjoint* model (D_2, I_2) which verifies $B_0 - \neg e_{1-n} \cup \{\neg u\}$. We join these models into $(D, I) := (D_1 \cup D_2, I_1 \cup I_2)$. Since $B_0 - \neg e_{1-n} \cup \{\neg u\}$ is a sentence set which contains no individual constants, no conflict between I_2 and the interpretations of individual constants by I_1 can arise. The joined model will still verify $B_0 - \neg e_{1-n}$ because this set consists of *simple* universal sentences which are verified by (D, I) iff they are verified by both (D_1, I_1) and (D_2, I_2). Moreover (D, I) will still verify e because e is singular, and $\neg u$ because $\neg u$ is existential. *Q.E.D.*

For more complicated universal hypotheses Theorem 3 does not go through, but more complicated theorems could be obtained. Note that none of my theorems entails that learning is impossible in be AGM-type belief revision; but the question which assumptions and conditions are necessary for reasonable learning in AGM-type belief revision has not be discussed in the dominant literature on belief revision.

Let us turn to some realistic examples from scientific belief revision. I represent a theory as a set T of characteristic axioms (so T is not deductively closed). T is the union of a theory core C and a series of auxiliary hypotheses $a_1, \ldots, a_n, T = C \cup \{a_1, \ldots, a_n\}$, *partially* ordered by a relation of epistemic preference ($>_e$) such that $C >_e a_i$ for all $a_i (1 \leq i \leq n)$. If T is revised by a sequence of evidences (e) which are unexplained by or do even contradict the full theory T, then as long as T does not entail learning sentences, AGM-type belief base revision may *remove* auxiliary hypotheses from T, but it will not create new or modified auxiliary hypotheses which are able to *explain* the new and/or conflicting data. Scientific theory revision will typically do this, as the next two examples show.

(8) **Example 3** When Adams and Leverrier in 1846 recorded a significant deviation of Uranus' orbit from the predicted orbit, they did not just remove the auxiliary hypothesis 'the only force acting on Uranus is that of the sun', but they replaced it by the new auxiliary hypothesis 'there exists an hitherto unobserved small planet, called Neptune, whose gravitational force deflects Uranus's orbit', which together with the remaining part of the theory could explain the observed positions of Uranus' orbit.

(9) **Example 4** Sickle-cell hemoglobin is an abnormal variant of hemoglobin, the oxygen-carrying protein in the red blood cells (cf. Ridley 1993, 110f). The allele (genetic variant) which is responsible for sickle hemoglobin, call it *s*, is lethal only in its homozygote form *ss*, but not in its heterozygote form *sh* (*h* for the normal hemoglobin gene). In the 1930s an increased frequency of sickle-cell hemoglobin was discovered in African populations, much higher than what was predicted by the auxiliary hypothesis that the fitness ordering of hemoglobin genotypes is $hh > hs > ss$. When Haldane reflected on these surprising data in 1949, he did not only remove this auxiliary hypothesis, but he replaced it by a new one, namely that apart from its fitness disadvantage sickle-cell hemoglobin has an additional fitness advantage for African populations, with the result that the fitness ordering favors the heterozygote form *hs*, i.e. the correct fitness ordering is $hs > hh > ss$; this would explain the increased frequency of sickle cell anemia in African populations. Haldane conjectured that the additional advantage could be an increased resistance against the Malaria virus, which was later confirmed by Allison.

Let me compare my diagnosis with that of Hans Rott (1992, ch. 8–9, 1994). He has suggested to interpret scientific theory revision as a kind of 'backward revision'. He represents the actual theory T as the revision of an ideally *true* theory T^+, which would have to be developed in the future, by a variety of *idealization* assumptions $i_1, \ldots, i_n$: $T = T^{+*} i_1{}^* \ldots {}^* i_n$. For example, if T^+ is the ideally true Newtonian theory of the planetary system, then the i_k assume that sun and planets are point masses, that neither planets nor sun rotate, that inter-planetary forces are neglectible, etc. (cf. 1994, p. 40). Rott describes theory progress than as an inverted ('backwards') revision process (p. 42): the data stream (e) forces scientists to remove successively more and more idealization assumptions until finally the ideally true theory stands 'naked' before their eyes. Rott's idea is logically ingenious. Unfortunately it is also unrealistic. Since the ideally true theory T^+ is *not known* to the scientists, their actual theory T cannot realistically be represented from the 'future viewpoint'. An ideally true and complete Newtonian theory of the planetary system has never be formulated; its complexity would exceed all reasonable bounds. An actual scientific theory has rather the structure $T = C \cup \{a_1, \ldots, a_n\}$ as explained above, where the 'naked' theory core is not an ideally true theory but a couple of general principles which without further assumptions are void of empirical content (cf. Sneed 1971, pp. 118, 127). Scientific theory revision does not remove the auxiliary assumptions, but replaces them by better and new ones.

So far we have illustrated the importance of learning mechanisms only for belief *revision* processes. But of course, learning mechanisms play also an important role in an adequate model of scientific (as well as common sense) belief *expansion*. Inductive generalizations as well as abductive conjectures accompany belief expansions by new observations, in science as well as in common sense cognitions. After observing several instances of a 'constant conjunction', humans almost automatically form the corresponding inductive generalization; and after performing a new experimental result sufficiently many times, experimental scientists proclaim the discovery of a new empirical law. Given that an adequate model of belief revision is required to include learning mechanisms, then the same should hold for belief expansions. In this respect, the diagnosis of AGM-type models of belief expansions is very simple. AGM-type expansion is not at all creative but merely *additive*: it simply adds the new information and forms the deductive closure, but never generates new (non-logically entailed) hypotheses.

4.1.3 Corrective Versus Creative, Input-Driven Versus Deliberate: Quo Vadis?

I call revision without learning ability *corrective* revision, and revision with learning ability *creative* revision. I have argued that belief revision of real agents in real environments (be it common sense of scientific) is never purely corrective: in face of new evidence we do not simply remove false hypotheses but replace them by better ones. The same holds for expansions: we do not simply add new data but – from time to time – create new hypotheses. Concerning the dichotomy 'corrective' versus 'creative' revisions (or 'additive' versus 'creative expansions, respectively), belief revision theory faces a *dilemma*. Modeling corrective belief change is a manageable and logically beautiful task, but with restricted applications to real-life belief revision (be it common-sense or scientific). Modeling creative belief change would have many such applications, but it is a very complex and in complete generality presumably an even impossible task.

Quo vadis? If we go for creative belief expansion and revision, we should be clear about the *task* which such a theory should fulfill. Should such a theory tell us under any circumstances how one should *rationally* change her belief system in the face of new evidence? Indeed, if we had a complete theory of this sort, all problems of epistemology were solved. But certainly a theory of belief revision cannot answer all epistemological questions. Rather, it presupposes certain standards and methods of rational believers, especially (a) criteria for reliable evidence, and (b) patters of reliable inference, which have to be justified independently from such a theory. In the following sections of this chapter I will assume these presuppositions.

Even under these presuppositions it is not clear which way a theory of creative belief change should go. It can model creative belief revision as an input-driven versus a deliberate process. A theory of *input-driven* creative belief revision would tell us which explanatory hypotheses one should form, given a background belief system K and a new evidence e. Hence such a theory would provide *discovery algorithms* for successful scientific hypotheses. For Popperians, such discovery algorithms are impossible – and I agree that they are impossible in complete generality. Many contemporary philosophers of science agree, however, that in certain cases such discovery algorithms exist, in the form of inductive or abductive inference patterns, and that they are extremely important (cf. Schurz 2008a). Peirce (1903, CP 5.171) has pointed out that abductively (or inductively) inferred hypotheses are never sufficiently confirmed by the mere fact that they are the best explanations which are available at the given time – they are always subject to *further* test operations. Therefore, a theory of input-driven creative belief revision must assume a more fine-grained structure of the belief system K. At least, three subsets of K have to be distinguished: the subset $E \subseteq K$ of accepted *evidences*, the subset $S \subseteq K$ of *settled* hypotheses which are taken as background beliefs, and the subset $H \subseteq K$ of *unsettled* hypotheses which require further testing. Whenever a creative belief revision $K{*}e$ generates a new of modified hypothesis h, h is first moved into the subset H; and only after sufficiently many further confirmations it will move at some later stage into subset S. I owe the distinction between 'unsettled' and 'settled' hypotheses to Levi (1980) and Olsson and Westlund (2006) (see also Section 4.2.3).

An alternative way of understanding creative belief revision has been suggested by Isaac Levi. Levi (1980, 35f; 1991, 71ff, p. 146) distinguishes between *routine* versus *deliberate* expansions (and similar for contractions and revisions). While routine expansions are input-driven, deliberate expansions enrich a belief system by a hypothesis which is the conclusion of a non-deductive (e.g. inductive or abductive) inference. Thus according to Levi in creative belief revisions one revises K directly with a new or modified hypothesis h, $K{*}h$.

In the move from input-driven to deliberate expansion (or revision) the crucial questions have changed completely. While in input-driven revision, only evidences can be revisers, in deliberate revision arbitrary hypotheses can be revisers. While in input-driven revision the reviser (evidence e) is given and the creative aspect is contained in the *effect* of the reviser on the belief system, in deliberate revision the creative aspect is entirely contained in the reviser (hypothesis h), and the crucial question is to choose the right reviser of K, while the effect of revising K by h is described as a purely corrective revision. Levi (1980, 52f) models deliberate expansion as a *rational choice* process in which the epistemic agent choices that hypothesis out of a partition of possible hypotheses which she regards as the best one. Note that according to Levi, as soon as an epistemic agent deliberately expands or revises K with h, she regards h as settled (1980, pp. 28, 41) – thus in contrast to our approach, Levi rejects the introduction of unsettled beliefs in K.

Again, quo vadis? In this chapter I follow the route of input-driven creative revision, rather than that of deliberate revision. The reasons for my choice of this route are explained in Section 4.2.3 and can be summarized as follows: (1) the assumption of partitions of hypotheses from which one chooses best elements is often unrealistic, while (2) in many (but not all) cases, input-driven rules for discovering promising explanatory hypotheses are epistemically available and cognitively feasible. Let me emphasize that the *deliberate* element of *accepting* a hypothesis has not vanished in a model of input-driven creative belief revision; it has just shifted. It is not contained in the input-driven (abductive) revision step which produces a most promising but still unsettled hypothesis, but in the follow-up decision to settle this hypotheses, i.e. move it from subset H to subset S.

4.2 Two Ways of Incorporating Learning Ability into Input-Driven Belief Revision and an Alternative

4.2.1 Martin and Osherson: Joining Belief Revision with Formal Learning Theory

My notion of a 'learning sentence' has been motivated by the work of Martin and Osherson (1998). Their formal setting is as follows: formulas of a countable 1st order language L are interpreted by models over a countably infinite domain $|N = \{1, 2, \ldots\}$; L contains standard names $\underline{k}$ for all $k \in |N$; and a data stream (e) enumerates all closed literals (atomic formulas or their negations) which are true in a given model mod((e)).[5] Formal learning theory is interested (among other things) in the question which kinds of hypotheses can be learned *in the limit* given K_0, which means there exists a computable function which for all data streams compatible with K_0 conjectures one of $\{h, \neg h\}$ at every time, and which stabilizes after some finite time to the true element of $\{h, \neg h\}$. For purely universal hypotheses $\forall xFx$ such a method may consist, for example, in conjecturing $\forall xFx$ as soon as at least one input of the form $F\underline{k}$ (for some $k \in |N$) and no input of the form $\neg F\underline{k}$ has occurred in the data stream. This method can be implemented into AGM-type belief revision by including a learning sentence of the form $F\underline{k} \supset \forall xFx$ in K_0. If $\forall xFx$ is true, $\forall xFx$ will enter the belief system at the first time at which $F\underline{k}$ has entered the data stream, and remain there forever, while if $\forall xFx$ is false, $\forall xFx$ will be removed from the belief system (provided it was there) at the first time at which a sentence of the form $\neg F\underline{m}$ $(m \in |N)$ has entered the data stream, and will remain removed there forever. In the case of $\exists$-$\forall$-hypotheses with binary relations, learning sentence are more complicated than simple implications from evidences to hypotheses (cf. Martin and Osherson 1998, pp. 151–157). The general result can be summarized as follows:

$$\textit{Result of Martin and Osherson}\ (1998, (63), \text{p.}153): \qquad (10)$$

Terminology (see ibid, chs. 3–4)*:* (1) A *problem* is an n-tuple $(K_0, \{h_1, \ldots, h_n\})$ where K_0 is consistent and $\{h_1, \ldots h_n\}$ is a partition of K_0 (i.e., $\bigvee_{1 \leq i \leq n} h_i$ is logically

[5] Martin and Osherson (1998, pp. 62–64) use variable assignments instead of standard names.

equivalent with K_0, whence $\{\text{mod}(h_i)\colon 1 \leq i \leq n\}$ is a partition of mod(K_0)). (2) A K_0-data stream is an infinite sequence (e) which enumerates all true atomic formulas of some K_0-model; $E(K_0)$ denotes the set of all K_0-data streams, and $(e) \uparrow n$ denotes the initial n-segment of $(e) \in E(K_0)$. (3) A problem is solvable iff there exists an algorithm which computes for each $(e) \uparrow n$ (with $(e) \in E(K_0)$ a conjectured hypothesis in $\{h_1, \ldots h_n\}$, and which after some time n always conjectures that h_i which is true in the model underlying the sequence (e). (4) A belief system $B \supseteq K_0$ solves $(K_0\{h_1, \ldots, h_n\})$ via belief revision iff for every $(e) \in E(K_0)$ with true hypothesis h_i, $h_i \in K^*((e){\uparrow}n)$ for some $n \in |N$. (5) A belief base contraction function is stringent iff for each B and s, $B{-}s$ is a maximal B-subset not implying s, which comes first in an assumed total ordering of Pow(L) (cf. Martin and Osherson 1998, 132, (5+6)).

Theorem: For each solvable problem $(K_0, \{h_1, \ldots, h_n\})$ there exists a set of learning sentences L such $K_0 \cup L$ is consistent and $K_0 \cup L$ solves this problem via stringent belief base revision.

Martin and Osherson show that the stability of this result is destroyed if belief base contraction is replaced by ordinary AGM-contraction: there will of course be some AGM-contraction functions which yield the right result, even without learning sentences, but Martin and Osherson (1998, 169, (100)) prove that there exists solvable problems such that no belief system B_0 which is closed under the tautological rule 'p, $q/p \supset q$' can solve these problems via stringent revision functions.

In conclusion, Martin and Osherson provide a fascinating way of combining the theories of formal learning and of belief revision with help of prior learning sentences. Of course, their account has also its problems, and I list two of them:

(1) Prior learning sentences are somehow unnatural: we do not literally believe 'if this (and this ...) raven is black, then all raven are black'. Moreover, learning does not need particular learning sentences such as $F\underline{k} \supset \forall x Fx$, but learning *schemata* such as $\psi\underline{k} \supset \forall x \psi x$ for *every* predicate ψ. Sentence *schemata* are not believed; what they reflect are non-deductive *inferential* moves rather than beliefs. So the natural alternative to learning sentences is to expand belief systems by non-deductive inferences. This possibility is investigated in the remaining sections.

(2) Formal learning theory is restricted to hypotheses formulated in an observational language, whose true atomic sentences occur in the data stream. The reason for this restriction is that formal learning theory concentrates on hypotheses which are guaranteed to be learnable in the limit. For hypothesis involving theoretical concepts (concepts not occurring in the data stream) there cannot be such a guarantee. Even in the domain of observational hypotheses, only problems consisting of ∃-∀-hypotheses can be solved in the limit. However, if one drops this restriction of formal learning theory, then one may try to extend this account to learning hypotheses of any sort, including hypotheses with new theoretical concepts.

4.2.2 *Pagnucco 1996: Abductive Belief Expansion and Revision in the AGM-Tradition*

In his dissertation Pagnucco (1996) attempted to combine belief expansion and revision in the AGM tradition with a logical theory of abduction. His underlying idea fits our observations about belief revisions in science: a belief system K confronted with a new evidence e, be it conflicting or not, does not only add e to the data record but tries to *explain* e by expanding K by an abductive inference step. To model abductive belief revision by simple logical principles in the AGM tradition, Pagnucco (1996, p. 57) starts from an extremely general and weak notion of abduction: an abductive hypothesis (an 'abduction') for a new evidence e (a 'goal') with respect to a background belief system K (a 'domain') is *any* sentence s such that s is consistent with K and entails e together with K. Pagnucco defines an abductive belief *expansion* function as any function $++ : |K(L) \times L \rightarrow | K(L)$ satisfying

$$\begin{aligned}&\textit{Abductive belief expansion after Pagnucco}: K + +e = \mathrm{Cn}(K \cup \{s\})\\&\text{for some } s \in L \text{ such that } K \cup \{s\} \text{ is consistent and entails } e\text{, provided } K \cup \{e\}\\&\text{is consistent; otherwise } K + +e = K.\end{aligned} \tag{11}$$

Pagnucco (1996, pp. 101–105, 204–208) proves a first (and easy) representation theorem which holds for *finite* languages and says that '++' is an abductive expansion function according to (11) iff '++' satisfies the following axioms: (Closure:) $K + +s$ is a belief system, (limited success): if $\neg s \notin K$, then $s \in (K + +s)$, (inclusion:) $K \subseteq K + +s$, (4) (failure:) if $\neg s \in K$, then $(K + +s) = K$, and (consistency:) if $\neg s \notin K$, then $\neg s \notin (K + +s)$. A similar definition of abductive belief expansion, but without a representation theorem, was proposed by Aliseda (2006, pp. 74, 184).

In my view, the problem with Pagnucco's account in (11) is not that there is anything wrong with it, but that the underlying notion of abduction is too weak to be useful from the viewpoint of philosophy of science. An abductive expansion is a hypotheses h which *explains* a new evidence e in the given background system K. It is well-known from the philosophy of science literature that not any sentence which logically entails e in K is an scientific explanation of e given K. To obtain a representation theorem for (11), Pagnucco admits even the completely trivial 'explanation' $e \Vdash e$ (cf. the proof of lemma B.1 on p. 204). Even if one restricts explanations to what Pagnucco calls 'non-trivial abductions' (and Aliseda 'explanatory adductions'; 2006, 186), namely derivations $K, h \Vdash e$ such that $K \nVdash e$ and $h \nVdash e$, this would not change the diagnosis. The literature on explanation, which is largely ignored by Pagnucco (1996) and Aliseda (2006), is full of additional requirements which a non-trivial derivation must satisfy to count as an explanation: e.g., the explanatory premises must contain lawlike statements as well as factual statements, the latter ones must have causal relevance for the conclusion, all explanatory premises must be deductively or statistically relevant, etc. (for overviews cf. Salmon 1989, Schurz 1995/1996).

Apart from this weakness, Pagnucco's requirements are at the same time too strong in two other respects. First, the requirement that all explanations must deductively entail the explanandum is too strong, as there exist probabilistic explanations which merely increase the explanandum's conditional probability. Second, the requirement that the abduced hypothesis h must not entail e alone is too strong, as there exist cases in which the abduction produces a completely new explanation $h := \bigwedge H$ which does not make use of any beliefs which are already accepted in K (cf. Schurz 2008a, §7). Further philosophical problems of Pagnucco's account – for example, confusions of justification and explanation relations – are discussed in Páez (2006, ch. 2.1).

My remarks are not intended to criticize Pagnucco's interesting work, but to point towards fundamental problems. Pagnucco's ignorance of stronger requirements on explanations coming from philosophy of science, which he shares with many researchers in computational science, is quite understandable: this ignorance is even necessary if one wants to have some simple logical representation theorems.

To obtain a more non-trivial notion of abductive belief expansion, Pagnucco utilizes the notion of maximal expansions which are a modification of Levi's potential expansions in (1991). A *maximal expansion* of K w.r.t. e is a maximally consistent extension of K implying e. Max(K,e) denotes the set of all maximal expansions of K w.r.t. e. Pagnucco assumes some transitive ordering relation $\leq$ over the elements of Max(K,e) in regard to their 'explanatory quality' and considers abductive expansion functions which are obtained as the intersections of all elements of Max(K,e) which are maximal w.r.t. $\leq$. Pagnucco's result is the following (1996, pp. 107–114):

(12) *Pagnucco's second representation theorem:*

Definition '$++ : |K \times L \rightarrow | K$' is a transitively relational partial meet (t.r.p.m.) abductive expansion operation iff there exists a transitive relation '>' over Max(K,e) such that $K + +e = \cap\{K' \in \mathrm{Max}(K, e)\colon \neg\exists K'' \in \mathrm{Max}(K, e)(K'' > K')\}$.

Result: '++' is a t.r.p.m. abductive expansion operation iff ++ satisfies the five axioms mentioned below (11) plus the following three: (extensionality:) if $K \Vdash\!\!-\; e \equiv f$, then $(K{+}{+}e) = (K{+}{+}f)$, (axiom 7:) $K{+}{+}e \subseteq \mathrm{Cn}(K{+}{+}(e{\vee}f)){\cup}\{e\}$), and (axiom 8:) if $\neg e \notin K + +(e \vee f)$, then $K + +e \vee f \ \subseteq\ K + +e$.

This latter result of Pagnucco is certainly non-trivial. Unfortunately, the additional axiom 7 which results from the definition via partial meets of selected maximal expansions turns out to be strongly inadequate. Axiom 7 says that if we expand K by the best explanation of a disjunctive fact '$e \vee f$', e.g., Peter or Paul will win the race, and then add the fact that Peter has won the race (e), the resulting belief system will also entail the best explanation of this fact. But this need not be so: the best explanation of why Peter or Paul will win the race may be that Peter and Paul dominate their competitors by far and are equally good runners; in this case the mere addition of the fact that Peter has won the race does certainly *not* explain why Peter has won the race. Moreover, let us assume the following principle which is very reasonable for input-driven abductive expansion (where T stands for 'tautology'):

$$K + +T = K, \tag{13}$$

because tautologies are *not* in need of explanation. Then axiom 7 leads to the result that $K + +e \subseteq K + e$, which is means that abductive expansion *collapses* into ordinary expansion. *Proof:* By $\Vdash e \vee \mathrm{T} \equiv \mathrm{T}$, extensionality and (13), $K + +(e \vee T) = K + +T = K$. Hence by axiom 7, $K + +e \subseteq \mathrm{Cn}(K \cup \{e\}) = K + e$. Q.E.D.

In the final step of his work, Pagnucco reduces abductive belief revision '**' to the notion of abductive belief expansion '++' by a suitable contraction function '–' via Levi-identity:

$$\textit{Abductive belief revision after Pagnuco}: K^{**}a := (K - \neg a) + +a. \tag{14}$$

The same proposal is made by Aliseda (2006, 185). Pagnucco prefers a Levi-contraction function which does not satisfy the axiom of recovery (1996, 143ff), although he admits also AGM-contractions (p. 163). He shows that the abductive belief revision function as defined in (14) satisfies certain axioms which are similar to the axioms for ordinary belief revision except that the axiom $K^{*}s \subseteq K + s$ is no longer valid for belief revisions, i.e. $K^{**}s$ need *not* be a subset of $K + +s$. On the same reason, Pagnucco's axioms for belief revision are only necessary conditions but not sufficient ones.

In conclusion it seems that logically nice representation theorems for adequate *and* non-trivial notions of abductive expansions or revisions are not possible. Nevertheless Pagnucco's work is of central importance because he discovered these problems (unfortunately his work is so far unpublished, although it is quoted in several places). The fact that representation theorems in the AGM tradition are presumably not possible for realistic models of abductive belief revision need not be considered as a problem. In Section 4.3 I go some steps towards realistic models of abductive belief revision. Before that, I discuss alternative approaches to input-driven abductive belief revision.

4.2.3 Levi's Deliberate Expansions and Olsson–Westlund's Research Agenda: Alternatives to Input-Driven Revision

First I should mention that I understand the notions 'induction' and 'abduction' different from Levi. For Levi the major task of abduction is to generate a partition of possible answers, while the task of induction is to choose a best element from this space (Levi 1980, 42f). In contrast, I restrict the notion of induction to Humean inductive generalizations, and I understand abductions as inferences to a most promising explanation of a new fact. Next, I think that Levi's notion of deliberate expansion in which the epistemic agent chooses a 'best' hypothesis from a complete partition of possible hypotheses in terms of epistemic utilities is often unrealistic. In many cases of abductive belief expansion and revision, partitions of possible explanatory answers are neither *known* nor *needed*. What scientists have instead are heuristic input-driven abductive strategies. Let me give two examples:

(15) **Example 5** If a detective has to solve a murderer case, initially virtually everyone could be the murderer. The initial partition of possible answers is *too large* to be useful: the detective doesn't choose persons randomly from the telephone book and starts to interview them. He rather proceeds in collecting evidence. When he has acquired some evidences, e.g. a foot-print, or a testimony, he abduces *one or a few* real explanatory possibilities by the factual abduction method of Section 4.3.4. Now the detective works with a small and *incomplete* space of possible hypotheses (a or b or 'someone else' is the murderer). The incompleteness of this space is contained in the *default* element 'someone else' whose epistemic utility is completely unknown.

(16) **Example 6** When scientists seek for a theoretical model which explains an observed empirical regularity, e.g. the general fact that wood swims on water but stone sinks in it, then no space of possible theoretical models is given at all. The difficulty of theoretical model-abduction (cf. Schurz 2008a, §5) does not consist in the elimination of possible explanations, but to find just *one* plausible theoretical model which allows the derivation of the phenomenon to be explained. When Archimedes had found such a model in terms of buoyancy, this was celebrated as a great success, without looking at alternative models.

These two examples are not intended to refute Levi's account of selecting best hypotheses from partitions, but merely to point out that *besides* deliberate selection procedures one needs also input-driven abductive revision procedures in science. In other words, the option 'deliberate' versus 'input-driven' is complementary.

Insofar my account relies on the distinction between unsettled and settled hypotheses, it is related to the research agenda account of Olsson and Westlund (2006). They formalize research agenda as research questions which are in turn represented by partitions of their possible answers; questions are *settled* if their cardinality is reduced to one. Unsettled (vs. settled) hypotheses in my sense (recall Section 4.1.3) are elements of unsettled (vs. settled) research agenda in the sense of Olsson and Westlund. I do not assume, however, that every unsettled hypothesis must be element of a complete partition of hypotheses.

If Olsson and Westlund's account is coupled with ordinary AGM-type belief revision, the problem of the missing learning ability of Section 4.1.2 is not solved. For example, without learning sentences the question $\{\forall xFx, \neg\forall xFx\}$ will never be solved by a data stream emerging from a model in which $\forall xFx$ is true, because no finite initial string of data, $\bigwedge_{1\leq i\leq n} Fa_i$, verifies $\forall xFx$ or falsifies $\neg\forall xFx$. To settle this question it is necessary to furnish the belief system with the ability of inductive learning. Erik Olsson told me that he assumes K to be furnished with this ability. I am not sure, however, what Olsson has in mind here: deliberate expansions by inductive generalizations $(K + \forall xFx)$ in the sense of Levi, or input-driven inductive or abductive expansions $(K + + \bigwedge_{1\leq i\leq n} Fa_i)$ in my sense.

Independent from that question, I would recommend to modify Olsson and Westlund's interesting account in one respect. For Olsson and Westlund new research agenda are only generated by contractions, but never by expansions. However, in science new questions are typically generated through expansions by new or surprising evidences, on the simple reason that the most important scientific questions are *why-questions*: why did this person die?, why can iron be magnetized?, etc.

4.3 Steps Towards a Theory of Abductive Belief Revision

4.3.1 Patterns of Abduction

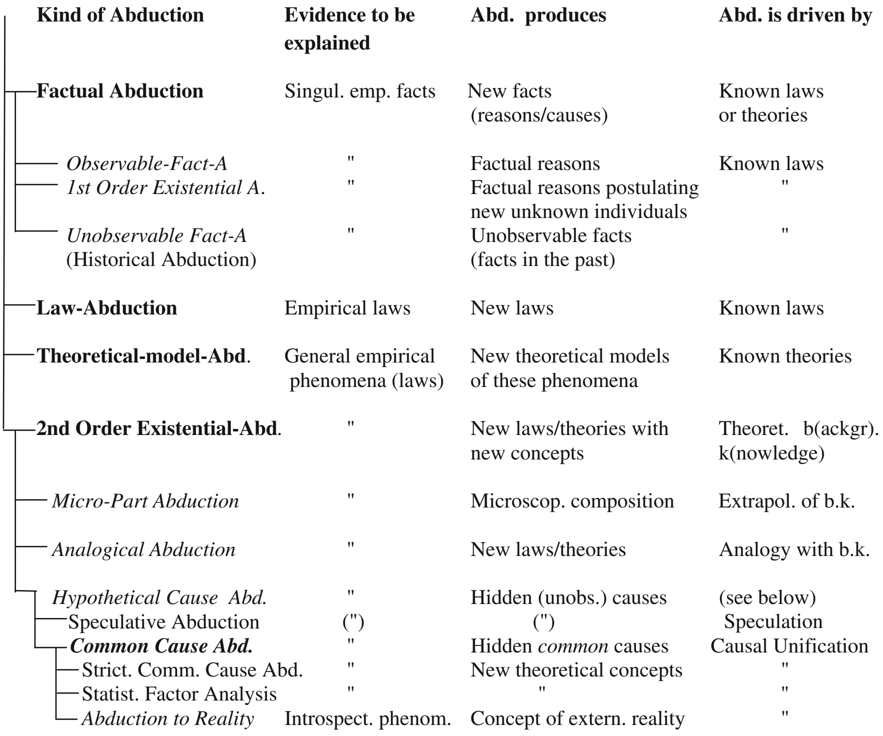
Peirce once remarked there are sheer myriads of possible hypotheses which would explain the experimental phenomena, and yet scientists have usually managed to find the true hypothesis after only a small number of guesses (cf. CP 6.5000). But Peirce did not tell us any abductive rules for conjecturing promising explanatory hypotheses; he rather explained these miraculous ability of human minds by their abductive instincts (CP 5.47, footnote 12; 5.172; 5.212). In Schurz (2008a) I introduce such rules, in form of a family of *local* and *specific* abductive inference patterns, which in certain (but not in every) kinds of epistemic situations specify a most promising but still unsettled abductive conjecture.

In Schurz (2008a) I classify patterns of abduction along three (not independent but related) dimensions: (1) along the kind of *hypothesis* which is abduced, (2) along the kind of *evidence* which the abduction intends to explain, and (3) according to the *beliefs* or *cognitive mechanisms* which *drive* the abduction. I signify the different kinds of abduction according to the first dimension. The classification is displayed in Fig. 4.1. The generating pattern as well as the evaluation criteria for abduced hypotheses depend crucially on the *kind* of abduced hypothesis and requires a *specific* discussion for each different pattern of abduction.

In what follows I will import this theory of abduction into my the theory of abductive belief *expansion* by assuming a certain *abduction function* ‘abd’ which outputs explanatory hypotheses in defined epistemic situations. My theory of abductive belief revision will go further than that. I will include not only proper abductions but also inductive generalizations as a special case.

4.3.2 Why Levi-Identity Fails For Abductive Belief Revision

Let abd: $|K(L) \times L \rightarrow L$ be an *abductive expansion function* which produces for certain epistemic scenarios in terms of a given (consistent) belief system K and a new evidential input e a most promising abductive conjecture abd(K,e) which explains e within K. I also admit ‘abd’ to be defined on sets of evidences, i.e. abd: $|K(L) \times \mathrm{Pow}(L) \rightarrow L$ (‘Pow’ for ‘power set’). Often, abd(K,e) is the conjunction of a finite set of hypotheses H, $\mathrm{abd}(K,e) = \bigwedge H$. In situations where no abductive strategy is available which generates an explanatory hypothesis meeting minimal scientific standards (cf. Schurz 2008a, §7.1), I set $\mathrm{abd}(K,e) := \mathrm{T}$ (T for ‘tautology’). $\mathrm{abd}(K,e) = \mathrm{T}$ shall also hold if e is K-inconsistent. The function ‘abd’ has to satisfy the following necessary but insufficient axioms:



Kind of Abduction	**Evidence to be explained**	**Abd. produces**	**Abd. is driven by**
Factual Abduction	Singul. emp. facts	New facts (reasons/causes)	Known laws or theories
Observable-Fact-A	"	Factual reasons	Known laws
1st Order Existential A.	"	Factual reasons postulating new unknown individuals	"
Unobservable Fact-A (Historical Abduction)	"	Unobservable facts (facts in the past)	"
Law-Abduction	Empirical laws	New laws	Known laws
Theoretical-model-Abd.	General empirical phenomena (laws)	New theoretical models of these phenomena	Known theories
2nd Order Existential-Abd.	"	New laws/theories with new concepts	Theoret. b(ackgr). k(nowledge)
Micro-Part Abduction	"	Microscop. composition	Extrapol. of b.k.
Analogical Abduction	"	New laws/theories	Analogy with b.k.
Hypothetical Cause Abd.	"	Hidden (unobs.) causes	(see below)
Speculative Abduction	(")	(")	Speculation
Common Cause Abd.	"	Hidden *common* causes	Causal Unification
Strict. Comm. Cause Abd.	"	New theoretical concepts	"
Statist. Factor Analysis	"	"	"
Abduction to Reality	Introspect. phenom.	Concept of extern. reality	"

Fig. 4.1 Classification of kinds of abduction (after Schurz 2008a)

$$\begin{aligned}
&\textit{Necessary axioms for } \text{`}abd(K,e)\text{'}:\\
&\text{Extensionality w.r.t.}e\text{: if } \Vdash e_1 \leftrightarrow e_2\text{, then abd}(K,e_1) = \text{abd}(K,e_2).\\
&\text{Consistency: } K \cup \{\text{abd}(K,e)\} \text{ is consistent.}\\
&\text{Explanation: If } \not\Vdash \text{abd}(K,e))\text{, then abd}(K,e) \text{ explains } e \text{ within } K.
\end{aligned} \tag{17}$$

The axioms are not sufficient, because abd(K,e) has to be not just any but a *most promising* explanation. Moreover, the precise conditions for an 'adequate' and a 'most promising' explanation can only be given for specific epistemic scenarios. If $K \cup \{\text{abd}(K,e)\} \Vdash e$ holds, we speak of a *deductive* explanation; but we also admit merely probabilistic explanations (recall Section 4.2.2). Given a function 'abd', the notion of abductive belief expansion '++' can be defined with help of the ordinary expansion operator '+' as follows:

$$\begin{aligned}
&\textit{Abductive belief expansion } \text{`}++\text{'}:\\
&K++e := (K+e)+\text{abd}(K,e)\text{, if } K+e\\
&\text{is consistent; otherwise } K++e = K.
\end{aligned} \tag{18}$$

Theorem 4 implies that Pagnucco's first representation theorem mentioned below def. (11) can be extended to our notion '++' in def. (18):

Theorem 4 '$++: |K(L) \times L \rightarrow| K(L)$' is an abductive expansion function according to (11) satisfying the axiom of extensionality, iff '++' is an abductive expansion function according to (18) for some abduction function 'abd: $|K(L) \times L \rightarrow L$' which satisfies the axioms in (17).

Proof: The case where $K + e$ inconsistent is trivial; so assume K+e is consistent. *Left-to-right:* If '++' satisfies def. (11), then we can identify abd(K,e) of def. (18) with the sentence s of def. (11), and the axioms of (17) for abd(K,e) are satisfied (extensionality by assumption). – *Right-to-left:* If '++' satisfies def. (18) for some function 'abd(K,e)' satisfying (17), then we can identify the sentence s of def. (11) with $e \wedge \text{abd}(K, e)$. So '++' satisfies def. (11), and extensionality by assumption. *Q.E.D.*

From the viewpoint of applications, Theorem 4 is not very interesting. In contrast to Pagnucco, I do not consider every function '++' satisfying (17) and (18) as an abductive expansion function (my axioms are only necessary conditions). The nontrivial part of the next sections will be to *define* the function 'abd(K,e)' for certain epistemic scenarios.

Abductive expansion of K by e is *harmless* in the sense that it is the superposition of an ordinary (corrective) belief expansion step '+e' and a pure abduction generation step '+ abd(K,e)'. In other words, the theory of abductive expansion is simply the 'sum' of the theory of ordinary expansion and the theory of abduction.

Abductive belief *revision*, however, is no longer harmless in this sense. The reason for this remarkable fact is that Levi's identity (recall (3) of Section 4.1.1) is not generally valid for abductive belief revision. Let '$-$' be a suitable (AGM- or Levi-) contraction function. If Levi-identity were valid, than also abductive belief revision would be decomposable into ordinary revision and abduction generation, since

(19) Abductive belief revision according to Levi-identity:

$$K^{**}_{\text{Levi}}e := (K - \neg e) + +e = (K - \neg e) + e + \text{abd}(K - \neg e, e) = K^{*}e + \text{abd}(K - \neg e, e)$$

However, I will show now that there are many situations were we have

$$\textit{Breakdown of Levi - identity}: K^{**}e \neq K^{**}_{\text{Levi}}e. \tag{20}$$

To avoid misunderstandings: of course, every belief change from K to $K^{**}e$ (where $K^{**}e \supseteq K^{*}e$) can be represented as a concatenation of a contraction $(K - \neg e)$ and an expansion by $\{e, h\}$ for a suitable conjunction of hypotheses h, $(K - \neg e) + e + h$ (cf. Levi 1991, 65; Schurz and Lambert 1994, §2.2). The problem is that h is not always determined as an abductive expansion of $K - \neg e$ by e. I see two main reason why Levi-identity fails for abductive belief revision and I call them 'the problem of old evidence' and 'the problem of incremental belief revision'.

The problem of old evidence: Attention – this is a different 'old evidence' problem than that of Bayesianism (for the latter one cf. Earman 1992, ch. 5). It goes as follows. Assume $K = \text{Cn}(E \cup \{h\})$ where h explains the old evidences in the set

E and a new evidence e falsifies h. Then $K - \neg e = \mathrm{Cn}(E)$, and $(K - \neg e) + +e$ would contain a new hypothesis h^{*} which explains e but *has forgotten* to explain the old evidence E. The new hypothesis h^{*} must not only explain the new evidence but at the same time keep explaining the old evidences. Therefore, $K^{**}e$ cannot be understood as the sequence $(K - \neg e) + e + \mathrm{abd}(K - \neg e, e)$ according to (19).

In an approach like Pagnucco's, $(K - \neg e) + +e$ could at least possibly be strong enough to explain in addition the old evidence E. But we have argued that Pagnucco's notion of 'explanation' as an arbitrary logical strengthening of the explanandum is inadequate, and we have decided to base abductive belief revision on functions of the form 'abd(K,e)' which generate a *specific* hypothesis explaining e given $(K - \neg e)$, but nothing else.

Levi-identity works only in the special case where the old evidence E is *explanans-separated* from e. This means that $h = \bigwedge(H \cup H')$ such that H explains E and H$'$ derives $\neg e$. In this special case the old explanations of the old evidence will be preserved in $(K - \neg e)$, i.e. $H \subseteq K - \neg e$, and $K^{**}e = (K - \neg e) + +e$ will hold.

If E is not explanans-separated from e, then Levi-identity can only hold in the following version, which I call ab initio abductive belief revision. We define $E(K - \neg e)$ as the *explanandum-loss* of $K - \neg e$, that is the set of all evidences which are explained in K but not in $K - \neg e$.

(21) *Ab-initio abductive belief revision* (where $E := E(K - \neg e)$):

$$K^{**}e = (K - \neg e) + +(E \cup \{e\}) = (K - \neg e) + e + \mathrm{abd}(K - \neg e, (E \cup \{e\})).$$

In ab initio revision, we generate the new hypothesis h^{*} which explains e as well as E from scratch: $h^{*} = \mathrm{abd}(K - \neg e, (E \cup \{e\}))$.

The problem of incremental belief revision: It is rather inefficient to remove h and generate h^{*} from scratch by ab-initio abductive expansion with $E \cup \{e\}$. It would be more efficient to obtain the revised hypothesis h^{*} by an direct (incremental) revision of the old hypothesis h given E and e, which automatically takes care of preserving the old explanations. In that case the new hypothesis h^{*} is obtained as a revision function 'rev' of the old hypothesis h given e and $K - \neg e : h^{*} = \mathrm{rev}(h, e, K - \neg e)$. The difference between ab initio and incremental revision is graphically displayed in the following Fig. 4.2:

$$\begin{tikzcd}
h \in K \arrow[rr, "h^{*}= \mathrm{rev}(h{,}e{,}K - \neg e)"] \arrow[dr, dotted, "- \neg e"'] & & h^{*} \in K^{**}e \\
& \begin{array}{c} K - \neg e\ (h \notin K - \neg e) \\ \text{state of explanatory ignorance} \end{array} \arrow[ur, dotted, "++\, E \cup \{e\}"'] &
\end{tikzcd}$$

Fig. 4.2 Ab initio revision $\cdot\cdot>$ versus incremental revision ($\rightarrow$): $h^{*} = \mathrm{rev}(h, e, K - \neg e)$

The *explanans-loss* $h(K - \neg e)$ of $K - \neg e$ is defined as the conjunction of all explanatory hypotheses which are in K but not in $K - \neg e$. Based on that notion, we define incremental belief revision $K^{**}e$ as follows, where rev: $L \times L \times |K(L) \rightarrow L$ is the incremental revision-function:

$$\textit{Incremental abductive belief revision}\,(\text{where}\, h := h(K - \neg e)):\qquad (22)$$

$$K^{**}e := (K - \neg e) + e + \text{rev}(h, e, K - \neg e).$$

(Later on we will apply 'rev' also to sets of new evidences E': $\text{rev}(h, E', K - \neg \bigwedge E')$.) We require from the function 'rev' that it outputs a revised hypothesis $h^* := \text{rev}(h, e, K - \neg e)$ which explains e as well as the explanandum-loss $E(K - \neg e)$. Since the revised hypothesis $\text{rev}(h, e, K - \neg e)$ depends on the old hypothesis h which is not in $K - \neg e$, Levi-identity is violated. One could defend Levi-identity by arguing that ab-initio revision $(K - \neg e) + e + \text{abd}(E \cup \{e\}, K - \neg e)$ leads always to the same result as incremental revision $(K - \neg e) + e + \text{rev}(h, e, K - \neg e)$. But apart from the computational advantages of incremental belief revision, I would regard this as unrealistic idealization.

Again, a successful incremental abductive belief revision does not exist for all types of epistemic problems, and if no improved hypothesis is at hand which does the required job, then we set $\text{rev}(h, e, K - \neg e) := \text{T}$. Our necessary (but insufficient) axiomatic requirements for 'rev' are these:

(23) *Necessary axioms for 'rev: $L \times L \times |K(L) \rightarrow L$':*

$$\begin{aligned}&\text{Extensionality}: {}'\text{rev}(h, e, K - \neg e)' \text{ is extensional w.r.t. } h \text{ and } e.\\ &\text{Consistency}: K - \neg e \cup \{\text{rev}(h, e, K - \neg e)\} \text{ is consistent.}\\ &\text{Explanation}: \text{If } \Vert\!\not\vdash \text{rev}(h, e, K - \neg e),\\ &\text{then } K - \neg e \cup \{\text{rev}(h, e, K - \neg e)\} \text{ explains } E \cup \{e\}.\end{aligned}$$

In the remaining sections ab initio as well as incremental ways of abductive belief revisions will be characterized for specific epistemic scenarios.

4.3.3 Inductive Belief Expansion and Revision

Induction can be considered as a logical subcase of abduction in the broad sense. In this section we discuss inductive belief revision as a simple example of incremental belief revision. For this purpose we restrict to *strict* (non-statistical) and *purely universal* generalizations, which are by definition formulas of the form $U := \forall x_1 \ldots \forall x_n M(x_1, \ldots, x_n)$ where M is free of quantifiers, and which are not entailed by any consistent set of singular sentences. First we characterize inductive expansions. According to Carnap's requirement of *total evidence* (1950, p. 211), inductive inferences have to be drawn in the light of all available and relevant evidence. A purely universal hypothesis U is called *elementary* iff it is not equivalent to a conjunction of purely universal hypotheses all of which are shorter than U (cf. Schurz 1991, 423). For example, $\forall xFx$, $\forall x(Fx \supset Gx)$, $\forall yGx$ are elementary, while $\forall x\forall y(Fx \wedge Gy)$ is not; the latter formula is logically equivalent with $\forall xFx \wedge \forall yGx$. We formulate the general procedure of inductive belief expansion as follows:

(24) *Inductive belief expansion* (for purely universal generalizations):

Format and assumptions: E is the (finite and consistent) set of all evidentially accepted (or 'known') closed literals, and K is a belief system which is consistent with E and does not contain any quantified hypotheses in terms of predicates occurring in E.

Definition: $\text{abd}(K,E) = U(E)$, where $U(E)$ is the set of strongest elementary purely universal hypotheses which are consistent with E and whose non-logical predicates occur in E.

Note: Definition of $K + +e$ as in (18), with 'E' instead of 'e': $K + +e = K + E + \text{abd}(K,E)$.

For example, if $E = \{Fa, Gb\}$, then $U(E) = \{\forall xFx, \forall xGx\}$; if $E = \{Fa, \neg Fb, Gb, Ga, \neg Gc\}$, then $U(E) = \{\forall x(Fx \supset Gx)\}$; etc.

Let us now turn to inductive belief revision. If scientists discover an exceptional instance to a hitherto well-confirmed purely universal hypothesis h, they do not simply remove h (in a Popperian fashion) but restrict the antecedent (or if-condition) of h by excluding the observed exceptional case. This can only be done in a *non ad-hoc* way if the exceptional instance has some specific properties by which it is distinguished from the confirming properties. As an example for such an exception-clause, consider the so-called anomaly of water: all liquids expand their volume when being heated, except water between 0 and 4°C. In the result, inductive belief revision proceeds incrementally and non ab initio. Our general explication of inductive belief revision is as follows:

(25) *Inductive belief revision* (for purely universal generalizations):

Format and assumptions: The total evidence E (recall (24) above) has the form:

$$E = \{Ma_i \wedge R_i a_i : 1 \leq i \leq n\},$$

where M is an (n-ary, possibly complex) predicate; a_i are (n-tuples of) individual constants, and R_i is a complex (n-ary) predicate which summarizes the remainder knowledge about the individual a_i. E.g. if a_i is a black swan, R_i is the knowledge that a_i has been observed in Australia.

The *hypothesis* $h = \forall xMx$ is the strongest purely universal hypothesis compatible with E ('x' is an n-tuple of variables if 'M' is n-ary.)

The *new evidence* $e = \neg Ma_{n+1} \wedge R_{n+1}a_{n+1}$ contradicts h.

Incremental revision algorithm: Search for some property H such that the new instance possesses property H but the old instances don't; i.e.: $Ha_{n+1}$is entailed by $R_{n+1}a_{n+1}$ and $\neg Ha_j$ is entailed by $R_j a_j$ for $1 \leq j \leq n$.

If found : replace h by the following $h^* = \mathrm{rev}(h, e, K - \neg e) : \forall x(\neg Hx \supset Mx)$.
Else : remove h; i.e. set $\mathrm{rev}(h, e, K - \neg e) = T$.   Definition of $K^{**}e$ as in (22).

Against (25) an AGM-defender could object that already AGM-revision alone would yield the exception-restricted hypothesis h^*, because h^* is entailed by h and is preserved in K^*e. However, the latter result does only hold for special entrenchment orderings. Apart from that, we have seen in (6) of Section 4.1.1 that this property of AGM-revision leads to inadequacies, and we have opted for belief base revision. Moreover, the entailment of h^* by h does no longer hold in a *quantitative* setting of inductive belief revision. Recall the curve fitting example (7) in Section 4.1.2. It is well known that every given set of data points $\{(x_i, y_i) : 1 \leq i \leq n\}$ in $|R^2$ can be approximated with arbitrarily high *precision* provided one takes a sufficiently complex *type* of curve (or function) $y = f(x)$; e.g. a polynomial curve of sufficiently high degree. But taking complex curves increases the danger of *overfitting* the data, whence one usually starts with a most simple (e.g. a linear) type of curve, and checks whether the achieved degree of approximation is sufficiently high; only if it isn't, one goes on to more complex curves (cf. Forster and Sober 1994). The procedure is formally described as follows:

(26) *Inductive belief revision for curve fitting:*

Format and assumptions: $K = \mathrm{Cn}(e \cup \{h\})$, where $e = \bigwedge\{f(x_i) = y_i : (x_i, y_i) \in D\}$, D is a set of data points $\{(x_i, y_i) \in \Big|R^2 : 1 \leq i \leq n\}$, and h expresses an optimal curve $y = f(x)$ of the polynomial family of degree n. Hence, h has the form $f(x) = c_0 + c_1 \cdot x + c_2 \cdot x^2 + \ldots + c_n \cdot x^n$ (with $c_i \in |R$), where h's parameters (c_i) are optimized by the method of curve fitting (minimizing squared deviations) such that the obtain standard deviation $s(h)$ satisfies $s(h) \leq s_{\max}$, where $s_{\max}$ is an assumed upper tolerance level for s. The new evidential *input* $e' := \bigwedge\{f(x_i) = y_i : (x_i, y_i) \in D'\}$ is given by a new set of data points D' deviating from h's predictions by far more than $s_{\max}$.

Incremental revision instruction: $h^* = \mathrm{rev}(h, e', K - \neg e')$ is the optimal curve which fits the total set of data points D ∪ D' from that polynomial family with smallest degree $m > n$ which yields a new standard deviation $s(h^*) \leq \mathrm{s}_{mx}$.

4.3.4 Factual Abduction in AI

The literature on abduction in A(rtificial) I(intelligence) is concentrated almost exclusively on abductions in the narrow sense of Peirce (1878) (cf. Josephson and Josephson 1994, Flach and Kakas 2000). This kind of abduction falls under the category of factual abductions of Schurz (2008a, §3). Given is a knowledge base $KB = (L[x], F[a])$ consisting of a finite set $L[x]$ of monadic implicational laws of the form $\forall x(\pm F_1 x \wedge \ldots \wedge \pm F_n x \supset \pm F_{n+1} x)$ going from conjunctions of open literals to open literals ("±" means "unnegated or negated"), and a finite set $F[a]$ of facts (closed literals) of the form $\pm Fa$ about the individual case a. Knowledge bases are not understood as deductively closed; so let us assume that $K = \mathrm{Cn}(L[x] \cup F[a])$. The task is to find 'potential explanations', i.e. derivations of a given explanandum fact

$e := Ga$ (the so-called 'goal') from K and factual abductive hypotheses. The set of possible factual abductive hypotheses (so-called 'abducibles') A[a] is defined as the set of all closed literals in K's language which are not further explainable in K, i.e. which are neither facts in $F[a]$ nor instantiations of consequents ('heads') of laws in $L[x]$ (cf. Paul 1993, p. 133). The *full* abductive task would be to find *all* possible explanations, i.e., all *minimal* sets $E[a]$ of literals such that (i) $E[a] \subseteq F[a] \cup A[a]$, (ii) $L[x] \cup F[a] \cup E[a]$ is consistent and (iii) $L[x] \cup E[a]$ logically implies Ga. Those elements of $E[a]$ which are not facts are the abductive hypotheses for Ga, abd(K,Ga). Abductive algorithms of this sort have been implemented in PROLOG by backward-chaining through implicational laws with backtracking to all possible solutions.

This kind of abduction problem is graphically displayed in Fig. 4.3 in form of a so-called *And-Or-tree* (cf. Bratko 1986, ch. 13). The *labelled* nodes of an And-Or-tree correspond to literals, unlabeled nodes represent conjunctions of them, and the directed edges (arrows) correspond to laws in $L[x]$. Arrows connected by an arc are And-connected; without an arc they are Or-connected. Written statementially, the laws underlying Fig. 4.3 are $\forall x(Fx \supset Gx)$, $\forall x(Hx \supset Gx)$, $\forall x(Q_1x \wedge Q_2x \supset Gx)$, $\forall x(R_1x \wedge R_2x \supset Fx)$, $\forall x(Sx \supset Hx)$, $\forall x(T_1x \wedge T_2x \supset Hx)$, $\forall x(Ux \supset Q_1x)$, $\forall x(Vx \supset Q_2x)$. Besides the goal Ga, the only known fact is T_1a. Since the task of finding *all* possible explanations has exponential complexity and, thus, is intractable, one is usually satisfied with algorithms which find some *most promising* explanation; this task has polynomial complexity and is tractable (cf. Josephson and Josephson 1994, ch. 7, p. 165, th. 7.1+7.2). The major method of finding a most promising explanation in the case of factual abductions is to furnish the abductive search space by probability (or plausibility) values, and to apply a simple *best-first* search: for each Or-node one processes only that successor node which has a highest plausibility value among all successors of this node. The route of a best-first abduction search is depicted in Fig. 4.3 by the bold arrow.

We summarize this method as follows:

(27) *Abductive belief expansion for factual abductions*:
Format and assumptions: $K = \mathrm{Cn}(L[x] \cup F[a])$; $e = G[a]$ *is consistent with K.*

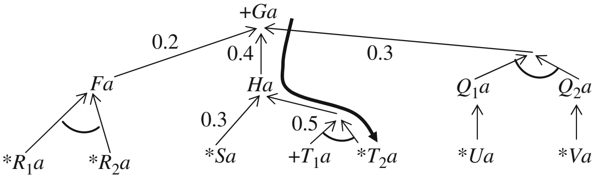


Fig. 4.3 Search space for a factual abduction problem. + indicates a known fact, * indicates possible abductive hypotheses. *Labelled* nodes correspond to literals, unlabeled nodes represent conjunctions of them, directed edges (arrows) correspond to laws in $L[x]$. The numbers are probability values (unknown residual probability). The bold arrow indicates the route of a best-first search

Definition: $\text{abd}(K, G[a]) = E[a] - F[a]$, where $E[a]$ is a *best-first possible explanation* of e given K as defined above, relative to some probability measure over conjunctions of literals.

We turn to the case of (iterated) factual abductive belief revision. The given belief systems is an expansion of the original belief system $K_0 := \text{Cn}(L[x] \cup F[a])$ by a set of goal-facts $G[a]$ which are or were in need of explanation, and a set of abductive hypotheses $\text{abd}(K, G[a])$ which together with $L[x] \cup F[a]$ explain the facts in $G[a]$ (hence, the goal-facts do not figure as explanatory premises for other goal-facts). We take base-facts, laws as well as the goal-facts as *granted* and revise only the abductive hypotheses in the light of new evidence. If a new evidence e contradicts K, some hypotheses in abd(K,G[a]) have to be revised in a way which covers the old explanandum facts $G[a]$ as well as the new evidence e. Consider the simple case $G[a] = \{e_1\}, e = e_2$, and $K \Vdash\!\!- \neg e_2$; hence $L[x] \cup F[a] \cup \text{abd}(K, e_1) \Vdash\!\!- \neg e_2$. In other words, the first-best explanation for e_1, $\text{abd}(K, e_1)$, was falsified by the new evidence e_2. Does there exist an incremental revision operation, for example, by backtracking to the second-best explanation for e_1? The general answer is no: apart from exceptional cases, the second-best explanation for e_1 will not be expandible to a first-best explanation for $e_1 \wedge e_2$. Rather, the first-best explanation for $e_1 \wedge e_2$ has to be searched once again from scratch. For example, the 1st and 2nd best explanations of the fact that person A was murdered may be that Ms. B or Ms. C were the murderer, but the first-best explanation of the facts that Mr. A was murdered by being strangled is that Mr. D was the murderer because Ms. B and C are not strong enough to strangle Mr. A. Thus, generally speaking, iterated factual abduction is a case of ab initio revision in the sense of (21). We summarize this as follows:

(28) *Abductive belief revision* (for iterated factual abduction):

Format and assumptions: $K = K_0 + +G = K_0 + G + \text{abd}(K_0, G)$, with $K_0 = \text{Cn}(\text{L}[x], F[a])$, where the elements of $L[x]$, $F[a]$ and G are taken as granted. The new evidential *input* is $e := e[a]$ such that $L[x] \cup F[a] \cup \text{abd}(G, K_0)$ derives $\neg e$.

Belief revision proceeds ab initio: Remove $\text{abd}(G, K_0)$. Generate $\text{abd}(K_0 - \neg e, G \cup \{e\})$ instead. $K^{**}e$ is defined by (21), as $(K_0 - \neg e) + (G \cup \{e\}) + \text{abd}(K_0 - \neg e, G \cup \{e\})$.

Only if the new explanandum e is explanans-separated from G in the sense of Section 4.3.2, one need not proceed by ab initio revision but may just *add* the new explanation $\text{abd}(K - \neg e, e)$ to the old explanations $\text{abd}(G, K_0)$.

4.3.5 Theoretical Model Abduction in Science

This kind of abduction generates a model within the framework of a given theory T which explains a possibly complex empirical phenomenon described by a (possibly complex and general) empirical sentence e. This situation is different from the situation of factual abductions, insofar one does *not* face the problem of a huge multitude of possible explanatory conjectures. The given theory T (which is represented by the set of its *axioms*) *constrains* the space of possible causes to a small class of basic parameters (or generalized 'forces') by which the theory models its intended applications. Every (mature) scientific theory is associated with a typical abduction pattern which specifies the kind of explanatory conjectures which ought to be sought for phenomena in T's domain of intended applications. In Schurz (2008a, §5) the abduction pattern of Newtonian mechanics is explicated as follows:

(29) *Abduction/Explanation Pattern of Newtonian Particle Mechanics:*

Explanandum e: a kinematical process involving (a) some moving particles whose time-dependent trajectories are known by observation, and (b) certain objects defining constant boundary conditions (e.g. a rigid plane on which a ball is rolling, or a large object which exerts a gravitational force, or a spring with Hooke force, etc).

=====================================

Generate the abduced conjecture as follows: (1) specify for each particle its mass and all non-neglectible forces acting on it in dependence on the boundary conditions and on the particle's position at the given time, (2) insert these specifications into Newton's 2nd axiom (sum-of-forces = mass times acceleration), and (3) solve the resulting system of differential equations such that the resulting predicted trajectories e_{pred} fit the given trajectory-observations e.

The explanatory conjectures in (1) describe what one calls a 'theoretical model', whence we speak of 'theoretical model abduction'.

Now let us turn the *revision* of theoretical models. The typical revision instruction associated with Newtonian mechanics can be explicated as follows:

(30) *Revision instruction for Newtonian particle mechanics:*

Given: e and e_{pred} as described in the Newtonian abduction pattern (29).

Assume new trajectory-observations e' produce a modified explanandum evidence $e^* = e \wedge e'$ which does no longer agree with e_{pred}.

Proceed as follows: (a) search for additional 'perturbing' forces (or boundary conditions) which have been overlooked so far; (b) add them to (1) in (29) above and proceed with (2) and (3) of (29) until a new predicted trajectory e^*_{pred} is generated which fits e^* sufficiently well.

Addendum: If the search is successful, T is strongly confirmed. Otherwise scientists will ask for revisions of T's core axioms.

Theoretical model revision proceeds more-or-less *incrementally* – only certain peripherical parts of T are modified. Thus we have $T = T_1 \cup T_2$ and $T^* = T_1 \cup \text{rev}(T_2, e', K - \neg e')$, where T_2 is a peripherical part of T which contains auxiliary assumptions about special theoretical models – recall the examples in (7) and (8) of Section 4.1.2. More generally, the revised part of T is located by a T-associated (partial) *preference* ordering over T's axioms.

I conclude this chapter with an attempt to formulate a generalized expansion and revision instruction for (deductive) theories about *dynamical systems* which work with differential (or difference) equations, as in mathematical physics, chemistry, or evolutionary biology. The *format* of such theories is explained in (31) below. Index of importance ranks from low to high: while theoretical model assumptions can easily be changed, central theoretical axioms (force laws) or explanatory promises concerning domains of application (cf. Kitcher 1981, 510ff) cannot be changed so easily. Derived consequences don't belong to *T*'s axiomatic part and, hence, are not ranked.

(31) *Format of (deductive) theories T describing dynamical systems:*

Structure and ranking of classes of axioms of T: index of importance

1. *Applicational part:* Consists of a *list* of several (types of) applications, i.e. empirically described systems S_i $(1 \leq i \leq n)$. 2-3

2. *Theoretical part:*

2.1 *Auxiliary (theoretical model) hypotheses* $A_k\,(1 \leq k \leq m;\ \text{for}\, m < n):$ 1

Each A_k describes a system (type of application) in terms of the theory. It consists of a ceteris-paribus-law cp(L_k) which asserts that *all* (non-neglectible) forces which act within the described system are contained in a list L_k which is listed by A_k.

2.2 *Special force laws* (e.g., gravitational force $= \gamma \cdot m_1 \cdot m_2/r^2$). 2

2.3 *General differential (or difference) equation(s)* with the variable expression 'sum-of-all-forces' (e.g. (2) of (29)). 3

Derived empirical consequences: predicted trajectories $e_{\text{pred},\,i}$

Every explanation provided by a theory of a dynamical system needs a cp-law because the 'heart' of the theory is a general differential (or difference) equation which is formulated in terms of the 'sum of all forces acting within the system'. The cp-law lists a couple of forces and asserts that these are indeed *all* forces; further perturbing forces are excluded (cf. Schurz 2002, §6).

Based on the example in (29) we explicate belief expansion for the kind of theories described in (31) by the following instruction:

(32) *Abductive belief expansion by models of theories about dynamical systems:*

Format and assumptions: $K = \text{Cn}(T \cup E)$ where *E* is the set of empirical phenomena explained by *T*. The new empirical input *e* describes a new empirical phenomenon.

Expansion instruction: abd(*K*,*e*) is a set of new auxiliary hypotheses describing a *T*-theoretical model about *e*, such that abd(*K*,*e*) explains *e* within *T* according to the *T*-associated abduction/explanation pattern. (*T* is expanded to $\text{T} + \text{abd}(K, e)$.)

This 'instruction' is not a full 'algorithm' because the definition of '*T*-associated abduction patterns' is left open.

T's theoretical model assumptions have to be revised when new evidences e' come in conflict with *T*'s empirical predictions e_{pred} – recall the Uranus–Neptune and the sickle cell examples in Section 4.1.2. We explicate the belief revision instruction for that kind of situation as follows:

(33) *Abductive belief revision for models of theories about dynamical systems*:

Format and assumption: $K = \mathrm{Cn}(T \cup \{e\})$. T entails e_{pred} which fits the observed data e, but new data e' entail $\neg e_{\mathrm{pred}}$. Proceed according to the following *instructions*:

(1.) Identify the auxiliary hypotheses A_k, i.e. the associated list L_k of assumed forces and the cp-law $\mathrm{cp}(L_k)$ which were needed for deriving e_{pred} within T. *Note:* Dynamical theories provide unique causal scenarios; they do not admit causal overdeterminations. Therefore, the unique list of assumed forces which wrongly 'explain' e within T is usually easily identified.

(2.1) *Either* e is explanans-separable from E, which means that A_k is merely relevant for the new application e (example: A_k asserts that for Uranus the only significant force is the gravitational force of the sun): then simply remove A_k.

(2.2) *Otherwise* A_k holds for other applications as well (example: A_k asserts that for all planets the only significant force is the gravitational force of the sun). In this case, restrict A_k by adding an exception clause to its antecedent that excludes the empirical (type of) system S_i about which the conflicting evidence e' speaks (e.g., Uranus) from the range of applications of A_k; call the result $A_{k,\mathrm{restr}}$ (for 'restricted'). Copy the list L_k into a new list $L_{k,i}$ (specially designed for system type S_i).

(3.) Try to *expand* $L_{k,i}$ to $L^*_{k,i}$ (in case 2.1, L_k to $L_k{}^*$) by searching for further (overlooked) forces which act in the system S_i , and add the cp-law $\mathrm{cp}(L^*_{k,i})$ (in case 2.1 $\mathrm{cp}(L^*_k)$), such that the new total evidence $e^* := e \wedge e'$ is (approximately) derivable from the so-revised theory $T^* = (T - \{A_k\}) \cup \{A^*_k\}$, where in case (2.1), $A^*_k = \mathrm{cp}(L^*_k)$, and in case (2.2), $A_k = A_{k,\mathrm{restr}} \wedge \mathrm{cp}(L^*_{i,k})$. Set $\mathrm{rev}(T, e', K^* - \neg e') := \mathrm{T}^*$.

Note: Only in case (2.1), theory revision is representable by Levi-identity, by assuming that $K — \neg e' = \mathrm{Cn}((T - \{A_k\}) \cup \{e\})$. In case (2.2) this is impossible because the revision modifies $\mathrm{A_k}$ which would have been forgotten in the contraction step.

If step (3.) is successful, the given theory in its new version T^* is confirmed. But if step (3.) fails repeatedly, scientists will attempt to revise central parts of T. This is no longer theoretical model revision but theory core revision, which doesn't lead to a new version of the same theory, but to a different theory.

Acknowledgement For valuable help I am indebted to Hannes Leitgeb, Erik Olsson, Issac Levi and Hans Rott.

References

Alchourrón, C.E., Peter Gärdenfors, and David Makinson. 1985. On the logic of theory change; partial meet contraction and revision functions. *Journal of Symbolic Logic* 50:510–530.

Aliseda, Atoacha. 2006. *Abductive reasoning*. Dordrecht: Springer.

Bratko, I. 1986. *Prolog programming for artificial intelligence*. Reading, MA: Addison-Wesley.

Carnap, R. 1950. *Logical foundations of probability*. Chicago: University of Chicago.

Earman, J. 1992. *Bayes or Bust?* Cambridge, MA: MIT.

Flach, P., and A. Kakas. (eds.). 2000. *Abduction and induction*. Dordrecht: Kluwer.

Forster, M., and E. Sober. 1994. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science* 45:1–35.

Gärdenfors, Peter. 1981. An epistemic approach to conditionals. *American Philosophical Quarterly* 18:203–211.

Gärdenfors, Peter. 1988. *Knowledge in flux*. Cambridge, MA: MIT.

Hansson, Sven O. (1999): *A textbook of belief dynamics*. Dordrecht: Kluwer.

Hansson, Sven O. 2006. Logic of belief revision, In *Stanford encyclopedia of philosophy*. plato.stanford.edu/entries/logic-belief-revision/.

Josephson, J., and S. Josephson. (eds.). 1994. *Abductive inference*. New York, NY: Cambridge University Press.

Kitcher, P. 1981. Explanatory unification. *Philosophy of Science* 48:507–531

Levi, Isaac. 1977. Subjunctives, dispositions, and chances. *Synthese* 34:423–455.

Levi, Isaac. 1980. *The enterprise of knowledge*. Cambridge, MA: MIT.

Levi, Isaac. 1991. *The fixation of belief and its undoing*. Cambridge, MA: Cambridge University Press.

Levi, Isaac. 2004. *Mild contraction*. Oxford: Clarendon.

Martin, Eric, and Daniel Osherson. 1998. *Elements of scientific inquiry*. Cambridge, MA: MIT.

Olsson, Erik J., and David Westlund. 2006. On the role of research agenda in epistemic change. *Erkenntnis* 65:165–183.

Páez, Andrés. 2006. *Explanations in K. An analysis of explanation as a belief revision operation*. Oberhausen: Athena.

Pagnucco, Maurice. 1996. *The role of abductive reasoning within the process of belief revision*, Dissertation, University of Sydney.

Paul, G. 1993. Approaches to abductive reasoning. *Artificial Intelligence Review* 7:109–152.

Peirce, C.S. 1878. Deduction, induction, and hypothesis. Peirce (CP) 2.619–2.644.

Peirce, C.S. 1903. Lectures on pragmatism. Peirce (CP) 5.14–5.212.

Peirce, C.S. (CP). *Collected papers*, ed. C. Hartshorne, P. Weiss, 1931–1935. Cambridge, MA: Harvard University Press.

Ridley, M. 1993. *Evolution*. Oxford: Blackwell.

Rott, H. 1992. *Reduktion und revision*. Bern: P. Lang.

Rott, H. 1994. Zur Wissenschaftsphilosophie von Imre Lakatos. *Philosophia Naturalis* 31(1):25–62.

Rott, H. 2000. Two dogmas of belief revision. *Journal of Philosophy* 97:503–522.

Rott, Hans, and Maurice Pagnucco. 2000. Severe withdrawal (and recovery). *Journal of Philosophical Logic* 29:501–547.

Salmon, W. 1989. *Four decades of scientific explanation*. Minneapolis, MN: University of Minnesota Press.

Schurz, G. 1991. Relevant deduction. *Erkenntnis* 35:391–437.

Schurz, G. 1995/96. Scientific explanation: A critical survey. *Foundation of Science* I(3):429–465.

Schurz, G. 2002. Ceteris paribus laws: Classification and deconstruction. In *Ceteris paribus laws,* eds. J. Earman, C. Glymour, and S. Mitchell. (2002, Hg.). *Erkenntnis* 57(3):351–372.

Schurz, G. 2008a. Patterns of abduction. *Synthese* 164:201–234.

Schurz, G. 2008b. Third-person internalism: A critical examniation of externalism and a foundation-oriented alternative. *Acta Analytica* 23:9–28.

Schurz, G., and K. Lambert. 1994. Outline of a theory of scientific understanding. *Synthese* 101(1):65–120.

Sneed, J.D. 1971. *The logical structure of mathematical physics*. Dordrecht: Reidel.

Chapter 5
A Structuralist Framework for the Logic of Theory Change

Sebastian Enqvist

5.1 Introduction

5.1.1 The Purpose of This Essay

The classical framework for the logic of theory change is the so-called AGM model, named after its creators Alchourron, Gärdenfors and Makinson (see Gärdenfors 1988 for a detailed introduction). The basic idea of the AGM model is to view a theory (or an epistemic state) essentially as a logically closed sets of sentences (the statements of the theory, often viewed as the *beliefs* of some agent), and to view theory changes (epistemic changes) as operations taking theories to theories, i.e. logically closed sets of sentences to logically closed sets of sentences. Theory changes are divided into three sorts: expansion, revision and contraction. Expansion simply adds information, contraction removes information. Revision adds information under the proviso that whenever necessary, enough information is removed to ensure that the new theory is consistent.

Formally, a theory (or *belief set*) is taken to be any logically closed set of sentences in a given language. If K is a theory and α is any sentence, then the result of expanding K with α, denoted $K + \alpha$, is given by the equation

$$K + \alpha = \mathrm{Cn}(K \cup \{\alpha\})$$

where Cn is a consequence operator over the language. The result of contracting α from K is denoted $K \div \alpha$, and is assumed to satisfy the following postulates, called the basic AGM postulates:

(Closure) $K \div \alpha = \mathrm{Cn}(K \div \alpha)$
(Inclusion) $K \div \alpha \subseteq K$
(Vacuity) If $\alpha \notin K$, then $K \div \alpha = K$

S. Enqvist (✉)
Department of Philosophy, Lund University, Kungshuset Lundagård 222 22 Lund
e-mail: Sebastian.Enqvist@fil.lu.se

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_5,

(Success) If α is not a tautology, then $\alpha \notin K \div \alpha$
(Recovery) If $\alpha \notin K$, then $K \subseteq (K \div \alpha) + \alpha$
(Extensionality) If α and β are logically equivalent, then $K \div \alpha = K \div \beta$

In order to actually construct a function which satisfies these postulates, a variety of strategies have been presented. The most common one is to associate with every theory a preorder over the language, a so-called *entrenchment order*. If $\leq$ is an entrenchment order for a theory K, then it is assumed to obey the following postulates:

(Transitivity) If $\alpha \leq \beta$ and $\beta \leq \chi$, then $\alpha \leq \chi$
(Dominance) If $\beta \in \mathrm{Cn}(\{\alpha\})$, then $\alpha \leq \beta$
(Conjunctiveness) Either $\alpha \leq \alpha \wedge \beta$ or $\beta \leq \alpha \wedge \beta$
(Minimality) If K is consistent, then $\alpha \notin K$ iff $\alpha \leq \beta$ for all β
(Maximality) If $\alpha \leq \beta$ for all α, then $\beta \in \mathrm{Cn}(\emptyset)$

Given a theory K together with an entrencment order $\leq$, we can construct a contraction function satisfying the basic AGM postulates, by the following equivalence (writing $\alpha < \beta$ as a shorthand for $\alpha \leq \beta$ and *not* $\beta \leq \alpha$):

$$\beta \in K \div \alpha \text{ iff } \beta \in K \text{ and either } \alpha \in \mathrm{Cn}(\emptyset) \text{ or } \alpha < \alpha \vee \beta$$

With expansion and contraction at our disposal, we can define the revision of a theory K with the sentence α, denoted $K^*\alpha$, by the following equation, called the *Levi identity*:

$$K^*\alpha = (K \div \alpha) + \alpha$$

This framework has in no way stood undisputed since its creation. For instance, the assumption that epistemic changes are *functions* on epistemic states, i.e. that every revision or contraction should be uniquely determined by the input and the epistemic state, has been challenged, giving rise to so-called *relational* belief revision. One argument in this direction, due to Lindström and Rabinowicz (1991), is based on rejecting one of the consequences of the above postulates for entrenchment, that the entrenchment order is connex, i.e. every pair of sentences is comparable w.r.t. entrenchment. The condition of logical closure on epistemic states has likewise been called into question, and has resulted in *belief base dynamics*. A recent suggestion is to equip every epistemic state with a set of *questions* that the agent wants answers to, a *research agenda* (Olsson and Westlund 2006).

Apart from various suppositions of the AGM theory that have been rejected by some authors, giving rise to somewhat diverging frameworks for the study of theory change, some different formal tools have also been tried. For instance, Adam Grove has suggested a modelling of epistemic change based on a space of possible worlds, where a theory is represented as the set of worlds that are in compliance with the theory, and likewise for individual propositions (Grove 1988). Dynamic Doxastic Logic (DDL), due to Krister Segerberg, is another alternative formal framework, where theory changes are modelled in a dynamic modal logic (see Segerberg 1998 or Segerberg and Leitgeb 2007).

All these variants of the logic of theory change are similar in one respect: they all agree that theories can, more or less, be equated with sets of statements, the statements that the theory claims are true. In this essay, I will try to develop a framework for the logic of theory change which is based on a different conception of theories. The basis of the framework will be structuralism's notion of a *theory net*.

According to structuralism, theories are not representable simply in terms of sets of statements; rather, theories are model theoretical constructions with a rather intricate structure, and these can then be *used* to make statements about the world. The structuralist model of theories is impressive in two respects: first, it presents a very detailed analysis of what may be called the *deep structure* of an empirical theory. Second, it has been shown that a range of actual scientific theories can be reconstructed as theory nets. This is the reason why I have chosen structuralism as the basis for my framework: the richness of the structuralist representation of theories will hopefully enable us to raise new and interesting questions about theory change, and also, it will allow us to ground the logic of theory change in an empirically adequate notion of "theory".

5.1.2 The Basic Features of Structuralism

The fundamental idea in structuralism has come to be known as the "non-statement" view of theories. The idea is that theories, per se, are not statements or sets of statements; rather they are mathematical structures, of a particular kind, that are used to make statements about the world. A theory can be said to consist of two components: a *theory core*, which describes a class of structures using the axioms of the theory, and a class of *intended applications*, which can be thought of as a collection of chunks of the real world to which the theory is supposed to apply. The empirical claim associated with a theory is, roughly, that the intended applications of the theory can be "embedded" into the core of the theory.

The basic notion in structuralism is that of a theory element. Theory elements are the smallest type of theories that structuralism speaks about. A theory element E is a pair $\langle K(E), I(E)\rangle$, where $K(E)$ is a theory core, and $I(E)$ is a class of intended applications.

The core of a theory element contains a number of components. First, it contains a class $M(E)$, which is called the class of *models* for the theory element. The models of a theory core are the structures where the axioms of the theory are true; rather than identifying a theory with its axioms, we thus identify it with the model-theoretical *content* of the axioms. Theories are supposed to be invariant under specific axiomatizations – if two theories are formulated differently, but from a model-theoretical point of view have the same content, then they are taken to be the same theory.

The second element of a theory core is intended to take into account the fact that theories are formulated within some *conceptual framework*. This component of the theory core, denoted $M_{\mathrm{P}}(E)$, is called the class of *potential models* for the theory.

The potential models of a theory can be thought of as any kind of structure that contains everything which is needed for it to be meaningful to ask: is this a model for the theory? It is assumed that for any theory element E, we have

$$M(E) \subseteq M_{\mathrm{P}}(E).$$

For instance, in the core of classical particle mechanics, the *potential* models are those structures that contain real valued functions that can be taken as interpretations for the concepts of mass, force and so on, and the *models* are those potential models that actually satisfy the Newtonian axioms for these functions. At this point, the empirical claim associated with the theory element is simply that

$$I(E) \subseteq M(E),$$

i.e. that all the intended applications of the theory are models for the theory.

Once we take the conceptual framework of a theory into account, a distinction can be made between the *theoretical* and the *non-theoretical* concepts of the theory. This distinction is notoriously hard to make in a precise way. However, structuralism goes some way in solving the problem: while the distinction between theoretical and non-theoretical concepts has traditionally been made in an *absolute* way, structuralism explicates this notion in a *theory-dependent* way. The idea is to say that a function is *non-theoretical* w.r.t the theory E, or "E-non-theoretical" if it can be measured without reference to any formerly succesful application of the theory E. It is theoretical, or "E-theoretical", if it is *not* non-theoretical. This distinction gives rise to a new component of the theory element, denoted $M_{\mathrm{PP}}(E)$. This is called the class of *partial potential models* of E. Intuitively, the partial potential models of a theory are those structures that contain interpretations for all the non-theoretical components of the conceptual framework. It is usually assumed that

$$I(E) \subseteq M_{\mathrm{PP}}(E).$$

Here, we have an apparent regress. For if we want to test a theory E against some empirical data concerning an application of the theory, then we need to determine the values of the theoretical functions of the theory in this application. But in order to do this, by the definition of "theoretical function", we have to make reference to some former succesful application of the theory. But in order to determine whether this application of the theory was "successful", we need to determine the theoretical values in this case too – and so on. Thus, there seems to be no way of testing the empirical claim of the theory.

This problem is solved by introducing a modified version of the empirical claim of a theory. Instead of claiming that the intended applications of the theory are also models for the theory, we say that they can be *enriched* with theoretical components to form models. For any partial potential model x, we say that a potential model y is an *enrichment* of x if y results from adding theoretical components to x. If this is the case, then we write $x \mathrm{\,e\,} y$. If each member of the class of partial potential models

X can be enriched to form some element of the class of potential models Y, then we write X e Y. The modified empirical claim of our theory element E is then as follows:

$$\exists X(X \subseteq M_{\mathrm{P}}(E) \wedge X \subseteq M(E) \wedge I(E) \text{ e } X)$$

This is called the *Ramsey-Sneed claim* of the theory E. Intuitively, it says that each intended application of the theory can be enriched to form a model for the theory. With this type of claim, the above mentioned regress disappears: we no longer have to determine the values of theoretical functions of a theory in order to determine whether some application is succesful, we only need to check that the application can be enriched, in some way, with values for the theoretical functions, so that the result is a model for the theory.

This kind of claim is usually not enough. We have, at this point, no way of transferring information about the theoretical functions from one application to another. This problem is solved by introducing the concept of a *constraint* on the theoretical functions of a theory. An intuitive example of a constraint is, for instance, that in any two applications for classical particle mechanics, if the same object occurs in the domain of two distinct applications, it should have the *same mass* in both instances. Formally, a constraint for a theory E can be taken to consist of a class of subclasses of the potential models of E. It is to be thought of as the class of classes of potential models of the theory that respect the constraint. That is, if C is a constraint for E, then

$$C \subseteq \wp(M_{\mathrm{P}}(E)).$$

Taking the intersection of all constraints of E, we get what is called the *global constraint* of E, denoted GC(E). At this point, the core $K(E)$ is a quadruple

$$\langle M_{\mathrm{P}}(E), M_{\mathrm{PP}}(E), \mathrm{GC}(E), M(E)\rangle,$$

where $M_{\mathrm{P}}(E)$ is a class of potential models, $M_{\mathrm{PP}}(E)$ is a class of partial potential models, GC(E) is the global constraint, and $M(E)$ is a class of models. The empirical claim of the theory now has the following form:

$$\exists X(X \subseteq \mathrm{M}_{\mathrm{P}}(E) \wedge X \subseteq M(E) \wedge X \in \mathrm{GC}(E) \wedge I(E) \text{ e } X)$$

It says that the intended applications of the theory can be enriched to form models for the theory, in a way that *respects* the global constraint.

Two further elements are usually taken into account. The first is a set of *intertheoretical links* for E. A link is a relation between the potential models of E and the potential models of some other theory; links are intended to account for the fact that the content of a theory can essentially involve its connections to other theories. Just like constraints allow us to transfer information from different applications of one and the same theory, links allow us to transfer information between applications of *different* theories. The links of a theory element E are usually "lumped together" to form a *global* link GL(E), in a way which is analogous to the construction of the global constraint GC(E).

Finally, it is noted that in reality, a perfect match between a theory and the world is not required in order to regard an application of the theory as successful. We usually accept a certain amount of *approximation* – it is enough that the theory *almost* fits the facts. For this reason, an additional component to the core of a theory element E is added, called the class of *admissible blurs* for E. We will denote it by $A(E)$. The formal construction of admissible blurs involves the topological notion of a *uniformity*, and it is rather complex. We will not introduce it here.

At this point, a theory element E is a pair $\langle K(E), I(E)\rangle$, where $I(E)$ is a class of intended applications, and

$$K(E) = \langle M_{\mathrm{P}}(E), M_{\mathrm{PP}}(E), \mathrm{GC}(E), M(E), \mathrm{GL}(E), \mathrm{A}(E)\rangle$$

is a theory core. With the notion of theory element in place, we will now introduce a *second* notion of theory in the structuralist framework, the notion of a *theory net*. This is the notion of theory that will form the basis for the framework for the logic of theory change which will be developed in Section 5.3.

First, we introduce the notion of *specialization*. Specialization is a relation over the class of theory elements, and if one element stands in this relation to another, then we say that the latter is a specialization of the former. Intuitively, a specialization of a theory element is the result of adding special assumptions, or *special laws* (and possibly also special constraints and links) to the theory core of the element, that are intended to hold only for a restricted subclass of its intended applications. Formally, we say that E_2 is a specialization of E_1, and write $E_1\ s\ E_2$, if and only if

$$\begin{array}{l}
\text{(i) } M_{\mathrm{P}}(E_2) = M_{\mathrm{P}}(E_1)\\
\text{(ii) } M_{\mathrm{PP}}(E_2) = M_{\mathrm{PP}}(E_1)\\
\text{(iii) } M(E_2) \subseteq M(E_1)\\
\text{(iv) } \mathrm{GC}(E_2) \subseteq \mathrm{GC}(E_1)\\
\text{(v) } \mathrm{GL}(E_2) \subseteq \mathrm{GL}(E_1)
\end{array}$$

It is easily shown that the specialization relation satisfies the properties of a partial ordering. A theory net is then a finite set of theory elements, usually assumed to have a least element under the specialization relation (such nets are called "tree-like"), and such that every element the net is connected to some other element by the specialization relation. The least element of the net can be thought of as the most "fundamental" theory element. Theory nets are, probably, the most "natural" level of speaking about theories; they take into account the fact that the laws of one and the same theory are more or less fundamental in the sense that their scope can be wider or narrower; a theory contains not only general laws, but also special laws with restricted scopes. The empirical claim of a theory net can be taken to be simply the conjunction of the claims corresponding to each of its elements.

Finally, structuralism speaks of theories on another, still larger level: these theories are called *theory holons*, and a theory holon can be thought of as the *total state of science* at some particular time. We do not introduce them formally here. For more on structuralism, see Balzer et al. (1987), Balzer and Moulines (1996) or Sneed (1971).

5.2 Preliminary Discussion

5.2.1 Expansion

Let us begin to think about how we may want to change a theory, when theories are represented in terms of structuralistic theory nets. For the purpose of our discussion, let us use an artificial theory with a very simple structure. The theory involves no distinction between theoretical and non-theoretical terms, no constraints or links, etc. The theory elements of the net are simply pairs of a class of models together with a class of intended applications. When we turn to develop our formal framework, our attention will be restricted to theories of this simple kind, leaving all the excluded elements aside from consideration.

The language of the theory involves a single function "$\otimes$" defined on some domain D such that $\otimes$: $D \times D \rightarrow D$, i.e. $\otimes$ is a function from pairs of objects to objects in a given domain. We will also make use of a constant "e". We shall now construct a theory net based on this simple framework. For convenience, we represent each core in terms of the axioms it states for $\otimes$ and e. A possible model for any theory element in our net can be taken to be any set with a binary function defined on it (i.e. a magma), with an intepretation of the constant e. Let E_1 be an element with the following axiom:

$$\text{A1: } \forall x \forall y \exists z (x \otimes y = z)$$

Let $K(E_1)$ denote the core of E_1 – to be identified, for now, with the logical closure of its axioms – and let $I(E_1)$ denote the class of intended applications for E_1. We will use this notation throughout this section. Now, a simple way of *expanding* E_1 is to add another axiom to the core, which is then taken to hold for the same class of intended intended applications. Say, for instance that we add the following axiom to $K(E_1)$:

$$\text{A2: } \forall x \forall y \forall z ((x \otimes y) \otimes z = x \otimes (y \otimes z))$$

Then the empirical claim of E_1, after the expansion, is that every intended application in $I(E_1)$ is a magma where the operator $\otimes$ is *associative*.

At this point, our theory is a one-element net N_1 with the single class of intended applications $I(E_1)$. We can expand this net in a second way, by adding a new axiom, which is intended to hold only for a restricted subclass of the applications $I(E_1)$.

In this case, we leave the element E_1 unchanged, and instead add another element E_2, which has a stronger core but a narrower class of intended applications. For instance, say we construct the core of E_2 by adding the following axiom to the core of E_1:

$$\text{A3: } \forall x(x \otimes e = e \otimes x = x)$$

We let $I(E_2)$ be properly contained in $I(E_1)$. The result is then a two-element net $N_2 = \{E_1, E_2\}$, where E_2 is a *specialization* of E_1, and

$$K(E_1) = \mathrm{Cn}(\{\mathrm{A1}, \mathrm{A2}\})$$
$$K(E_2) = \mathrm{Cn}(\{\mathrm{A1}, \mathrm{A2}, \mathrm{A3}\})$$

We can represent N_2 with the diagram shown in figure 5.1, where the downwards arrow represents the specialization relation:

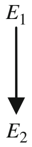


Fig. 5.1 The net N_2

The empirical claim of N_2 is that all structures in $I(E_1)$ are groupoids where the operator $\otimes$ is associative, and furthermore, that within all structures in the narrower class $I(E_2)$, the element e is an *identity* for the operator $\otimes$.

Just like in AGM, whenever we expand, either by addition or by specialization, there is always the risk of running into inconsistency. When this happens, we want to remove something in order to resolve the inconsistency; we want to *revise* rather than just expand. In order to do this, we need a method of contraction – we turn to this now.

5.2.2 Contraction

Consider the three element net $N_3 = \{E_1, E_2, E_3\}$ where E_1 and E_2 are as before, and E_3 is a specialization of E_2 in which the axiom

$$\text{A4: } \forall x \exists y(x \otimes y = y \otimes x = e)$$

is added to those of E_2. N_3 is a linearly ordered net, where $I(E_3) \subset I(E_2) \subset I(E_3)$, and

$$K(E_1) = \mathrm{Cn}(\{\mathrm{A1}, \mathrm{A2}\})$$
$$K(\mathrm{E}_2) = \mathrm{Cn}(\{\mathrm{A1}, \mathrm{A2}, \mathrm{A3}\})$$
$$K(\mathrm{E}_3) = \mathrm{Cn}(\{\mathrm{A1}, \mathrm{A2}, \mathrm{A3}, \mathrm{A4}\})$$

Diagrammatically, N_3 looks as follows (Fig. 5.2):

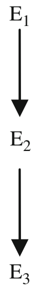


Fig. 5.2 The net N_3

A1–A4 are the axioms for *groups*, and so the empirical claim of the lowest element of our net is simply that the elements of $I(E_3)$ are groups. The element e governed by axiom A3 is called the *identity element*, and for any x, the element y guaranteed by A4 is called the *inverse* of x.

Now, it is well-known that for any group, the following so-called *cancellation laws* hold:

$$\text{CL1: } \forall x \forall y \forall z (x \otimes y = x \otimes z \rightarrow y = z)$$
$$\text{CL2: } \forall x \forall y \forall z (y \otimes x = z \otimes x \rightarrow y = z)$$

We also note that removing any one of A2, A3, or A4 from the axiom set A1–A4 yields a theory that has models where these laws fail.[1]

So let's say that we want to contract the sentence CL1 from the core of E_3. That is, we want to remove the claim that CL1 holds true for the intended applications of E_3. Then we have to remove one of the claims ascribing the axioms A1–A4 to the intended applications of E_3. It is sufficient to remove either one of A2–A3 from the core of E_3, so let's restrict our attention to those. Which one do we remove? In particular, to refer to a common principle in the logic of theory change, how do we contract in order to perform some kind of *minimal change*?

[1] Proof: we focus on CL1. As a model for {A1,A3,A4} where CL1 fails, take the set of natural numbers N, evaluate e as 0, and evaluate $\otimes$ as the function $\delta : N \times N \rightarrow N$ given by

$\delta(x, y) = 0$, if $x = y$,
$\delta(x, y)$ is the greatest number of x and y otherwise.

A1 is clearly satisfied, A3 is satisfied since if $x = 0$, then $\delta(x, 0) = \delta(x, x) = 0$, and if $x > 0$, then $\delta(x, 0)$ is the greatest number of x and 0, i.e. x. A4 is also satisfied; the inverse of any x is simply x itself. CL1 fails, for $\delta(4, 3) = \delta(4, 2) = 4$, but $3 \neq 2$ of course. As a model for {A1, A2, A4} where CL1 fails, take the set $\wp(N)$, and evaluate e as $\emptyset$, and $\otimes$ as $\cap$. A1 holds of course, A2 holds since intersections are associative, and A4 holds since the inverse of every set is simply $\emptyset$. CL1 fails, as $\{1\} \cap \{1, 2\} = \{1\} \cap \{1, 3\} = \{1\}$. Finally, as a model for {A1, A2, A3} where CL1 fails, take the set N and evaluate e as 1, and $\otimes$ as multiplication on natural numbers. A1 holds, A2 holds as multiplication is associative, and A3 clearly holds as well. But CL1 fails, since $0{*}1 = 0{*}2 = 0$, for instance. □

A first suggestion could be to use the specialization structure of the net N_3 to decide the matter. The idea is then to make changes as *low as possible* in the net. In the above case, this would mean contracting A4 – if we remove A3, then (given that we want to retain the specialization structure of the net), we would have to weaken not only E_3 but E_2 also, and if we remove A2, then we would have to change all the cores in the net.

Now, how does this proposed principle correspond to the notion of minimal change? We are using the specialization structure of a net as a basis for what to count as a minimal change here, and the intuition behind this strategy is that axioms that are introduced higher up in a net are in some sense more *fundamental* than those that essentially belong to lower parts of the net. But this is more than an intuition, since there is a perfectly straightforward sense in which it is true: the higher up an element lies in a net, the wider is its class of intended applications, and so the higher up in a net we find a certain axiom, the wider is its *scope*. Thus, it is the *generality* of an axiom that determines how fundamental we take it to be, and how seriously we regard the effect of removing it.

There are two connections one can make from all this: one is that the specialization ordering in a net is doing approximately the same job as entrenchment does in the AGM theory. However, while doing the same job, it is quite a different notion from that of entrenchment. The specialization structure of a net encodes the generality of certain principles, while entrenchment rather symbolizes the reluctance of some specific agent towards letting go of a belief. The other thing we may note is that the normative principle we stated above is essentially an explication of Imre Lakatos's idea of a *protective belt*. Lakatos sees theories as having a structure which is reminiscent of the structure of a theory net, with certain fundamental principles making up the *core* of the theory, and more specialized hypotheses that can be altered so as to *protect* the core from falsification (the protective belt). This is precisely what the principle tells us to do: while "cutting" as low as possible in the net, we save the higher elements so that they may be kept the same.

The principle we used here does not always give us as specific information for how to change a net as in the above case. For instance, it may be that a net does not have the convenient property of being *linearly ordered* like the net N_3 is. As an example, take the net N_4 shown in figure 5.3:

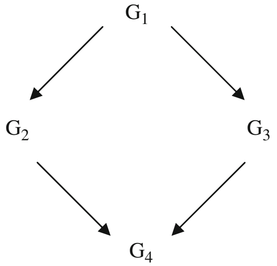


Fig. 5.3 The net N_4

where we have $I(G_4) \subset I(G_2) \subset I(G_1), I(G_4) \subset I(G_3) \subset I(G_1)$, neither of $I(G_2)$, $I(G_3)$ is a subclass of the other, and the cores correspond to our axioms as follows:

$$\begin{aligned} K(G_1) &= \mathrm{Cn}(\{\mathrm{A1}\}) \\ K(G_2) &= \mathrm{Cn}(\{\mathrm{A1}, \mathrm{A2}, \mathrm{A4}\}) \\ K(G3) &= \mathrm{Cn}(\{\mathrm{A1}, \mathrm{A3}, \mathrm{A4}\}) \\ K(G4) &= \mathrm{Cn}(\{\mathrm{A1}, \mathrm{A2}, \mathrm{A3}, \mathrm{A4}\}) \end{aligned}$$

Say we wish to contract CL1 from the core of G_4. In this case, our proposed principle of contraction gives us only partial guidance. It certainly tells us not to remove A1, since this is the only axiom which is present in the top element of the net. However, we are given no advice when it comes to deciding between A2, A3 and A4. Does this mean that we simply have to choose one of these arbitrarily?

In some cases at least, I think we can do better. We should try to preserve the most fundamental propositions of a theory, but this is not the *only* grounds we may have for wishing to keep some proposition in the course of a contraction. In particular, it may be that a certain proposition has been subjected to more empirical tests than another – for instance, if it has been part of the theory for a relatively long period in its course of evolution, and has survived empirical testing throughout this period. That is, it may be that certain propositions are more *corroborated* than others.

My suggestion is that we use corroboration as a second means for determining what to remove in the course of contraction. For instance, in the case above, if it turns out that both A3 and A4 are better corroborated than their rival A2 (or rather, the claims ascribing A3 and A4 to the intended applications of G_4 are better corroborated than the claim which states that A2 holds true for this class of intended applications), then we can make a principled decision of which one to remove: we should remove A2.

Once we take corroboration into account, an interesting problem presents itself. Let us go back to the linear net N_3, and let's imagine a possible story of the evolution of this theory net. *Prima facie*, there seems to be no reason to exclude the possibility that the axiom A2 came into the picture later in time than the other axioms; say, at one time, the highest element of the net had only the axiom A1, and A2 was later added to it by expansion (and thereby also to the cores of the two lower element). Now, if this previous theory had undergone some rigorous testing, but the new net N_3 has then not been tested as much, we might say that A2 is the least corroborated axiom among those of the bottom element E_3.

So let's go back to the problem of contracting CL1 from the core of E_3. We need to contract one of the axioms A1–A4, and it suffices to remove any of A2–A4. As we have already seen, the specialization structure of the net N_3 would urge us to remove A4. But if we instead decide to preserve the most *corroborated* axioms, then we should instead remove A2. This is thus a case where the two principles we have proposed for deciding what to keep and what not to keep in a contraction are in *conflict* with each other. Which of the two strategies we should follow seems to be something we cannot decide by purely a priori considerations – rather, in real cases, this would probably be decided by contextual factors, such as the current research interests at the time, the intended use of the theory, etc.

For one last remark, we note that there is another way in which we may change the empirical content of a theory net, apart from adding or removing axioms to a theory element, or generally, adding or removing the claim that a certain axiom holds true for some set of intended applications. As Peter Gärdenfors has noted (Gärdenfors 1992), we can also change the range of intended applications for some element in the net. Widening the range of intended applications makes the empirical claim of the theory element stronger, and narrowing it down weakens the empirical content of the element. This observation will be taken up in Section 5.3.3, where we will discuss the possibility of representing novel types of theory change within the structuralist framework, aside from the basic types of theory change in AGM.

5.3 Outlines of a Structuralistic Logic of Theory Change

5.3.1 The Formal Framework

The formal framework we will develop here is intended to be entirely semantical, i.e., we do not want to take the specific axiomatization of a theory into account – a theory should be equated with its model-theoretical content. In Section 5.2 we continuously referred to the axioms of a theory explicitly in the discussion – here, we will try to avoid any dependence on a specific linguistic representation of a theory.

For our purposes, the internal structure of a model or an intended application can be left out of consideration. Thus, since we want our framework to be as simple as possible, we will take as its basis simply a set U, the elements of which are intuitively to be thought of as *any structure a theory can talk about*. Propositions are taken to be subsets of this set; moreover, we will mimic the semantics for DDL (Dynamic Doxastic Logic) in that we assume a *Stone topology* over U, and we will take as *propositions* the clopen sets in this topology. Thus, the propositions form a Boolean algebra of sets – as we would want them to do, as long as we wish the propositions of the space to obey the laws of classical logic. Any Stone space $\langle U, T\rangle$ will in the following be referred to as a *structure space*. All the results and definitions stated should thus begin with the clause "for any structure space $\langle U, T\rangle\ldots$", but for brevity we usually do not state this explicitly.

We begin by defining the notion of *theory element* in this setting. The notion of a theory element that we use here will be considerably thinner than that used in structuralism; we will not take any account of the possible models of a theory, the distinction between theoretical and non-theoretical concepts, of constraints, links or admissible blurs. Basically, the core of a theory element will be equated with the models for the theory, and the possible models can be taken to be simply the entire space. This is not because these other aspects of theories are uninteresting in the present context – they certainly aren't. The model we develop here is only intended to be a first prototype for a structuralistic logic of theory change, and the task of developing it further to encompass all these elements is an important task for future research.

Definition 3.1.3 A *theory element* in $\langle U,T\rangle$ is a pair $E = \langle M(E), I(E)\rangle$, where $M(E)$ and $I(E)$ are both closed in $\langle U,T\rangle$, and $I(E)$ is non-empty.

We can now define the notion of specialization:

Definition 3.1.4 Let E_1,E_2 be any theory elements. We say that E_2 is a specialization of E_1, and write $E_2\ s\ E_1$, if

(i) $M(E_2) \subseteq M(E_1)$
(ii) $I(E_2) \subseteq I(E_1)$

Proposition 3.1.5 The specialization relation s is a partial ordering over the set of theory elements in $\langle U,T\rangle$.

We have written the specialization relation in the converse direction of how it is usually denoted. This is for transparency: with this way of writing it, *greater* elements will also be "higher" elements in the net, and "lower" elements will be *smaller*.

We can now define the notion of a theory net:

Definition 3.1.6 A *theory net N* is a non-empty finite set of theory elements $\{E_1, \ldots, E_n\}$ such that

(i) N has a greatest element under the specialization relation s,
(ii) if $E_i \in N$, then there is some $E_j \in N$ such that $E_i\ s\ E_j$ or $E_j\ s\ E_i$
(iii) Whenever $I(E_i) \subseteq I(E_j)$ for E_i,$E_j \in N$, we have $E_i\ s\ E_j$.
(iv) Whenever $M(E_i) \subseteq M(E_j)$ for E_i,$E_j \in N$, we have $E_i\ s\ E_j$.

We will need some way to identify the *empirical content* of a theory net, the statements that are asserted by someone who holds a certain theory. Let $\langle U, T\rangle$ be any structure space. An *empirical claim* in $\langle U, T\rangle$ is to be a pair consisting of a class of intended applications and a proposition, which is asserted to hold for that class according to the claim. Formally, a claim is a pair $\langle X, P\rangle$, where X is non-empty and closed in $\langle U, T\rangle$, and P is clopen. Whenever convenient, we will write claims as α, β, χ etc. We will denote the set of claims in a structure space S as CLM(S). If $\alpha = \langle X, P\rangle$ is a claim, then we call X the *range* of α, and P is called the *propositional content* of α.

We can define a simple logic (a logical closure operator) over the set of claims in a structure space. Let Γ be a set of claims, and let $\alpha = \langle X, P\rangle$ be a claim. Then we say that the set Γ *logically implies* α, and write $\Gamma \Rightarrow \alpha$, if there is a subset

$\{\langle X_i, P_i\rangle | i \in I\} \subseteq \Gamma$ such that

$$\text{(i) } X \subseteq \bigcap_{\text{i}\in\text{I}} X_i$$
$$\text{(ii) } \bigcap_{\text{i}\in\text{I}} P_i \subseteq P$$

Thus, $\Gamma \Rightarrow \alpha$ if the range of α is contained in the range of each member of some subset of Γ, and the propositional content of α contains the intersection of the logical contents of all members the same subset of Γ. In a sense, $\Gamma \Rightarrow \alpha$ means that the set Γ says at least as much as α about as least as large a chunk of the world as the range of α.

If Γ is a singleton $\{\beta\}$, then instead of $\beta \Rightarrow \alpha$, we write $\beta \Rightarrow \alpha$. We say that a claim $\langle X, P\rangle$ is *true* if $X \subseteq P$. Then clearly, if all claims in the set Γ are true, and $\Gamma \Rightarrow \alpha$, then α is true as well. Furthermore, we say that $\langle X,P\rangle$ is *trivially* true if $P = U$.

Let Γ be a set of claims in a given structure space S. Then we define the logical closure of Γ, denoted $\text{Cn}(\Gamma)$, by the following equation:

$$\text{Cn}(\Gamma) = \{\alpha \in \text{CLM}(S) | \Gamma \Rightarrow \alpha\}$$

Proposition 3.2.3 The operator Cn: $\wp(\text{CLM}(S)) \rightarrow \wp(\text{CLM}(S))$ satisfies the usual postulates for a logical closure operator:

$$\text{(i) } \Gamma \subseteq \text{Cn}(\Gamma)$$
$$\text{(ii) } \Gamma \subseteq \Phi \text{ implies } \text{Cn}(\Gamma) \subseteq \text{Cn}(\Phi)$$
$$\text{(iii)}\text{Cn}(\Gamma) = \text{Cn}(\text{Cn}(\Gamma))$$

Proof (i) and (ii) are straightforward. The only non-trivial verification is the right-to-left inclusion in (iii). Suppose that $\alpha = \langle X, P\rangle \in \text{Cn}(\text{Cn}(\Gamma))$ for some set of claims Γ. Then $\text{Cn}(\Gamma) \Rightarrow \alpha$. Hence there is a subset $\Phi \subseteq \text{Cn}(\Gamma)$, $\Phi = \{\langle X_i, P_i\rangle | i \in I\}$, such that $X \subseteq \bigcap_{i \in I} X_i$ and $\bigcap_{i\in I} P_i \bigcap \subseteq P$. But then $\Gamma \Rightarrow \beta$ for each $\beta \in \Phi$. Hence, for each i $\in I$, there is a subset $\Psi = \{Y_j, Q_j | j \in J_i\} \subseteq \Gamma$ such that $X_i \subseteq \bigcap_{j \in J_i} Y_j$ and $\bigcap_{j \in J} Q_j \subseteq Q_i$. Let

$$\Sigma = \cup\{\Psi \subseteq \Gamma | \Psi \Rightarrow \beta \text{ for some } \beta \in \Phi\}$$

We wish to show that $\Sigma \Rightarrow \alpha$. Let $\Sigma = \{\langle Z_k, R_k\rangle \,| k \in K\}$. We have

$$\text{(A)} X \subseteq \bigcap_{i\in I} X_i \subseteq \bigcap_{i\in I} \bigcap_{j\in J_i} Y_j = \bigcap_{k\in K} Z_k$$

$$\text{(B) } P \supseteq \bigcap_{i\in I} P_i \supseteq \bigcap_{i\in I} \bigcap_{j\in J_\text{i}} Q_j = \bigcap_{k\in K} R_k$$

It follows that $\Sigma \Rightarrow \alpha$, as desired. □

Proposition 3.2.4 $\mathrm{Cn}(\emptyset) = \emptyset$

This follows from our assumption that the range of any claim is non-empty. For our last fact about the consequence operator Cn, we say that a consequence operator is *finitary* if, for any set Γ and $\alpha \in \mathrm{Cn}(\Gamma)$, there is a finite subset $\Phi \subseteq \Gamma$ such that $\alpha \in \mathrm{Cn}(\Phi)$.

Proposition 3.2.5 For any structure space *S*, the consequence operator Cn over CLM(*S*) is finitary.

Proof let Γ be a set of claims and $\alpha = \langle X, P\rangle$ any claim such that $\alpha \in Cn(\Gamma)$, and let $\Psi \subseteq \Gamma$ be a witness to this fact. Then the range of α is contained in the intersection of all ranges of claims in Ψ. Hence, for any non-empty subset Φ of Ψ, the range of α is contained in the intersection of all ranges of claims in Φ. Thus, it suffices to find a non-empty finite subset $\Phi \subseteq \Psi, \Phi = \{\langle X_1, P_1\rangle, \ldots, \langle X_n, P_n\rangle\}$, such that $P_1 \cap \ldots \cap P_n \subseteq P$. This is equivalent to $P_1 \cap \ldots \cap P_n \cap -P = \emptyset$. But letting $\Psi = \{\langle X_j, P_j\rangle | j \in J\}$ we have $\bigcap_{j\in J} P_j \subseteq P$, i.e. $\bigcap_{j\in J} P_j \cap -P = \emptyset$. Since all P_j are clopen and hence closed, it follows by compactness of the Stone topology that there is a finite subset $\{j_1, \ldots, j_n\}$ of *J* such that $P_{j_1} \cap \ldots \cap P_{j_n} \cap -P = \emptyset$, as desired. □

Let *E* be a theory element in some structure space $\langle U,T\rangle$. Then we define the *empirical content* of *E*, denoted EC(*E*), by the equation

$$\mathrm{EC}(E) = \mathrm{Cn}(\{\langle X, P\rangle \in \mathrm{CLM}(S) | X \subseteq I(E) \text{ and } M(E) \subseteq P\})$$

Actually, the occurrence of the Cn-operator here is redundant, since a little thought shows that the set $\{\langle X, P\rangle \in \mathrm{CLM}(S) | X \subseteq I(E)$ and $M(E) \subseteq P\}$ is logically closed. If *N* is a net then we define the empirical content of *N*, denoted EC(*N*), by the equation

$$\mathrm{EC}(N) = \mathrm{Cn}(\cup\{\mathrm{EC}(E) | E \in N\})$$

The empirical content of a net is thus the logical closure of the set of elements of the empirical contents of its elements (here, the occurrence of the Cn-operator is *not* redundant).

If Γ is a set of claims, then we say that Γ is *logically closed* if $\mathrm{Cn}(\Gamma) = \Gamma$. Hence, for any theory element *E* and any net *N*, EC(*E*) and EC(*N*) are logically closed.

We will now begin by setting the basis for the fundamental epistemic actions of adding or removing a claim from the empirical content of a theory net (expansion and contraction). Revision can then be defined by the Levi identity. Every type of theory change will be represented as a function from empirical claims to relations over the set of theory nets in a given structure space. For instance, if N_1 and N_2 are nets, and α is a claim, then we write $\langle N_1, N_2\rangle \in R^+(\alpha)$ to say that "N_2 results from expanding the net N_1 with the claim α". R^+ thus denotes the function which we use to represent expansion.

We can translate the AGM characterization of expansion to the language of our formal framework as follows:

$$\text{(Expansion) If } \langle N_1, N_2 \rangle \in R^{+}(\alpha), \text{ then } \mathrm{EC}(N_2) = \mathrm{Cn}(\mathrm{EC}(N_1) \cup \{\alpha\})$$

In the same way, we let $R^{\div}$ be the function corresponding to contraction; thus, $\langle N_1, N_2 \rangle \in R^{\div}(\alpha)$ means: "N_2 results from contracting the claim α from the net N_1". We can then translate all the basic AGM postulates for contraction as follows:

(Closure) If $\langle N_1, N_2 \rangle \in R^{\div}(\alpha)$, then $\mathrm{Cn}(\mathrm{EC}(N_2)) = \mathrm{EC}(N_2)$
(Incusion) If $\langle N_1, N_2 \rangle \in R^{\div}(\alpha)$, then $\mathrm{EC}(N_2)) \subseteq \mathrm{EC}(N_1)$
(Vacuity) If $\langle N_1, N_2 \rangle \in R^{\div}(\alpha)$, and $\alpha \notin \mathrm{EC}(N_1)$, then $\mathrm{EC}(N_2) = \mathrm{EC}(N_1)$
(Success) If $\langle N_1, N_2 \rangle \in R^{\div}(\alpha)$, and α is nor trivially true, then $\alpha \notin \mathrm{EC}(N_2)$
(Recovery) If $\langle N_1, N_2 \rangle \in R^{\div}(\alpha)$, and $\alpha \in \mathrm{EC}(N_1)$, then $\mathrm{EC}(N_1) \subseteq \mathrm{Cn}(\mathrm{EC}(N_2)) \cup \{\alpha\})$

One postulate is missing here: the *extensionality* postulate. First of all, since the framework we are working with here does not assume that contraction is functional, the extensionality principle should rather be expressed along the lines of something like the *substitutivity* postulate for revision in a relational framework (Lindström and Rabinowicz 1991). Then the postulate should look as follows:

$$\text{(Substitutivity) If } \alpha \Rightarrow \beta \text{ and } \beta \Rightarrow \alpha, \text{ then } \langle N_1, N_2 \rangle \in R^{\div}(\alpha) \text{ iff } \langle N_1, N_2 \rangle \in R \div (\beta)$$

But this postulate does not really make sense here. The reason for this is that equivalence between claims comes to identity in our framework. If $\alpha \Rightarrow \beta$ and $\beta \Rightarrow \alpha$, then $\alpha = \beta$, as is easily shown. We have already assumed a purely extensional representation of empirical claims, so any form of extensionality principle would be out of place.

We note, also, that the *Closure* postulate can be omitted, since it is trivially satisfied by the definition of the empirical content of a net. Lastly, we should add the following postulates for expansion and contraction (adapted from Lindström and Rabinowicz 1991):

(Seriality of expansion) For any net N_1, and any $\alpha \in \mathrm{CLM}(S)$, there is a net N_2 such that $\langle N_1, N_2 \rangle \in R^{+}(\alpha)$

(Seriality of contraction) For any net N_1, and any $\alpha \in \mathrm{CLM}(S)$, there is a net N_2 such that $\langle N_1, N_2 \rangle \in R^{\div}(\alpha)$

At this point, we have simply recreated the existing notions of expansion and contraction within a structuralistic framework. If this were the end of the story, then of course there would not be much interest in adopting this framework rather than any of the existing ones. We want to gain something new from the framework. The following two sections will be devoted to this – in Section 5.3.2, we show how the specialization structure of nets may shed some new light on contraction. In Section 5.3.3, we shall see that we can identify new *types* of theory change in the structuralistic framework developed here.

5.3.2 Contraction: Two Dimensions of Minimal Change

Let us take a closer look at contraction. What further conditions should we impose on contraction as a relation between theory nets? When answering this question, what we are looking for is more information about *what to keep and what not to keep* in the course of a contraction. The AGM postulate *Recovery* gives us some guidance in this matter: it is a first explication of the idea that we should attempt to *keep as much information as possible* when contracting. We should strive for a *minimal change*, in the common parlance.

But there is certainly more to the intuitive principle of "minimal change" than what is captured by the *Recovery* postulate alone. In the AGM theory, the principle of minimal change is further explicated in terms of the entrenchment ordering – contractions in AGM are minimal in the sense of *retaining the most entrenched beliefs*, thus making as "small" a change as possible of the belief set.

Can we reconstruct something similar to the entrenchment order in our framework? That is, can we construct some order over the set of asserted claims of some theory net, that guides us in the decision of what to retain and what to remove in the course of a contraction? As we noted in Section 5.2.2, one prima facie plausible motivation for keeping certain propositions of a theory is that they are particularly *fundamental* in the theory. With the notion of a structuralistic theory net, we can understand precisely what this means: the most fundamental principles of a theory are those that can be located higher up in the net, and this simply means that they have a wider *range* than the "lower" principles.

Let N be a net in some given structure space, and suppose that $\alpha = \langle X, P\rangle \in \mathrm{EC}(N)$. Say that we wish to remove α from the empirical content of the net N when performing some contraction on N. Then – just as in the AGM theory – we need to remove enough of the empirical content of N so that α is not entailed by any claims asserted in the new theory. In particular, for any claim $\beta = \langle Y, P\rangle \in \mathrm{EC}(N)$ such that $X \subseteq Y$, i.e. such that β asserts the same proposition as α about a wider class of applications, we have $\beta \Rightarrow \alpha$, and hence β must be removed as well. This means that, for a claim in the empirical content of a net, the propositional content of which is so to speak present high up in the net, removing this claim forces us to make changes within the more fundamental levels of the net. Let us try to capture this formally.

Let N be any theory net. Then we shall say that a *filter* in N is any subset S of N such that, if $E_1 \in S$, and E_1 s E_2 for some $E_2 \in S$, then $E_2 \in S$. Clearly, each non-empty filter in N contains the top element. Using filters, we shall now define the *depth of a in N* for a claim $\alpha \in \mathrm{EC}(N)$, or rather an *ordering* of the claims of a net w.r.t their relative depth compared to each other. If one claim is at least as large as another in this ordering, then we shall say that the former is at least as deeply grounded in the net as the latter. We begin with an auxiliary definition:

Definition 3.2.1 let N be any net, and let $\alpha \in \mathrm{EC}(N)$. Then an *a-free filter* in N is any filter S in N such that $\alpha \notin \mathrm{Cn}(\cup\{\mathrm{EC}\ (E)|\ E \in \mathrm{S}\})$

Clearly, since $\mathrm{Cn}(\emptyset) = \emptyset$, for each claim $\alpha \in \mathrm{EC}(N)$ there exists an α-free filter in N – it may, however, be empty.

Definition 3.2.2 Let N be any theory net, and suppose that $\alpha, \beta \in \mathrm{EC}(N)$. Then we say that β is *at least as deeply grounded in N as a*, and write $\alpha \leq_N \beta$, if every β-free filter in N is an α-free filter in N.

The intuitive content of this definition is, roughly, this: if $\alpha \leq_N \beta$, then in order to find a "subnet" of N that does not have β as part of its empirical content, we have to cut off at least so much of the net as to loose α as well.

Proposition 3.2.3 For any net N, $\leq_N$ is a preorder over the set $\mathrm{EC}(N)$.

Proof reflexivity is obvious. For transitivity, suppose that $\alpha \leq_N \beta$ and $\beta \leq_N \chi$ for $\alpha, \beta, \chi \in \mathrm{EC}(N)$. Let S be a χ-free filter in N. Since $\beta \leq_N \chi$, S is β-free. Since $\alpha \leq_N \beta$, S is α-free as well. This shows that $\alpha \leq_N \chi$, as desired. €

Proposition 3.2.4 For all $\alpha, \beta \in \mathrm{EC}(N)$, if $\alpha \Rightarrow \beta$ then $\alpha \leq_N \beta$

Proof if $\alpha \Rightarrow \beta$, then clearly each β-free filter in N is α-free. □

The relation $\leq_{\mathrm{N}}$ is not, in general, antisymmetric, i.e. it is not generally a partial order. Furthermore, it is not difficult to construct an example of a net N for which $\leq_N$ is not *connex*, i.e. we do not have $\alpha \leq_N \beta$ or $\beta \leq_N \alpha$ for all claims $\alpha, \beta \in \mathrm{EC}(N)$. Connexity of the ordering w.r.t depth corresponds to the specialization order being linear.

We wish to use this order over the set of claims asserted by some net to further constrain the relation of contraction between theory nets. In analogy with how the entrenchment order functions in the AGM framework, the order w.r.t groundedness in a net should make sure that the most deeply grounded claims asserted by a net are prioritized for being kept through the course of contraction.

Given any net N, we say that a *specialization-based fallback* of $\mathrm{EC}(N)$ (adapted from Lindström and Rabinowicz 1991) is a non-empty subset $\Gamma \subseteq \mathrm{EC}(N)$ which is logically closed, and such that if $\alpha \in \Gamma$, and $\alpha \leq_N \beta$, then $\beta \in \Gamma$. As a convention, we shall also always count the logical closure of the set of trivial claims in $\mathrm{EC}(N)$ as a fallback, although it may not be closed upwards under the ordering $\leq_N$. We say that a fallback Γ of $\mathrm{EC}(N)$ is *α-free*, for a claim α, if $\alpha \notin \Gamma$ (equivalently, $\alpha \notin \mathrm{Cn}(\Gamma)$). An α-free fallback is called maximal if it is not properly contained in any α-free fallback.

Proposition 3.2.5 for any net N and any claim α which is not trivially true, then there exists a maximal α-free specialization-based fallback of EC(N).

Proof we first show that there is at least one α-free fallback of $\mathrm{EC}(N)$, given that α is not trivially true. Let $\langle X, P\rangle$ be any claim in $\mathrm{EC}(N)$ (clearly, $\mathrm{EC}(N)$ is non-empty for any net N). Then $\langle X, U\rangle \in \mathrm{EC}(N)$, and this claim is trivially true. So the set of trivially true claims in EC(N) is non-empty, and by convention, the logical closure of the trivially true claims in EC(N) is a fallback of EC(N). It is easy to see that this fallback is α-free. In order to show that the set of α-free fallbacks of EC(N) contains maximal elements under set inclusion, we use Zorn's lemma. Take any chain C of α-free fallbacks in EC(N). We show that Cn($\cup C$) is an upper bound of C in the family of α-free fallbacks. It is logically closed, and it is easy to show that it is closed upwards under the relation $\leq_N$. It is also clearly contained in EC(N). Thus, it suffices to show that $\alpha \notin$ Cn($\cup C$). Suppose, on the contrary, that $\alpha \in$ Cn($\cup C$). Since Cn is finitary, there is a finite subset $\{\alpha_1, \ldots, \alpha_n\} \subseteq \cup C$ such that $\alpha \in \text{Cn}\{\alpha_1, \ldots, \alpha_n\}$. Using the fact that C is a chain, and that each member of C is logically closed, we find that $\alpha \in \Gamma$ for some $\Gamma \in C$, contrary to assumption. This proves the proposition. □

We can now use the specialization order of the net to constrain contractions, with the following postulate:

(*) If $\langle N_1, N_2 \rangle \in R^{\div}(\alpha)$, then if α is not trivially true, EC(N_2) is a maximal α-free specialization-based fallback of EC(N_1).

However, there seems to be a good reason not to accept this principle, in its current unrestricted form. As it stands, the postulate says that sufficiently deeply grounded claims of a net should always be kept, if possible, in the course of a contraction. But this principle puts a bit *too* much weight on the depth of a claim as a guidance for contracting. It seems reasonable that the depth of a claim should have some bearing on the decision of whether or the claim should remain in the empirical content of a net after contracting, but it is not the *only* factor which is relevant for such a decision. In particular, as we noted in Section 5.2.2, the notion of *corroboration* must also be taken into account.

Even if some part of the empirical content of a net is only part of some rather low level of the net – that is, even if a claim is not essentially tied to the most basic elements of the net – we may still have good reason to try to preserve it, since it may be very well *tested*. If an empirically testable claim has survived a great deal of tests, then that presumably makes it more likely to be *true*. Of course, if we believe a claim to be rather likely to be true, then we would be more reluctant to give it up.

So let us try to account for corroboration in our framework. We begin with the following definition:

Definition 3.2.6 Let S be any structure space. Then we shall say that a pair $\langle N, \sim \rangle$ is a *corroborated net* in S if

(i) N is a theory net in S,
(ii) $\sim$ is a preorder over EC(N)
(iii) for all $\alpha, \beta \in$ EC(N), if $\alpha \Rightarrow \beta$, then $\alpha \sim \beta$

If $\alpha \sim \beta$, then we say that β is *at least as well corroborated* as α.

That the corroboration relation $\sim$ should at least be a preorder seems rather intuitive – reflexivity is definitely reasonable, and transitivity also seems in order: if α is at least as well corroborated as β, and β is at least as well corroborated as χ, then α must be at least as well corroborated as χ. It should not be a partial order – of course, there is nothing that precludes two distinct empirical claims to be equally well corroborated. Clause (iii) also seems reasonable – if β is a logical consequence of α, then any piece of empirical evidence in favor of α is also evidence in favor of β.

We now make a small adjustment in our formal framework. Rather than taking R^{+} and $R^{\div}$ as functions from claims to relations over the set of theory nets in a space, we take them to be functions from claims to relations over the set of *corroborated* nets in a space. All the previous postulates can be adapted straightforwardly to fit with this modification, and when we refer back to one of the previously mentioned postulates, we take this to refer to the suitably adjusted version of the postulate.

Let $\langle N,\sim\rangle$ be any corroborated net. Then we say that a *corroboration-based fallback* of EC(N) (w.r.t $\sim$) is a logically closed non-empty subset G of EC(N) such that if $\alpha \in \Gamma$ and $\alpha \sim \beta$, then $\beta \in \Gamma$. As before, by convention, we count the logical closure of the set of trivially true claims in EC(N) as a corroboration based fallback. A corroboration-based fallback Γ is called *α-free* if $\alpha \notin \Gamma$. An α-free corroboration-based fallback of EC(N) is *maximal* if it is not properly contained in any α-free corroboration based fallback of EC(N). Using the same proof procedure as before, we have

Proposition 3.2.7 For any net N, and any claim α which is not trivially true, there is a maximal α-free corroboration-based fallback of EC(N).

Consider now the following postulate:

(#) Let $\langle N_1, \sim_1\rangle$, $\langle N_2, \sim_2\rangle$ be corroborated nets in any structure space, and suppose that $\langle\langle N_1, \sim_1\rangle, \langle N_2, \sim_2\rangle\rangle \in R^{\div}(\alpha)$. Then if α is not trivially true, EC(N_2) is a maximal α-free corroboration-based fallback of EC(N_1).

This is a straightforward translation of the postulate (*) for the corroboration order. Now, I claimed earlier that there is good reason not to assume the postulate (*) in its current form, and the motivation, it was hinted, is that we need to take also the corroboration order into account. What is not clear at this point is why we cannot do this while keeping the postulate (*) as it is. The most straightforward way to incorporate *both* the specialization order of a net and the notion of corroboration into contraction would be to simply assume both the postulates (*) and (#). So what prevents us from this?

We will give an example to illustrate this. Consider a theory net N, consisting of two distinct theory elements E_1, E_2 such that $E_2 \; s \; E_1$. E_2 is the least element in N, and E_1 is the top element. We shall make some assumptions about this net, and I leave it to the reader to check that none of these assumptions are inconsistent with any of the postulates or definitions we have given sofar. Suppose that α and β are two claims, which are not trivially true, and such that EC(E_1) = Cn($\{\alpha\}$), EC(E_2) = Cn($\{\alpha, \beta\}$). Suppose further that χ is a (non-trivial) claim such that $\{\alpha, \beta\} \Rightarrow \chi$, but neither $\alpha \Rightarrow \chi$ nor $\beta \Rightarrow \chi$. We then have $\chi \in$ EC(E_2), and hence $\chi \in$ EC(N).

Since the net N is linearly ordered by the specialization relation, the ordering $\leq_N$ is connex, and hence the set of specialization-based fallbacks of EC(N) are linearly ordered under set inclusion. We have not $\alpha \leq_N \chi$. Now let $\sim$ be a corroboration order over EC(N), which we assume to be connex, and such that *not* $\beta \sim \chi$. Since the corroboration order $\sim$ is connex, the set of corroboration-based fallbacks of EC(N) w.r.t $\sim$ is linearly ordered under set inclusion.

We show that we cannot successfully contract $\sim$ from the empirical content of the corroborated net $\langle N,\sim\rangle$, given that we assume both postulates (*) and (#). Since not $\alpha \leq_N \chi$, for any claim δ such that $\alpha \leq_N \delta$, we have not $\delta \leq_N \chi$ by transitivity of $\leq_N$. This shows that there exists a χ-free fallback of EC(N) which contains α. Hence, since the set of specialization-based fallbacks of EC(N) are linearly ordered under set inclusion, α is in every maximal χ-free specialization-based fallback of EC(N). For the same reason, β is in every maximal χ-free corroboration-based fallback of EC(N). But then, there can be no subset of EC(N) which is both a maximal χ-free specialization fallback and a maximal χ-free corroboration-based fallback, for by the above argument such a subset of EC(N) would have to contain both α and β, and therefore also χ, since $\{\alpha, \beta\} \Rightarrow \chi$. So we cannot find a subset of EC(N) which fulfils both postulates (*) and (#), and so we cannot successfully contract χ – which the postulates *Success* and *Seriality of contraction* tell us we should always be able to do.

Corroboration and specialization are two different respects in which we may regard a certain part of the empirical content of a theory as particularly desirable to keep in the course of contraction, and these may very well be in strict conflict with each other. In fact, this is not just a technical possibility – it seems rather reasonable to suspect that it would often be the case in the real world. For the more specialized principles of a theory have smaller ranges of intended applications, and generally one should expect it to be easier to formulate viable laws for smaller ranges of phenomena than more abstract laws with wide ranges. Thus, we can expect that it is quite often the case that the more specialized principles of a theory are formulated first, and only later subsumed under more general hypotheses. Of course, the longer we have held a certain claim to be true, the more likely it is that we have exposed it to a great deal of empirical testing, i.e. the more likely it is that the claim is relatively well corroborated.

So, say that we need to remove one of two claims from the empirical content of a net in order to contract succesfully, one being well corroborated and the other being deeply grounded in the net. Then a choice to remove the former would have the virtue of "protecting the core" in Lakatos' sense, but it would have the disadvantage of removing something which seemed likely to be true, something which was believed relatively firmly. Choosing to remove the latter would avoid this disadvantage, but instead it would force us to change the theory net in a more fundamental way. What we are faced with here, it seems, is a pragmatic factor in our model – the appropriateness of a choice between the two alternatives will most likely be dependent on contextual factors. For instance, if the core principles of a theory are regarded as very useful or beneficient in some sense, then we might expect a researcher working within the theory to be rather cautious about making deep changes in the theory. If a theory is rather young and in a relatively unstable state, then we might be more willing to make fundamental changes and preserve the specialized claims of the theory if they are well corroborated. Different strategies for changing a theory may have different virtues, and the choice between them may be a matter of choosing your priorities.

With this in mind, we should try to set up a principle of contraction which allows corroboration and specialization to restrict our range of admissible strategies for changing a net, but which leaves room for this kind of pragmatic decisions. One way of doing this would be to split contraction into two distinct types of theory change, say "specialization based" versus "corroboration based" contraction, treating these as distinct operations altogether. One would aim to preserve deeply grounded claims, and the other would aim to preserve corroborated ones. We can then keep both postulates (*) and (#), but let each of them correspond to its own species of contraction. The pragmatic factor is then present in the choice between these two types of contraction. But this solution does not seem quite right, since we would then always either take *only* specialization or *only* corroboration into account in any contraction. There seems to be no *prima facie* reason to exclude the possibility that we may take both factors into account in the same process of contraction, preserving corroborated claims in some instances and deeply grounded ones in other.

Instead, I would suggest that we keep treating contraction as a single type of theory change, and try to find some middle ground between the two postulates (*) and (#). What we need then, it seems, is some way of "weighing" the two orderings over the empirical content of a net against each other. For this purpose, we introduce the notion of a *merging* of the two relations:

Definition 3.2.8 Let $\langle N, \sim\rangle$ be a corroborated net. Then a *merging* of the orderings $\sim$ and $\leq_N$, is a relation R such that

$$\begin{aligned}
&\text{(i)}\, R \subseteq \sim \cup \leq_N \\
&\text{(ii)}\, \sim \cap \leq_N \subseteq R \\
&\text{(iii)}\, R \text{ is a preorder over } \mathrm{EC}(N) \\
&\text{(iv)}\, \text{for all } \alpha, \beta \in \mathrm{EC}(N), \text{ if } \alpha \Rightarrow \beta \text{ then } \alpha R \beta
\end{aligned}$$

The first clause in this definition says that a merging of $\sim$ and $\leq_N$ is "built up" strictly from the relations $\sim$ and $\leq_N$. The second clause says that any merging should contain all those pairs on which the corroboration relation and the ordering w.r.t groundedness are in agreement. The last two clauses say that the merging should have the formal properties of a corroboration relation or an ordering w.r.t depth.

Proposition 3.2.9 For any corroborated net $\langle N, \sim\rangle$, both $\sim$ and $\leq_N$ are mergings of $\sim$ and $\leq_N$.

This proposition shows that there are always two trivial possibilities for merging the relations $\sim$ and $\leq_N$: simply ignoring one of them and taking only the other into account gives a merging. However, it would seem reasonable to try to incorporate as much as possible from the two orderings when merging; the more information we retain concerning what to keep and what not to keep in a contraction, the better. This suggests that we should look for *maximal* mergings when contracting. Maximality here means maximality w.r.t set inclusion; a merging is said to be maximal if it is not properly contained in any merging. We have the following result:

Proposition 3.2.10 For any corroborated net $\langle N, \sim \rangle$, there exists some maximal merging of $\sim$ and $\leq_N$.

Proof Proposition 3.2.9 shows that the set of mergings of $\sim$ and $\leq_N$ is non-empty. To show that there exist maximal mergings, we ue Zorn's lemma. Let C be a chain in the set of mergings of $\sim$ and $\leq_N$. We show that $\cup C$ is an upper bound of C in the set of mergings of $\sim$ and $\leq_N$. It suffices to show that $\cup C$ is indeed a merging of $\sim$ and $\leq_N$. Clauses (i) and (ii) are obvious. For clause (iii), suppose that for some $\alpha, \beta, \chi \in \mathrm{EC}(N)$, we have $\langle \alpha, \beta \rangle \in \cup C$ and $\langle \beta, \chi \rangle \in \cup C$. Then there are $R_1, R_2 \in C$ such that $\langle \alpha, \beta \rangle \in R_1$, $\langle \beta, \chi \rangle \in R_2$. Since C is a chain, we have $R_1 \subseteq R_2$ or $R_2 \subseteq R_1$. In either case, we have $\langle \alpha, \beta \rangle$, $\langle \beta, \chi \rangle \in R_i$, for some $i \in \{1, 2\}$. Hence, since R_i is a merging, $\langle \alpha, \chi \rangle \in R_i$, and so $\langle \alpha, \chi \rangle \in \cup C$. So $\cup C$ is transitive. It is easy to show that $\cup C$ is reflexive over EC(N), since $\langle \alpha, \alpha \rangle \in \sim \cap \leq_N$ for each $\alpha \in \mathrm{EC}(N)$, clearly. Similarly, we obtain (iv). This proves the proposition. □

Let $\langle N, \sim \rangle$ be any corroborated theory net, and let R be a maximal merging of $\sim$ and $\leq_N$. Then, as before, we say that an *R-based fallback* of EC(N) is a logically closed subset $\Gamma \subseteq \mathrm{EC}(N)$ such that if $\alpha \in \Gamma$ and $\alpha R \beta$, then $\beta \in \Gamma$. The logical closure of the set of trivially true claims in EC(N) is an R-based fallback by convention. We say that an R-based fallback Γ is *α-free*, for any claim α, if $\alpha \notin \Gamma$. An α-free R-based fallback is called *maximal* if it is not properly contained in any α-free fallback. As before, we have

Proposition 3.2.11 For any corroborated net $\langle N, \sim \rangle$, and any maximal merging R of $\sim$ and $\leq_N$, for any claim α which is not trivially true there exists a maximal α-free R-based fallback of EC(N).

We are now ready to set the desired middle road between the postulates (*) and (#):

(†) Let $\langle N_1, \sim_1 \rangle$, $\langle N_2, \sim_2 \rangle$ be corroborated nets in any structure space, and suppose that $\langle \langle N_1, \sim_1 \rangle, \langle N_2, \sim_2 \rangle \rangle \in R^+(\alpha)$. Then if α is not trivially true, EC(N_2) is a maximal α-free R-based fallback of EC(N_1), for some maximal merging R of $\sim$ and $\leq_N$.

Let us conclude this section with a reflection upon one of the philosophical consequences of our framework, in particular the notion of contraction we have developed here. What the postulate (†) tells the agent to do, when contracting a claim α from the empirical content of a net, is to look both to the specialization structure of the net and the relative corroboration of the claims of the net for guidance concerning what to remove and what to keep, and then look for a best possible way of reconciling these two pieces of information (i.e. a maximal merging) which is then used as a basis for the contraction. Now, this principle makes sure that the agent is *constrained* in his decision to contract in this or that way, both by the specialization structure of the net and the relative corroboration of the claims of the net. But he is also left with a certain amount of freedom: he must base his decision on some maximal merging of the two orderings, but clearly, there are generally several "best possible" ways of merging the orderings. The agent may put more weight on corroboration, or he may put more weight on specialization. Several strategies are open to the agent who wishes to contract some claim from the empirical content of a net, but the range of admissible strategies is limited by certain constraints.

In the classical AGM theory, there is simply one way to contract, and it is determined by the entrenchment order. If an agent wishes to remove some piece of information from his corpus, then if he is an AGM type agent, he has only one strategy to follow – he is to retain the most entrenched beliefs. In our framework, however, contracting one and the same claim can be done in a variety of ways. First, of course, even if we had followed either the postulate (*) or (#), contraction would not in general be functional, since theory nets do not have to be linearly ordered under the specialization order, and we have not assumed the corroboration order to be connex. But adopting the postulate (†), it seems we have a deeper or at least rather different reason why contractions are not in general functional: the agent is free to determine the degree to which he assigns highest priority to corroboration or specialization as a means to determine which claims are to remain in the empirical content of a net after a contraction. Thus, even with a linearly ordered net, and a connected corroboration order, contraction may not be functional. Thus, the primary motivation for having a non-deterministic notion of contraction is here rather different from, for instance, the argument given by Lindström and Rabinowicz (1991), based on incomparabilities in the entrenchment order. The problem is not that certain alternatives for performing a contraction are *incomparable*, it is that they may be comparable on several *different, conflicting grounds*.

5.3.3 Further Types of Theory Change?

We have, at this point given an account of expansion and contraction of a net with an empirical claim. Thus we have in effect given an account also of revision, under the proviso that revision can be defined in terms of the Levi identity in our setting; and there seems to be nothing to prevent this. So the basic AGM types of theory change can be modelled within our structuralist framework.

An interesting question now presents itself: with the additional structure we obtain when representing theories in terms of structuralist theory nets, can we identify further types of theory change, beyond the three basic AGM-types of expansion, contraction and revision? That is, can we obtain a *finer typology* of theory changes in our setting? This question actually splits into two questions: first, we may ask whether we can make finer distinctions *within* the three basic types of theory change, which cannot be made in the AGM framework. Second, we may ask whether it is possible to identify new types of theory change that fall *outside* the scope of expansion, contraction and revision. That is, is it possible to identify (natural) types of theory change that cannot in general be represented as series of expansions, contractions and revisions? Since we obtain revision by the Levi identity, this question is equivalent to the question whether we can find types of theory change that cannot be represented as series of expansions and contractions.

Turning to the first question, we need to make clear what it would mean to identify finer distinctions within the categories of expansion, contraction and revision. In order for us to plausibly affirm a posistive answer to this question, I think we should at least have the following condition: we should be able to point to a subtype of expansion or contraction (leaving revision aside) for which we need to postulate some *additional postulates* beyond the AGM postulates, and these additional postulates should not be possible to formulate in the AGM framework. This last restriction is of course a bit problematic; it is not entirely clear what it means for it not to be possible to formulate a certain principle in the AGM framework. Therefore, it does not seem possible to present any straightforward proof that a certain type of theory change is not representable in the AGM framework. The best we can do is, probably, to identify a type of theory change which is subject to constraints which seemingly depend on the additional features of our framework in an essential way – and then leave the burden of proof to one who wishes to maintain that these types of theory change can be reconstructed with the conceptual tools of AGM in a plausible manner.

With this in mind, I think that we can in fact answer the first question, of whether we can identify subtypes of the basic AGM postulates that are not representable in the conceptual framework of AGM, in the affirmative. We have already laid the foundation for such a distinction: in our treatment of contraction, we saw that there are two distinct orderings which the agent may take into account when making a decision to keep or not keep some part of the empirical content of a net through the course of a contraction. The agent is required to weigh these orderings against each other, and construct a merged ordering which is then taken as the basis for the contraction. When doing so, the agent is free to put more weight on specialization, or on corroboration, according to his preferences.

However, we could of course restrict the range of admissible mergings somehow, so that the agent is not free to choose *any* maximal merging of the two orderings as a basis for contraction. Such a restriction may not be suitable for contraction in *general*, but it could still determine some subtype of contraction which we may want to consider. For instance, we could restrict the agent in the following way: when merging the two orderings, whenever there is a conflict between the corroboration order and the ordering w.r.t depth, the latter is to take priority. This principle can be seen as a strong form of Lakatos' principle of protecting the core: when we contract something from the empirical content of a net, we must protect the most fundamental principles of the theory, *even at the expense* of loosing well corroborated claims. This principle does not seem reasonable to accept in general – it would mean that we are always to protect the core of a theory, "at any price", and this I think we should not assume. But still, this principle determines a subtype of contraction, and the additional constraint of prioritizing the depth of a claim over its corroboration in the decision to keep or remove it might be suitable in certain circumstances. In fact, something like this principle seems to be present in what Kuhn would call normal science: a period in the evolution of a theory where the core principles are unquestioned, and protected even at the expense of claims for which there are good evidence.

Let us try to distinguish this type of contraction formally. Let us call it *core-protecting contraction*, and denote it by the symbol $R^{©}$. Thus, if a net $\langle N_2, \sim_2\rangle$ results from a core-protecting contraction of the empirical content of a net $\langle N_1, \sim_1\rangle$ by the claim α, then we write $\langle\langle N_1, \sim_1\rangle, \langle N_2, \sim_2\rangle\rangle \in R^{©}(\alpha)$. The postulates for this type of contraction are then the postulates for contraction that we have given up to this point, plus the following modified version of the postulate (†):

(©) Let $\langle N_1, \sim_1\rangle$, $\langle N_2, \sim_2\rangle$ be corroborated nets in any structure space, and suppose that $\langle\langle N_1, \sim_1\rangle, \langle N_2, \sim_2\rangle\rangle, \in R^{©}(\alpha)$. Then if α is not trivially true, EC(N_2) is a maximal α-free R-based fallback of EC(N_1), for some maximal merging R of $\sim$ and $\leq_{N_1}$. Furthermore,

$$\leq_{N_1} \subseteq R$$

The last added clause in the postulate is intended to capture the idea that the ordering of claims w.r.t depth is prioritized over corroboration. The range of admissible mergings for the agent is restricted to those that contain the entire ordering w.r.t depth; in case of a conflict between the two orderings, the ordering w.r.t depth must thus be put in first place. In the same way, of course, we could construct a subtype of contraction that prioritizes corroboration.

Have we here identified a type of contraction which is not representable in the AGM-framework? One of our desiderata is fulfilled, at least: the postulate (©) is restricted to this subtype of contraction, and does not seem plausible to assume as a principle for contraction in general. The question remains whether or not this type of theory change is in some sense distinguishable using the conceptual apparatus of AGM. But it seems reasonable to answer this question negatively; for the distinction between the two orderings over the claims of a net is only possible once we acknowledge that the principles of a theory have different ranges of intended applications, and this is an intrinsic part of the structuralist framework which is simply absent in the AGM-style representation of theories as logically closed sets of sentences.

Now, one could argue that we could make the distinction by simply replacing the entrenchment order in AGM by two distinct orderings, one with respect to corroboration, and one with respect to depth. Then we would make a change of the AGM-framework which is significantly smaller than the adoption of a structuralist representation of theories, but which would still allow us to distinguish core-protecting contraction from contraction in general.

But this, I think, would not really amount to the same thing. For the making the distinction between the corroboration order and the ordering w.r.t depth over the claims of a net is something deeper than just assuming two distinct orderings; it is important to realize that we do not *identify* a theory with its empirical content, its set of claims. We identify theories with theory nets, and the empirical content of a net is distinct from the net itself, which contains the specialization order as an essential component. The point is that the ordering w.r.t depth, brought about by the specialization structure of the net, is not something external to the theory: it is an *intrinsic part of the logical structure of the theory*. Thus, the distinction between the ordering w.r.t depth and the corroboration order is a distinction between something which is a *part* of the theory, and something which is external to it. In order to make *this* distinction, we would have to make a more drastic revision of the AGM framework than just adding another ordering to the epistemic states.

We now turn to the second question, of whether we can find types of theory change that lie outside the scope of the basic AGM-types of theory change. I believe this question can also be answered in the affirmative, and once again the key to this result is the presence of intended applications as a part of the logical structure of theories. The basic observation that we shall use to confirm this fact was noted in Section 5.2.2: as we take intended applications into account, there are two distinct ways of strengthening the empirical content of a theory element. In order to strengthen what a certain theory says about the world, we may either add further axioms to it, or we may widen its class of intended applications. In a sense, we may add to what the theory says about some chunk of the world, or we may let the theory say the same thing about a *larger* chunk of the world. Likewise, we may weaken the empirical content of a theory either by weakening some of its axioms, or by dropping some previously accepted applications of the theory.

Let us focus on the latter possibility, of removing some part of the class of intended applications of a theory element as a means to weaken its empirical content. We want to represent this formally as a type of theory change, and then we shall ask the question of whether this type of theory change is definable from the AGM-style theory changes we have provided sofar. We will not attempt to give a full account of this type of theory change; we shall only provide some basic postulates for it.

Let N be any theory net, and let $E \in N$. Suppose then that X is some subset of the set $I(E)$ of intended applications for E. What we are after is a way of removing X from the class of intended applications for E. This kind of operation on a theory is quite different from what we have encountered before: expansion and contraction were both defined as operations on the empirical content of a net, consisting in the addition or subtraction of some claim. Here, we are rather operating directly on the theory net itself, and the "input" of the theory change is a class of intended applications, not an empirical claim. We shall denote this type of theory change by the symbol R^{app}, and we shall model it as a function from pairs of theory elements and closed sets to relations over the set of theory nets in a given structure space. If N_1, N_2 are theory nets, E is a theory element and X is some closed set, then we shall write $\langle N_1, N_2 \rangle \in R^{app}(E, X)$ to say that N_2 results from N_1 by removing the all members of the set X from the class of intended applications of E. Now, of course, if E is not a member of the net N_1, then removing the members of X from the intended applications of E should have no effect on the net N_1. Also, if the intended applications of E contain no member of X, then there should be no effect on the net N_1 of removing the members of X from $I(E)$, since there is nothing to remove. We thus introduce the following postulate, to deal with these trivial cases:

(R^{app} 1) Let N_1 be any theory net in some structure space, and let E be any theory element. Let X be any closed set in the space. Then if $E \notin N_1$, or if $X \cap I(E) = \emptyset$, then

$$\langle N_1, N_2 \rangle \in R^{\text{app}}(E, X) \text{ iff } N_2 = N_1.$$

There is another case in which we cannot remove the class X from the intended applications of E successfully: it may be that $I(E) \subseteq X$, and then we cannot successfully remove the applications in X from $I(E)$, since the set of intended applications for the resulting element would be empty, which we do not allow. More generally, in order to be able to remove the elements of X succesfully, $I(E)$ must contain a non-empty closed set Y such that $Y \cap X = \emptyset$. If this is the case, however, the action of removing the elements of X should be succesful. We sum this up in the following postulate:

(R^{app}2) Let N_1,N_2 be theory nets in a structure space such that $\langle N_1, N_2 \rangle \in R^{\text{app}}(E, X)$, for $E \in N_1$, and where X is some closed set such that $X \cap I(E) \neq \emptyset$. Then

(i) N_2 contains an element E^* such that $M(E^*) = M(E)$ and $I(E^*)$ is a closed non - empty subset of $I(E)$ such that $I(E^*) \cap X = \emptyset$, if such a set exists,
(ii) $N_2 = N_1$ otherwise.

Like with the other types of theory change, we introduce a seriality postulate:

(R^{app}3) For any net N_1, theory element E and closed set, X, there is some net N_2 such that $\langle N_1, N_2 \rangle \in R^{\text{app}}(E, X)$.

These postulates are far from being a complete characterization of this type of theory change – in particular, we have said nothing of how to change the rest of the net while we remove a class of applications from one of its elements. But let us rest content with them for the moment, and ask whether this type of theory change is representable in the AGM framework.

There is an immediate reason to answer this question negatively: intended applications just aren't part of the AGM framework, and so the notion of removing intended applications from a theory would not seem to make any sense. But I think we can go even further – we have reconstructed the AGM operations of expansion of contraction in our framework, and so we may ask whether we can define some relation over theory nets which has the above properties from the already given relations of expansion and contraction; and so, in a certain sense, recreate this type of theory change without going outside the typology of theory changes of the AGM framework. It is not entirely clear what this would mean, but a reasonable way of explicating the notion is this: in order to answer the question affirmatively, for any net N_1, theory element E and closed set X we should be able to find a *finite* series of contractions and expansions which result in a net N_2 such that N_1 and N_2 are related in the way described by the above postulates. Can we always do this?

I cannot at this point present any definitive proof either way, but I conjecture that the answer is *no*. The reason is the following: say that N_1 is a net which contains an element E, such that for some closed set X, there is a closed non-empty subset Y of $I(E)$ such that $Y \cap X = \emptyset$. According to ($R^{\text{app}}3$), we should be able to find some net N_2 such that $\langle N_1, N_2 \rangle, \in R^{\text{app}}(E, X)$, and according to ($R^{\text{app}}2$), N_2 should contain some element E^* such that $M(E^*) = M(E)$ and $I(E^*) \subseteq I(E) - X$. How would we go about to find such a net through a series of contractions and expansions? The most likely strategy would be something like this: first remove all claims of the form $\langle Y,\ P \rangle$, where $M(E) \subseteq P$ and Y is any closed subset if $I(E)$ which intersects X, through a series of contractions, and the expand again in order to regain all claims of the form $\langle \mathrm{Z}, P \rangle$ where $M(E) \subseteq P$, and Z is a fixed non-empty closed subset of $I(E)$ which does not intersect X.

But $M(E)$ might not be finitely axiomatized, that is, there might not be a finite family of clopen sets the intersection of which equals $M(E)$. In this case, in order to regain an element E^* which has exactly the same models as E (but a different class of intended applications), we might have to resort to an infinite series of expansions. The point is that we are here not working with single claims, we are defining an operation on entire theory elements, and such operations may not be possible to break up into a finite series of operations using single claims. Therefore, I believe we have good reason to think that the conjecture above is true. Hence, we answer also the second question of this section, whether there are types of theory change representable in this framework that cannot plausibly be recreated using the tools of AGM (or any closely related system, for that matter), affirmatively. For one who wishes to dispute this conclusion, the challenge is to prove that my conjecture above is false.

5.4 Conclusion

The goal in this essay was to develop a structuralistic framework for the logic of theory change. We used only a rather small part of the structuralist model of empirical theories as a basis for the framework; the main difference with the AGM representation of theories and the notion of theory used here is that the principles of a theory are always taken relative to a range of *intended applications*. What have we gained by adding this element to the representation of theories?

First of all, we showed in Section 5.3.2 how the specialization structure of theory nets gives rise to an ordering over the claims of a net, which can then play essentially the same role as the entrenchment order in AGM, as a guide to what claims are to be kept or removed in the course of a contraction. But there is an important difference between this order and the entrenchment order of AGM: the entrenchment order is simply postulated, as something external to the theory held by an agent at a certain time. In AGM, there is no distinction between the beliefs of an agent and the theory the agent holds at a particular time: theories just are sets of beliefs. In our framework, however, we have not equated theories with sets of beliefs – theories give rise to what we have called *empirical contents*, which can be seen as what an agent holding the theory believes about the world. But the theory itself has a deeper structure, and this structure itself gives rise to an ordering over the claims of the theory. The agent holding a theory can gain information about which beliefs to keep or remove through a contraction simply by examining the logical structure of his theory, the specialization structure of the net. And there is nothing *arbitrary* about this structure – it determines the ranges of different principles of the theory, and thus what the theory says about the world. What we have obtained is a kind of analysis of the entrenchment order – or perhaps better, an *explication* of it in terms of the logical structure of empirical theories.

But we also noted that this ordering of the claims of a theory, which is *internal* to the theory, should also be supplemented by the *external* relation of *corroboration*. This gives rise to an interesting new possibility: we have two distinct orderings over the claims of a theory, with essentially different informal contents, and these two orderings may in fact be in conflict with each other. In this case, the agent has to find a way of weighing the two senses in which claims may be desirable to keep against each other (our notion of *merging*). This gives rise to a pragmatic aspect of contraction: the choice to merge the two orderings in this or that way is up to the preferences of the agent. The agent has a certain amount of freedom in how he chooses to contract, and this depends on how he sets his priorities between the two orderings of his beliefs. This sheds some new light on the discussion concerning relational vs. functional theories of belief revision; the primary reason why our framework gives rise to a relational theory of contraction is not that there may be incomparabilities in one order over the beliefs, it is that we may have several conflicting reasons to regard certain beliefs as less admissible to give up.

Lastly, we argued in Section 5.3.3 that the additional structure lets us distinguish new types of theory change. For instance, we saw that we could distinguish subtypes of contraction by placing restrictions on the admissible mergings of the corroboration order and the ordering w.r.t depth. But more radically, we saw that we could define types of theory change that lie outside the scope of the basic AGM operations of contraction and expansion. Our example was the possibility of changing the empirical content of a theory element simply by changing its set of intended applications, keeping the core intact (an observation due to Gärdenfors).

So, although we used only a small part of structuralism, i.e. the addition of intended applications to the structure of theories, we were able to reach some substantial consequences for the logic of theory change. My hope is that, once we extend the framework to encompass more of the structuralist representation of theories, the ramifications for the logic of theory change will be all the greater. In particular, when we take the possible models of a theory into account, along with the distinction between theoretical and observable concepts of the theory, and the theoretical constraints, we have an account of the theoretical *framework* of a theory.

This would, first, give rise to further types of theory change: changes in the conceptual framework of a theory. It should be interesting to investigate the properties of such theory changes, and how they differ from the types of theory change we have considered here. This goes beyond the scope of the current essay – but hopefully, I have at least succeded in providing a basis for further investigations, in which the full apparatus of structuralism can be brought into the normative study of theory change.

References

Balzer, W., and C.U. Moulines. (eds.). 1996. *Structuralist theory of science (focal issues, new results)*. Berlin, NY: Walter de Gruyter.

Balzer, W., C.U., Moulines, and J.D. Sneed. 1987. *An architectonic for science*. Dordrecht: Reidel.

Gärdenfors, P. 1988. *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge, MA: MIT.

Gärdenfors, P. 1992. Sneeds rekonstruktion av teoriers struktur och dynamik. In *Metod eller anarki*, ed. B. Hansson, 93–105. University of Lund.

Grove, A. 1988. Two modellings for theory change. *Journal of Philosophical Logic* 2(17):157–170.

Levi, I. 1991. *Fixation of belief and its undoing: Changing beliefs through inquiry*. Cambridge, MA: Cambridge University Press.

Lindström, S., and W. Rabinowicz. 1991. Epistemic entrenchment with incomparabilities and relational belief revision. In *The logic of theory change*, eds. A. Fuhrmann and M. Morreau, 93–126. Berlin: Springer.

Olsson, E.J. 2006. Lindström and Rabinowicz on relational belief revision. Festschrift to Wlodek Rabinowicz on the occasion of his 60th birthday.

Olsson, E., and D. Westlund. 2006. On the role of the research agenda in epistemic change. Retrieved December 3, 2007, from Lund University, Department of Philosophy web site: http://www.fil.lu.se/publicationfiles/pp91.pdf

Segerberg, K. 1998. Irrevocable belief revision in dynamic doxastic logic. *Notre Dame Journal of Formal Logic* 39(3):287–306.

Segerberg, K., and H. Leitgeb. 2007. Dynamic doxastic logic: Why, how and where to? *Synthese* 155*:*167–190.

Sneed, J.D. 1971. *The logical structure of mathematical physics*. Dordrecht: Reidel.

Chapter 6
Using Conceptual Spaces to Model the Dynamics of Empirical Theories

Peter Gärdenfors and Frank Zenker

6.1 Introduction

The aim of this paper is to apply *conceptual spaces* as developed by Gärdenfors (2000) to give a new account of the dynamics of scientific theories. We compare this account to the *structuralist view* of empirical theories (Sneed 1971, Stegmüller 1976, Balzer et al. 1984, 1987, 2000, Moulines 2002). Using a reconstruction of *Newtonian Particle Mechanics* (NPM) as a paradigmatic example, we explain how the structuralists' terms are applied. Our claim is: By using conceptual spaces, one can recover *most* of the key concepts of the structuralist view without the set-theoretical overhead. There is also some loss in comparison with structuralism, because we are not after a *mathematically general* model. We do not situate our approach with respect to the statement vs. non-statement dichotomy, as this appears (to us) as an overstressed distinction.

Our aim is to provide philosophy of science with a new tool by which to comprehend in general terms the dynamics of empirical theories. We argue that our approach, which is based on geometrical notions, is more suited as a general framework for representing theories and their dynamics than structuralism is. We believe it also fits better with the intuitions of practicing scientists. By means of examples, it is shown that conceptual spaces provide a clearer account of different kinds of changes of scientific theories.

We start with a summary of the structuralist view of empirical theories (Section 6.2), followed by an outline of our modeling tool, *conceptual spaces* (Section 6.3). In Section 6.4, we show how the central notions of structuralism

P. Gärdenfors (✉)
Department of Philosophy, University of Lund, Kungshuset, Lundagard, 22 222 Lund, Sweden
e-mail: peter.gardenfors@lucs.lu.se

Forthcoming in: Olsson, E.; S. Rahman, T., and Tulenheimo. *Science in flux*: *Philosophy of science meets belief revision theory* (Logic, Epistemology, and the Unity of Science Series). Berlin: Springer.

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_6,

can be expressed in terms of conceptual spaces. We then summarize the structuralist reconstruction of theory changes and point out its limitations in the application to *radical* or *revolutionary* changes (Section 6.5). In Section 6.6, we let conceptual spaces prove their mettle by presenting four types of increasingly more severe changes to empirical theories.

6.2 A Brief Summary of the Structuralist View

For a structuralist, an empirical theory is a set of *mathematical (set-theoretical) structures* – hence the name. These structures happen to satisfy some *axioms* that can be expressed as set-theoretical predicates (see Suppes 1957: ch. XII, Sneed 1971). Since the structures and not the axioms are central, structuralists characterize their program as a *non-statement* or *semantic view* of empirical theories. Ultimately, the aim of the program is to provide a framework for the detailed representation of the logical structure of an empirical theory in order to achieve a rigorous reconstruction of *changes* either to the theory or the conditions of its application to empirical phenomena, for example, a reconstruction of how NPM changed over time.[1]

We now turn to a brief presentation of the basic concepts of structuralism. According to Sneed's (1971) account (further developed by Stegmüller (1976)), an empirical theory is represented as a pair $<K, \ I>$, consisting of a *formal core*, *K*, and a set of *intended applications*, *I*. The intended applications are identified pragmatically, while *K* is specified via the set-theoretic structures which systematize its parts, most notably its *models* (see below). Although structuralism has developed over time,[2] we choose to focus on Sneed's original account since it contains the most essential components.

The mathematical structure of a theory core is described, firstly, by a set of *measures* (variables) for different magnitudes of the objects that are studied. The measures are functions in the set-theoretic sense of sets of ordered pairs. For example, in Newtonian Particle Mechanics (NPM) the relevant variables of an object are *position* (location in space), *time*, *mass* and *force*. Secondly, there are *constraints* for these measures. Thus, in every model of NPM, mass is supposed to be a magnitude which is *conservative* (any object has the same mass in all applications) and *additive* (the mass of a complex object is the sum of the masses of its components).[3]

[1] The construction and application of an empirical theory can naturally be seen as the paradigmatic example of rational human belief and its use. In this sense, structuralism also becomes a framework, albeit limited, for doxastic dynamics.

[2] See Moulines (2002) for a brief outline of the current state of the structuralist program and Balzer et al. (1987) for a full account.

[3] Balzer et al. (1987: 105) call these the constraints of *equality* and *extensivity*. Compare their ensuing discussion with respect to the simplifying assumption that gives rise to a third constraint: In any subsystem considered, e.g., moon and earth, masses are assumed to be impressed upon by the *same* forces, if these masses were related to the system as a whole, i.e., the cosmos. "Although

Magnitudes which cannot be measured without the theory being applied are called *T-theoretical.* Those which do not presuppose the theory for their measurement are called *T-non-theoretical* (Sneed 1971). Thus, in NPM, force and mass are theoretical, while position and time are non-theoretical, because we are not required to use NPM to determine values of the former. Instead we rely on an "antecedently accepted" theory for measuring space and time.[4]

The distinction between theoretical and non-theoretical magnitudes motivates a corresponding one among a theory's models: An empirical structure, also called *data-structure* of values for the measured variables (space and time) is called a *partial potential model* and the set of these structures is denoted *Mpp – partial*, because the model *lacks* theoretical functions. If values for non-theoretical terms (*time* and *location* in NPM) are specified, but the values for the theoretical terms are unconstrained, one obtains the set of *potential models Mp – potential*, because all possible values for theoretical functions appear in some member of *Mp*. Thus, in *Mp*, kinematical descriptions (co-ordinations of indexes for time over Euclidean space) are set in relation to a system's masses and the forces impressed upon them, that is, the dynamical factors. Of the potential models, only some will satisfy the central axioms of the theory, for example Newton's second law $F = ma$ in NPM. A potential model that satisfies the axioms of the theory is called a (full) *model* of the theory. The set of models is denoted *M*. Hence: $M \subseteq Mp$, while the relation between *Mp* and *Mpp* is formally expressed by a "forgetful functor" (alternatively: a restriction function (projection) from *Mp* onto *Mpp*).

In order to address changes to empirical theories, it is useful to focus on Sneed's (1971) notion of the *core of a theory*. It consists of the following five parts: (i) a set of variables of the theory, (ii) a set of constraints for the theoretical variables, (iii) a set of models determined by the central axioms for the theory, (iv) a set of potential models and (v) a set of partial potential models.[5] The core captures what remains *constant* in a theory during the time of its use.

Some changes of a core are so-called *core expansions* (Stegmüller 1976, p. 107). These are reached from the core through a process of *specialization*, thus adding special laws of a theory, e.g., the Newtonian law of gravitation (see below). The core of NPM, for example, does not yet establish any *quantitative* relation between given masses, forces and accelerations. Rather, a core specifies so-called *basic laws* of a theory.[6]

this is not so if we look at things quite accurately, physical calculations work with such and similar assumptions" (1987: 106).

[4] NPM does, of course, presuppose its own theory of space-time, namely that of *absolute* (Euclidian) *space* and *absolute simultaneity*, see DiSalle (2006: 17–35, 98–130).

[5] This is not exactly Sneed's definition, which may be found in his (1971: 171). Our above characterization of a core, however, captures what is essential about this concept.

[6] This is Sneed's term, cf. Sneed (1979: 300); Stegmüller (1976: 107f.) uses *fundamental laws*. In addition, there are assumed characterizations of the single components of the model, so-called *frame conditions*. E.g., for NPM, that the set of particles is finite or that mass is a positive real

In this sense, basic laws underlie – thus come to be valid in, thus unify – all of a theory's models. Of course, the core does not yet exhaust a theory. Rather, it spells out presuppositions with which *more specific* characterizations – formulated as core-expansions and by means of *special* or *substantial laws* – must be consistent. Only the latter are (sought to be) applied to real world situations.[7]

Through a *specialization* of the core, that is, by adding restricting characterizations, the structuralist can specify a theory's *internal structure* hierarchically as a partial order over so-called *theory-elements*.[8] For example, in the case of NPM, from the basic law $F = ma$ (expressed in theory element T_0) one may "carve out" (Balzer et al. 1987: 169) first, T_1, the *actio-reactio* principle (Newton's third law), followed by specifying *conservative forces* in T_2, then *central forces*, T_3, followed by *forces inversely proportional to distance*, T_4, and, finally, *forces for which the gravitational constant*, G, *holds*, T_5.[9]

Thus, one reaches the law of gravitation, $F = \mathrm{G}{\cdot}Mm/r^2$, as a five-fold specialization from the core of NPM. This way, one has construed a series of core expansions and thereby specified the core. Note, again, with only $F = ma$ in the core, we do not have any information about the interaction of particles. Models that satisfy $F = ma$ might just as well be about single particles. It is only with special laws such as the law of gravitation that connections between objects are introduced.

The decision on what to reconstruct as the core of a theory is relevant, because the potential models are characterized *only* by the core structure. Therefore, every more specialized theory T_{n+1} must have the *same* set of potential models as the less specialized T_n, i.e., $Mp(T_{n+1}) = Mp(T_n)$.[10] Differences arise, among other things, with respect to a theory-element's full models, *M*, and its intended applications *I*, that is, $M(\mathrm{T}_{n+1}) \subseteq M(\mathrm{T}_n)$ and $I(\mathrm{T}_{n+1}) \subseteq I(\mathrm{T}_n)$.

Since there is, in principle, more than one way in which a partial potential model can be completed to a full model, the *empirical claim* of a theory is rendered as the contention that every partial potential model (representing data structures arrived at by measurement) *can* be successfully enriched to a full model. This claim *may* very well turn out to be false, in which case one would say: A particular theoretical enrichment of a partial potential model constitutes an *anomaly* for the theory.

function. These conditions "do not say anything about the world (or are not expected to do so) but just settle the formal properties of the scientific concepts we want to use" (Moulines 2002: p. 5).

[7] Hence, both basic laws and frame conditions are not *open* to refutation in the same sense that special laws are. Certainly, they can be *revised*, but not (without using stronger assumptions) *falsified*.

[8] The term "theory-element" denotes the set of the sets *M*, *Mp*, *Mpp*, *L* (links to theory elements specialized from different cores), *C* (constraints), *I* (intended applications) and blurs, *B*, used for approximation (see Balzer et al. 1987).

[9] Gähde (1997, 2002).

[10] *A fortiori* for the so-called *partial potential models, Mpp*, i.e., $Mpp(T_{n+1}) = Mpp(T_n)$.

Among the partial potential models are, most notably, the *paradigmatic applications*. These are systems to which the theory has already been successfully applied and which are considered central to the theory. They form a subset of the *intended applications I*. The latter are (in a non-formalized sense) *similar* to paradigmatic applications.

With all the parts of the core and the intended applications now in place, the structuralist disposes of a seemingly powerful terminology to describe changes within one theory as well as connections across several empirical theories.

6.3 Conceptual Spaces

With this brief summary of structuralism as a background, we next turn to a presentation of our modeling tool. Conceptual spaces represent information by *geometric* structures rather than by set theory. Information is represented by points in the space (standing for objects or individuals), and regions (standing for properties and relations) in dimensional spaces. A great deal of the structure of a theory can be modeled in a natural way by exploiting *distances* in the space. These distances represent degrees of similarity between objects.

A conceptual space consists of a number of *quality dimensions*. Psychological examples of such dimensions connected to sensory impression are color, pitch, temperature, weight, and the three ordinary spatial dimensions. However, in scientific theories the dimensions are determined by the variables presumed by the theory. We have already noted that within NPM the relevant dimensions are three dimensions of space, time, mass and three dimensions of force.[11]

The primary role of the dimensions is to represent various "qualities" of objects in different *domains*. The notion of a domain can be given a more precise meaning by using the notions of *separable* and *integral* dimensions. These concepts are adapted from cognitive psychology (see e.g. Garner 1974, Maddox 1992, Melara 1992). In that context, certain quality dimensions are said to be integral if, to describe an object fully, one cannot assign it a value on one dimension without giving it a value on the other.

For example, one cannot give an object a hue without also giving it a brightness value. Or the pitch of a sound always goes along with its loudness. Dimensions that are not integral are said to be *separable*, as for example the size and hue dimensions. Within the context of scientific theories, the distinction should rather be defined in terms of *measurement procedures*. If two dimensions (or sets of dimensions) can be measured by independent methods, then they are separable, otherwise they are integral.

[11] Strictly speaking, forces need not be represented as separate dimensions. After all, $F = ma$. Therefore, the dimensions *mass*, *space* and *time* are sufficient.

In NPM, space and time are separable. In contrast, a relativity theory construes space-time as an integral set of dimensions. In the kinematics (descriptions of moving bodies) presupposed by Newtonian mechanics, the measurement procedures for location in space (measuring rods or optical signals) and time (pendulum motions, i.e., clocks) were considered to be independent of each other. By defining velocity and acceleration as the first and second derivates of position with respect to time, Newton proposed a theory that *coordinated* spatial and temporal measurement. In this sense, Newton too presupposes a theory of space-time. Yet, trigonometry and chronometry (the measurement of *distance* and of *duration*) were thought to be independent.

Using this distinction, the notion of a *domain* can now be defined as a set of integral dimensions that are separable from all other dimensions. In NPM, the domains are four in number: *space*, *time*, *mass* and *force*. The domains form the framework used to assign properties to objects and to specify relations between them (see below). The dimensions are taken to be independent of symbolic representations in the sense that we can represent the qualities of objects, for example by vectors, without presuming an explicit language in which these qualities are expressed.

The notion of a dimension should be understood literally. It is assumed that each of the domains (integral set of dimensions) is endowed with certain *topological* or *metric* structures.[12] It is part of the meaning of "integral" dimensions that they share a metric. Considering NPM, space is a three-dimensional Euclidean space, time is a one-dimensional structure that is isomorphic to the line of real numbers, mass is a one-dimensional structure that is isomorphic to the positive half-line of real numbers, and force is isomorphic to a three-dimensional Euclidean (vector) space.[13]

A consequence is that the topological structure of different quality dimensions entails that certain statements will become *analytically true*. For example it follows from the linear structure of the length dimension that comparative relations like "longer than" are *transitive*. This is thus an analytic feature of such a relation (*analytic-in-S*, that is). Similarly, it is analytic that everything that is green is colored (since "green" refers to a region of the color space) and that nothing is both green and blue. Analytic-in-S is thus defined on the basis of the topological and geometrical structure of the conceptual space S.[14] However, different conceptual spaces will yield different notions of analyticity, which leads to a form of *relativism* that would be foreign to a classical notion of analyticity.

[12] For examples of different topological and metric assumptions within psychological domains, see Gärdenfors (2000).

[13] We use "isomorphic to" rather than "homomorphically embedded in", as is the norm in standard accounts of measurement. Clearly, a very small difference, e.g., between in the lengths of two objects, will fall below the threshold of our cognitive capacity or our measurement apparatus, and is thus not measurable. Still, such differences are part of a conceptual space (see Batitsky 2000: 96).

[14] The "phenomenological" assumptions that Carnap (1971: 78f.) formulates are validated by the very structure of the space and need not be added as meta-linguistic constraints.

6.4 Correspondence Between Structuralism and Conceptual Spaces

After the brief accounts of structuralism and conceptual spaces, this section will show how most of the structuralist notions can be expressed in terms of conceptual spaces. Since our account does without the set-theoretic paraphernalia of the structuralists, we believe that it is more palatable for practicing scientists and fits better with their intuitions. Furthermore, conceptual spaces will allow us to highlight new aspects of the structure and dynamics of theories. In particular, we will show in Section 6.6 that our approach generates a new way of classifying theory changes.

For each of the five components of a theory core (see Section 6.2), we will present its correspondence in terms of conceptual spaces. Let us begin with the measurements (variables) that form the building blocks of a theory core. In general, they correspond to the domains of a conceptual space. For example, in the structuralist account, *space* and *time* were the two T-non-theoretical terms of NPM. The distinction between T-theoretical and T-non-theoretical terms is basically the same in conceptual spaces, except that we put emphasis on the separability of measurement procedures. The importance of this will show up in Section 6.6.3.

Secondly, let us consider the constraints on the theoretical variables. In general, they are determined by the assumptions concerning the metric (or scale) that is connected with the domain. NPM introduces as theoretical terms the variables *mass* and *force*. In the conceptual space of NPM, mass is a separable dimension isomorphic to the non-negative real numbers. The *extensivity* or *additivity* of mass, i.e., $m_1 + m_2 = m_{1+2}$ (where 1+2 denotes the object composed of objects 1 and 2), is accounted for by the assumption that, like many other physical magnitudes, mass is measured on a ratio scale (Stevens 1946, Ellis 1968).

Another constraint on the mass dimension is that it be *conservative*, which means that mass is a property of an object which is constant over different applications. *Force* is represented as three integral dimensions isomorphic to Euclidian 3-D space. A constraint on this variable is that the component forces of a body must be *independent* of the system to which the body belongs.

In NPM, any object (particle) is represented as a point in the eight-dimensional space spanned by the *space*, *time*, *mass* and *force* dimensions. Once an object has been assigned a value for all of these eight dimensions, it is fully described as far as the conceptual apparatus of Newtonian particle mechanics is concerned.

We then turn to the three kinds of models. Whenever the structuralist speaks of various models, the conceptual space framework generally speaks of sets of *vectors* or *points* in dimensional space. Each point represents the properties of an object. Thus, a *partial potential model* of NPM is a set of points (partial vector) in a 4-dimensional space (3-D for space, 1-D for time). A *potential model* is a set of points in the entire 8-dimensional space of NPM. Finally, a full model is a set of points (full vectors), the values of which satisfy the core axioms.

In NPM, the partial potential models are construed from partial vectors with values for space and time, while a full model involves vector values for all eight dimensions such that they always satisfy Newton's force law and may also satisfy a special law. In particular, Newton's second law $F = ma$ determines a hyper-surface in the eight-dimensional space. However, it should be noted that the law in itself only concerns single objects. Only when specialized laws are added will NPM introduce forces that connect several objects in an application.

The upshot is that once the conceptual space is specified and the core axioms formulated, the three kinds of models fall out very naturally as:

(i) sets of points in the subspace with values for T-non-theoretical dimensions (*partial potential model*);
(ii) sets of points in the space with values also for T-theoretical dimensions (*potential model*);
(iii) sets of points in the space with values for all dimensions such as to satisfy the core structure (and perhaps some special law) (*full model*).

As can be seen from this reconstruction, the three kinds of models, which play such a central role for structuralism, are not really required as separately specified entities when the conceptual space plus the theoreticity distinction are given.

How are the values for the various vectors determined? Just like the structuralists, we assume that the values for non-theoretical dimensions are determined by observations and that this is done by careful measurement according to established procedures. The values for the theoretical dimensions are obtained by presuming that the applied theory yields an empirically correct prediction. This parallel is not changed by taking a different view on the description of the conceptual apparatus of a theory.

As regards the intended applications, the framework of conceptual spaces has little new to offer.[15] Naturally, theories lend themselves to making predictions which are based on special laws. For example, the law of gravitation is applied to a given partial potential model of NPM, i.e., data on the time- and location-function of a number n of objects. In our way of speaking, this comes out as n trajectories in the 8-D space (described above) as constrained by the law of gravitation. It should be noted that, unlike Newton's second law, the law of gravitation introduces forces that *connect* several objects in an application.

Finally, the empirical claim of a theory is rendered as follows: *Any* partial vector (partial potential model) can be completed to a *set of points* (one for each object in the application) in the eight-dimensional space (potential model) such as to satisfy the constraints and axioms of the core (full model). In particular, in NPM the points are predicted to lie on the hyper-surface spanned by $F = ma$. Certain applications will also be expected to satisfy further special laws.

[15] However, it should be investigated whether the similarity naturally offered by distances in conceptual spaces can be exploited to give an account of the similarity of different applications.

Of course, at a given moment, one will start by considering only the finite number of special laws that have already been established. Yet, strictly speaking, *any* special law consistent with the frame-conditions will qualify, *including those not yet forwarded* (Stegmüller 1976: 105). Thus, our rendering of the empirical claim is just as *weak* as that offered by structuralism. This, we think, is as it should be. We have thereby shown that the components of an empirical theory as identified in structuralism can also be identified in terms of conceptual spaces.

6.5 Structuralist Change Operations and Their Limitations

Having shown the correspondences between structuralism and conceptual spaces, we shall next discuss the change operations that structuralism can identify and point to a limitation in the reconstruction of so-called *radical theory change*.

On the pragmatic side, structuralism can reconstruct the following changes by respecting the set of *intended applications I* as a part of a theory element: By correcting earlier measurements, the numerical values of a partial potential model for, say, Mercury's orbit within the application of NPM to the solar planet system – which had, so far, been *successfully completable* to a full model – may no longer be completable without offsetting a neighboring application, say, Venus's orbit (Roseveare 1982, Gähde 1997, 2002). In this case, a particular intended application, unless it is simply *retracted* from the set of intended applications, becomes an *anomaly*.[16]

Furthermore, Mercury's orbit (which had been calculated with the aid of Newton's law of gravitation) can "move" to a *new* theory element, by means of which one might calculate the application, after all. For example, this had been the case when a correction term to the exponent of Newton's gravitational law was proposed or when Clairaut or Hall proposed alternative gravitation laws featuring yet different correction terms (Roseveare 1982).

On the formal side, the basic change operations that structuralism offers are the *addition* or *deletion* of any of the parts that constitute a theory element, i.e., any change to the elements of the sets M, Mp, Mpp, C, L (see Section 6.2) – possibly including so-called *blurs* (Moulines 2002). Thus, the addition/deletion of an "entire" theory element to/ from the structure of a theory – e.g., the proposal of a new law which is specialized from an already present element – is merely a special case of the changes to the parts of a theory.

Note that, through the identification of the set of potential models via the satisfaction of the basic law plus constraints, there is – despite appearances to the contrary – not a lot of room for the structuralist to trace changes of a more *radical* kind. While the internal (logical) structure of a theory (as revealed through a

[16] A data structure which can no longer be successfully completed to a full model may itself be *hypothetically* enriched, such as to include so far unobserved additional objects. E.g., postulation of the planet Vulcan is such a case. This, however, does not seem to be a relevant change, because the new hypothetical data structure – *qua* also obeying $F = ma$ – had already been among the set of partial potential models. It had merely not been proposed as suitable for *theoretical* enrichment.

structuralistic reconstruction) may change in the most minute ways, every system to which *the same* theory is applied will be "forced" to co-operate within the conditions that are spelled out in the most basic theory element, such as $F = ma$ in NPM.[17] Thus, one defines the potential models of NPM into which we can "squeeze" a given kinematical system. At the same time, one thereby *excludes* any alternative which does not obey $F = ma$.

Clearly, structuralism provides a very fine-grained view on changes to *one* basic structure. However, in the case of a so-called *revolutionary theory-dislodgement* – e.g., the transition from *Newtonian Particle Mechanics* (NPM) to *Einstein's General Relativity* (GR) – the structuralist must resort to speaking of a *core-replacement*, because $F = ma$ simply is a basic law that is *not* valid in all of GR's potential models.[18] Speaking of core-replacement or *theory dislodgement* (Stegmüller) strongly suggests that there are "jumps" in theory evolution – an assumption we deny.

On our view, symbolic formulations of a theory, i.e., *equations*, specify *quantitative relations* between the ranges of values that this theory's terms can take. If two theories that quantitatively relate the *same* terms in a *different* way or – as is the more likely case – quantitatively relate a subset of these terms to *terms not included in the former theory*, one is well advised to take a step back from the equations. Instead, it is worthwhile to consider the conceptual spaces that the theories span. In this way – or so we submit – one may better understand how the old and the new space are connected.

This way of viewing the matter provides – we think – an interesting approach to the incommensurability issue that, however, we will not take up here.[19] For our present purposes, it is sufficient to have shown that we can fruitfully address theory-dislodgement. This very issue appears not to be answered satisfactorily by the structuralists' endeavor. *Nolens volens*, in speaking of core rejections, the structuralist will have to admit that she cannot trace the *continuities* between, e.g., NPM and GR by means of her reconstructive apparatus. Thus, she cannot reconstruct the transition, but only the theory's "initial and final sets of generalizations" (Kuhn 1987: 19).

[17] This is the sense in which Sneed can explicate "having a theory" as "being committed to use a certain mathematical structure, together with certain constraints on theoretical functions to account for the behavior of a, not too precisely specified, range of phenomena" (1971: 157).

[18] See Diederich (1996: 80) for the claim that "the classical problems of incommensurability have [thereby] been circumvented", rather than resolved. Also see Balzer et al. (1987, pp. 306–319) for attempts at tackling the incommensurability issue by relating two theories, *T* and *T**, through their sets *Mp* and *Mp**. For Kuhn's largely negative reaction to the structuralist's endeavor, see Kuhn (1976).

[19] Following Kuhn, incommensurability has been predominantly identified as a problem that occurs in relating the symbolic forms of two theories or frameworks. With our shift away from the symbolic and towards the conceptual level, we may end up not finding incommensurability at all. This sounds odd, but it is how it should be. After all, the practicing scientists that we know do not admit to any problems whatsoever in making a transition from one set of generalizations to the other.

6.6 Four Types of Theory Change Within the Framework of Conceptual Spaces

Compared with structuralism, we know of no comparable attempts in the literature at a similar formal representation of concrete cases of theory evolutions. However, this account of theory dynamics appears weak when we look at its ability to completely model radical theory change. The rather heavy set-theoretical apparatus that has been developed for the purpose gives little substantial insight into the processes of such changes of theories in return. In this section, our aim is to show that the framework of conceptual spaces fares better.

Given the notion of an empirical theory *as* a conceptual space (presented in section 6.4), changes to a theory core (including expansions) can naturally be divided into four types:

(i) addition and deletion of special laws;
(ii) change of scale or metric as well as the salience of the dimensions;
(iii) change in the separability of dimensions;
(iv) addition and deletion of dimensions which make up the space.

We shall argue that this ordering of the changes represents increasing *degrees of severity*. To show that these are generally applicable distinctions, we consider examples from the history of science. In the following, we discuss which change operation a case exemplifies. To be clear, the hypothetical completion of an application to one that features additional data does not constitute a change to the theory in question (see Section 6.5).

6.6.1 Addition of Special Laws

In general, the addition of a special law to a theory core only further specifies the class of models of the theory and thus increases the empirical content of the theory. Historically, it is not unusual that special laws, e.g., Hooke's law of the spring or the law of the pendulum, are formulated as specializations of the theory core *after* the latter had been specialized – in this case to the law of gravitation. Such a process only extends the range of applications of a theory core, without causing any significant change in the hitherto available theory structure.

Generally, the addition of special laws seems to be characteristic of what Kuhn (1962/1970) has called *normal science*. It should be regarded as the mildest form of change to an empirical theory since it does not involve any change in the theory core. In line with the expression 'expansion' from Sneed/Stegmüller, it could also be seen as an *expansion* of the theory core in the sense of belief revision (Gärdenfors 1988). In the present terminology, the dimensions presupposed in an empirical theory are simply used in new quantitative ways within the same qualitative space.

Special laws could also be *deleted*, but this is often not quite what happens when a theory core encounters anomalies. Rather, if an application of a special law results in predictions that do not fit the data, the application itself may be *retracted* from the set of intended applications and, possibly, be moved to a new theory in which it may be more successful. For example, Newtonian mechanics was once presumed to apply to the phenomenon of light. It was later realized that this would not work, whence light phenomena ceased to be intended applications of NPM. However, any "problematic" special law may persist as part of the theory without any application being assigned to it, as one may hope to find a new application for it in the future. Hence, special laws are never really deleted. Rather their intended applications may be temporarily suspended.[20]

6.6.2 Change of Metric/Scale

It is part of the description of a conceptual space to assign every domain (set of integral dimensions) its own metric. In theories that use classifications based on features that depend on several domains, the metrics for the domains must be weighed together. In other words, their relative salience must be determined (see Gärdenfors 2000, Section 4.7.2). For example, pre-Linnaean botany was based on holistic features of the flowers, such as *size* and *color*, while Linnaeus' classification made the *numbers of pistils* and *stamens* the most salient features. What is involved in this kind of theory change is the principle for *combining* different domains of a conceptual space.

However, even *within* a single domain there may occur changes of metric. It is trivial that temperature can be measured on both the Celsius and the Fahrenheit scales. These scales are equivalent since they both involve an interval scale (invariant under all positive linear transformations (Stevens 1946, Ellis 1968). However, temperature can also be measured on the Kelvin scale, which is stronger since it is a ratio scale (and thus has less invariance). The change from Celsius to Kelvin leads to different predictions concerning temperature. It is part of the theory associated with the Kelvin scale that no object can have a temperature below absolute zero, while no such prediction could be made only by assuming that temperature is measured on an interval scale. Thus a change of scale can lead to a change in the empirical contents of the associated theories of temperature.[21]

Another example of a more severe form of changing metric is obtained when, in reaction to experimental findings in early chemical theory, it was argued that the "fire substance" (*phlogiston*) would need to have *negative* mass in order for the theory to square with experience. Here, the negative range of the mass scale needs to be introduced – a rather radical, but possible move to bring the theory in line with the empirical results.

[20] Of course, such changes are traceable and, therefore, not without repercussions in the theory. Thus, if T_n is the respective theory element, its set of applications $I(T_n)$ will simply be zero.

[21] A science-historical account of the process leading to a current concept of temperature as *mean kinetic energy* is provided in Chang (2004).

A change of metric involves changes of the predictions of the theory and is thus a form of *revision* in the sense of Gärdenfors (1988). It is therefore a more drastic change than adding new special laws. However, the change is still relatively mild, since the basic framework of the conceptual space and the core axioms are maintained.

6.6.3 Change in Integrality or Separability of Dimensions

In NPM, time and space are separable domains. A remarkable change occurred as a reaction to the Michelson Morley null result on ether drift. It was hypothesized that the rods by which one measured length are *shortened* in the direction of the ether drift, resulting in the values of the Lorentz–Fitzgerald contraction. Effectively, one thereby "squeezed" the length-scale to account for the null result within an ether theory.

A most basic assumption today is that light signals propagate at a finite velocity. An ether theorist, on the other hand, expected a difference in the speed of light signals as a function of their direction of motion relative to the ether. To uphold the ether hypothesis after the null result, it was proposed that the rod – along which the light beam traveled and then returned after being reflecting by a mirror positioned at the end of the rod – was *shortened* in the direction of the "ether-wind". Moreover, shortened just enough to let the (predicted, yet unobserved) drag of the ether-wind onto the light beam cancel out.

This is an effective way of interpreting an experiment in favor of one's theory. However, it can hardly be called a plausible hypothesis, if – as Einstein did – one doubts that there be (any necessity for) an ether to begin with and is committed to the constancy of lengths on observational grounds. In fact, Einstein's solution does not presuppose a *mechanism* by which the length-scale is squeezed. From suitable assumptions, he could rather deduce the contraction factor such that there will be only an *apparent* contraction.[22]

The Lorentz–Fitzgerald contraction combines the domains of space and time into 4-dimensional space-time. This seems to be an exceptional case of integrating domains within the framework of a theory core (i.e. NPM). In general, dimensions are not separated, nor are unconnected dimensions integrated, unless some more severe change also takes place in the form of adding or deleting dimensions.

6.6.4 Addition and Deletion of Dimensions

The most fundamental change of a conceptual space occurs when dimensions are added or deleted. Most, perhaps all, of the historical changes Kuhn (1962/1970)

[22] If L_0 is the length of an object in a rest frame, and L_1 the length measured by an observer, then the contraction is given by $L_1 = L_0/\gamma$, where γ is defined as $\gamma = \sqrt{}(1 - u^2/c^2)^{-1}$ (with u for relative velocity between observer and object, and with c for the speed of light).

calls revolutions can be analyzed as changes in the fundamental dimensions of a scientific area.

A paradigmatic case for the addition of a dimension is Newton's introduction of the *mass* dimension as distinct from the *weight* dimension of Galilean physics and adding the dimension of *force* in his mechanics. Given the distinction between weight and mass, an object of a given weight is now analyzed as an object of a given mass under the influence of a given gravitational force.[23] Effectively, the weight dimension is deleted and replaced by the separate dimensions of mass and force, which function as theoretical variables in Newton's theory.

For a second example, in order to save electro-magnetic phenomena, Einstein introduced the energy dimension as a theoretical variable in his relativity theory and eliminated force as a fundamental dimension. Energy, in the form of kinetic energy, was a derived variable of NPM, but in GR it becomes a fundamental variable.

As a third example, following Chen (2003), one can characterize the particle theory of light by assuming that it postulates *at least* two integral dimensions for *velocity* and *size* (both taking continuous values) and one dimension for *side* (taking a binary value), separable from velocity and size. In a wave theory of light, the *velocity* dimension remains, but it now becomes integral with the dimensions *amplitude* and *wavelength*, while the dimension of *size* is deleted and replaced by *phase difference* (also taking a binary value).

6.6.5 Discussion

We will leave a full discussion of the incommensurability issue for future work, because we are first required to have a good *definition* of incommensurability. The extent to which our approach provides an interesting answer to this issue will largely be a function of the definition we chose.

However, the main point may already have been anticipated: If one accepts as plausible the idea of literally *developing* a conceptual space into a new one by the change operations identified above, then the description of this development as one of the above four types of change *is* the answer to the incommensurability issue. In other words, the traditional *problem* of incommensurability – finding ruptures between the symbolic forms of predecessor and successor theory – is a consequence of treating the symbolic level of representation as primary. On the conceptual level, the four kinds of change operations establish the continuities between an old and a new theory.

It appears to us that a set-theoretical apparatus (such as structuralism) is prone to conceal insights among precise yet cumbersome formulations. This is especially so in cases of so-called large-scale changes, i.e., theory dislodgement, for which, in our

[23] In 1901, the *Bureau International des Poids et Mesures* conventionally defined as much "to put an end to the ambiguity which in current practice still exists on the meaning of the word *weight*, used sometimes for *mass*, sometimes for *mechanical force*" (BIPM 1901: 70).

opinion, structuralism only offers a pseudo-reconstruction by offering the concept of a core-rejection. This ultimately seems to miss out on tracing continuity.

We have provided several examples of theory dynamics that can be analyzed in terms of changes of conceptual spaces; other examples need to be studied. The reader might have noticed that we do not put much stress on the mathematical symbolism that usually accompanies accounts of theory change. In fact, we suggest that the inclusion of such symbolism in the presentation of the history and philosophy of science is a myopic outgrowth of established standards fostered by a philosophical training which is almost exclusively devoted to the symbolic level of representing theories and other forms of knowledge.

This, we hold, should not persist as the *only* framework in which to discuss or understand changes to empirical theories. The formulas themselves do not reveal the underlying assumptions concerning the variables involved: their geometrical structures, their integrality and their determining measurement procedures. Of course, studying formulas is indispensable in the analysis of empirical theories, but understanding conceptual change should not proceed exclusively on this level.

It is important to realize that the conceptual structure of an empirical theory as well as its dynamics can be approached by *abstracting* from the quantitative relations (formulae) and by focusing on the qualitative relations between the terms postulated by the theory – in our terminology: on the *dimensions* that constitute an empirical theory. We do not thereby deny that exact science should be expressed in the language of mathematics, but we deny that insights into the conceptual development of theories are generated by staring at formulas. Rather, insights arise from having a clear and simple geometrical conception of what it means to be a (separate or integral) dimension and from having defined change operations which, when applied to the space, can transform it into a new one. Furthermore, in our opinion, the role of the measurement procedures associated with the central variables has been underestimated.

Taking stock of what is gained by our approach when compared with structuralism, then – a full answer on the incommensurability issue pending – we claim that a reconstruction of empirical theories in terms of conceptual spaces will generate the insights of structuralism (without the set-theoretical apparatus) and provides a fruitful way of describing different kinds of theory changes which goes beyond what is possible within structuralism. Already one can say that our approach bears evident benefits for educating a wider audience in theory change. Future work should treat additional cases, the relation between measurement and the separability of dimensions, and the similarity relations between applications that can be formulated in terms of conceptual spaces.

6.7 Conclusions

Conceptual spaces allow us to present a more unified view than structuralism of empirical theory cores and their expansions and, in particular, theory-dynamical aspects in a way that fits better with actual scientific practice. The geometrical notions developed in Gärdenfors (2000) provide insights into how a theory develops by classifying changes according to the operations that have been identified in the previous section. On our account, the fact that the symbolic formulations of a theory change over time is but an effect of the dynamics of the underlying conceptual space. While we believe that our approach will allow a more fruitful approach to the problem of incommensurability than structuralism has been able to offer, we must leave this to future work.

We have shown how the structuralist's set theoretical constructs can find their correspondences in the theory of conceptual spaces. In particular, it poses no major difficulty to recover the T-theoreticity distinction and account for the distinction into partial potential, potential and full models of a core (and its expansions). Basically, these distinctions are reached by separating among a theory's dimensions those that are grounded in antecedently available measurement processes from those that are not.

In general, the significance of measurement procedures for our account is the following: When it comes to Kuhn's revolutionary change, we hypothesize that any introduction of a new or any deletion of an old dimension will also reveal a change in the measurement procedure. In this sense, the measurement procedures also come out as the pragmatic links between concepts and the empirical world that they represent.

Acknowledgements We would like to thank the organizer and the audience of the Science in Flux workshop at Lund University and, especially with respect to our exposition of structuralism, J. D. Sneed and C. U. Moulines for helpful comments and criticism. Frank Zenker acknowledges funding from the *Swedish Institute* (SI) and Peter Gärdenfors from the Swedish Research Council.

References

Balzer, W., D.A. Pearce, and H.-J. Schmidt. 1984. *Reduction in science. Structure, examples, philosophical problems* (Synthese Library Vol. 175). Dordrecht: Reidel.

Balzer, W., C.U. Moulines, and J.D. Sneed. 1987. *An architectonic for science. The structuralist program*. (Synthese Library Vol. 186). Dordrecht: Reidel.

Balzer, W., C.U. Moulines, and J.D. Sneed. (eds.). 2000. *Structuralist knowledge representation. Paradigmatic examples*. Amsterdam: Rodopi.

Batitsky, V. 2000. Measurement in Carnap's late philosophy of science. *Dialectica* 54:87–108.

BIPM. 1901. Comptes Rendus de la 3^e Conférence Générale des Poids et Mesures, Paris 1901. http://www1.bipm.org/en/CGPM/db/3/2/

Carnap, R. 1971. A basic system of inductive logic, part 1. In *Studies in inductive logics and probability, vol. 1*, eds. R. Carnap, and R.C. Jeffrey, 35–165. Berkeley, CA: UCP.

Chang, H. 2004. *Inventing temperature. Measurement and scientific progress*. Oxford: OUP.

Chen, X. 2003. Why did John Herschel fail to understand polarization? The differences between object and event concepts. *Studies in the History and Philosophy of Science* 34:491–513.

Diederich, W. 1996. Pragmatic and diachronic aspects of structuralism. In *Structuralist theory of science. Focal issues, new results*, eds. W. Balzer, and C.U. Moulines, 75–82. New York, NY: De Gruyter.

DiSalle, R. 2006. *Understanding space time. The philosophical development from Newton to Einstein*. Cambridge: CUP.

Ellis, B. 1968. *Basic concepts of measurement*. Cambridge: CUP.

Gähde, U. 1997. Anomalies and the revision of theory-elements: Notes on the advance of Mercury's perihelion. In *Structures and norms in science. vol. 2* (Synthese Library Vol. 260), eds. M. L. Dalla Chiara. et al., 89–104. Dordrecht: Kluwer.

Gähde, U. 2002. Holism, underdetermination, and the dynamics of empirical theories. *Synthese* 130:69–90.

Gärdenfors, P. 1988. *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge, MA: MIT Press.

Gärdenfors, P. 2000. *Conceptual spaces*. Cambridge, MA: MIT Press.

Garner, W.R. 1970. *The processing of information and structure*. Potomac: Wiley, New York, NY: Erlbaum.

Kuhn, T. 1962/1970. *The structure of scientific revolution*. Chicago: CUP.

Kuhn, T. 1976. Theory change as structure change. Comments on the sneed formalism. *Erkenntnis* 1**0**:179–199.

Kuhn, T. 1987. What are scientific revolutions? In *The Probabilistic revolution. vol. 1*, eds. L. Krüger, et al., 7–22. Cambridge: MIT Press.

Maddox, W.T. 1992. Perceptual and decisional separability. In *Multidimensional models of perception and cognition*, ed. G. F. Ashby, 147–180. Hillsdale, NJ: Lawrence Erlbaum.

Melara, R.D. 1992. The concept of perceptual similarity: From psychophysics to cognitive psychology. In *Psychophysical approaches to cognition*, ed. D. Algom, 303–388. Amsterdam: Elsevier.

Moulines, C.U. 2002. Introduction: Structuralism as a program for modeling theoretical science (special issue on structuralism). *Synthese* 130:1–11.

Roseveare, N.T. 1982. *Mercury's perihelion from LeVerrier to Einstein*. Oxford: Clarendon.

Sneed, J.D. 1971. *The logical structure of mathematical physics*. Dordrecht: Reidel.

Stegmüller, W. 1976. *The structuralist view of theories*. Berlin: Springer.

Stevens, S.S. 1946. On the theory of scales of measurement. *Science* 103:677–680.

Suppes, P. 1957. *Introduction to logic*. New York, NY: Van Nostrand.

Chapter 7
A Note on Theory Change and Belief Revision

Bengt Hansson

7.1 Conceptual Closure

For a long time, scientific theories were usually characterised (by philosophers) as sets of laws, holding of the world. In the last decades, this view has been repeatedly challenged, with structures,[1] capacities,[2] mechanisms,[3] and perhaps also ontologies and processes competing to take the place of laws. But rather than taking part in that debate, I wish to look into a matter that is epistemologically prior to it, namely how various sciences select, define, and develop their concepts.

If this topic is dealt with at all in the law view, it is usually by some general remarks that concepts are implicitly defined by the laws, thus making them secondary to these laws. My claim is instead that concepts must come before laws, that they are involved already in the conception of the identity of a particular science, and that it is only the fine grinding that remains when laws (or structures, or capacities, or mechanisms, or ontologies, or processes) come into the picture.

For the aim of a particular science is not to describe or explain the entire world, but only one special aspect of it: physical, chemical, biological, economic, political, or organisational, as the case may be. And it separates that aspect from others,

B. Hansson (✉)
Lund University, Lund, Sweden
e-mail: bengt.hansson@fil.lu.se

[1] Structuralism or the non-statement view, originating in works by Patrick Suppes and developed by Joseph Sneed. For an overview see *An architectonic for science: the structuralist program* by Wolfgang Balzer, C. Ulises Moulines and Joseph D. Sneed, Dordrecht 1987. The theory has since been further developed by several German philosophers. For a more recent account, see *Structuralist theory of science: Focal issues, new results*, ed. by Wolfgang Balzer and C. Ulises Moulines, Berlin 1995.

[2] See Nancy Cartwright's *Nature's capacities and their measurement*, Oxford 1989.

[3] See e.g. Jon Elster's *Nuts and bolts for the social sciences*, Cambridge 1989.

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_7,

in the first approximation, by the sort of phenomena it intends to describe and explain: physicists are interested in matter and motion, chemists in transformations of substance, economists in the production and exchange of goods, and so on.

In order to capture the essence of these phenomena, the sciences develop concepts, carefully selected and adjusted to form a coherent whole with certain closure properties. Let us, by way of introduction, take a brief and simplified look at physics:

To describe motion you need the concepts of position and time. For certain purposes this has been considered sufficient, for example for recording planetary movements before Kepler. But for explanatory purposes you need to add forces as a kind of invisible capacities that cause motion or change in motion. And in order to specify the size of forces you also need the concept of mass.

We thus arrive at full Newtonian mechanics. It forms a closed whole in the following sense: as long as you are interested in how objects move, you need to know nothing but positions, times, forces and masses, and if you need to calculate any of these magnitudes it is sufficient to know the other ones (and their rates of change). The fact that some derived concepts, like energy and momentum, are helpful in explaining some transformations changes nothing in principle. It seems as if Newtonian mechanics succeeded in delineating a closed substructure of the world.

The closure condition on a theory is all-important for its applicability and, I think, a neglected topic in the philosophy of science. If only some changes of motion were caused by forces and not all, or if masses only sometimes exerted gravitational pull, then the laws of mechanics would be inapplicable to the real world; there would always be the possibility of an unexplained outside influence. A successful theory somehow marks off a closed subworld. It is usually not intended to apply to the whole world – there are other phenomena than motion in the world – but if it is so intended, all other features of the world must supervene on the features described by the theory.

The closure condition on a theory can be mathematically described in terms of homomorphism and commutative diagrams. Let us imagine different stages of the full, real world arranged horizontally, like in the diagram below, with world w being transformed into world u by some function m_w reflecting motion in the real world. We thus have $u = m_w(w)$.

Let us further assume that each world is described as completely as possible in the theory's language. This mapping, from worlds to descriptions, is denoted by the function d, so that $d(w)$ is everything the theory can say about the world w. We assume that this description function will to be a homomorphism, but not necessarily an isomorphism, that is, it will preserve those structural features of a world that the theory is intended to account for, but not necessarily cover every aspect that world. There may well exist distinct w and w' with $d(w) = d(w')$, so many different real worlds may have identical descriptions – the language of a scientific theory is normally construed to capture only those features of the world that belong to that science, and worlds that differ in other respects need not be distinguishable in that language.

Diagram with tracking function m_t

$$\begin{tikzcd}
w' \arrow[dr] & w \arrow[r, "m_w"'] \arrow[d, "d"] & m_w(w) \arrow[d, "d"] \\
& d(w) \arrow[r, "m_t"'] & m_t(d(w))=d(m_w(w))
\end{tikzcd}$$

The closure condition can now be expressed in the following way: there exists a function m_t from descriptions to descriptions which reflects motion within the theory and which correctly tracks motion in the real world, i.e. such that $d(m_w(w)) = m_t(d(w))$. The above diagram should be read accordingly: We start with a given world w (up left), describe it in the theory's vocabulary (go down to $d(w)$), and calculate its motion according to the theory (go right to $m_t(d(w))$). Then we end up in the same place as we would if we had first observed the real motion (gone right from w to $m_w(w)$ and then described the resulting state (gone down to $d(m_w(w))$).

There are two essential points about this closure condition: the practical one that the theory can correctly track the real world and therefore be used for predictions, and the theoretical one that we feel satisfied that our theory has captured everything there is to say about motion.

7.2 What's in a Law?

In this connection it might be useful to distinguish between two types of Newtonian laws. Some are completely general and existentially uncommitted, like the force equation, saying that the force acting on a body is equal to its mass times its acceleration. Although the equation as such is time- and directionless, it is obvious that it is meant to convey the idea that the existence of a given force causes a mass to have a certain acceleration (and not that the presence of an acceleration causes a force to come into existence). The law can therefore be regarded as an implicit definition of the existence and magnitude of a force, refining the previous qualitative concept of "force" as that which causes change of motion. You can also say that it is structural in character, spelling out an internal theoretical relation between the three basic concepts time, (derivative of) position and change. It is existentially uncommitted in that it does not entail the existence of any particular force, or indeed the existence of any force at all, for it is compatible with a world void of accelerations.

Other laws make existential claims, like the law of gravitation, saying that between two masses there exists a force which is proportional to each of the masses and inversely proportional to the square of the distance between them. Here it is clear that it is the masses that bring about the force, and the law asserts not only that a force of specific type exists (given the existence of masses) but also specifies its magnitude. In a world where there exist masses it necessitates the existence of forces. While acceleration was merely a *sign* of force in the case of the force equation, and therefore ontologically secondary to it, masses are ontologically prior to forces in the case of the law of gravitation.

We have thus found that there is more to a law than meets the eye, and therefore also to theories containing them. In two respects, they have additional implicit content over and above the numerical relationships expressed by the formulas. The first is that some laws involve no existential commitments but serve to establish connexions between basic concepts of the theory, whereas others specify how and when these concepts are instantiated in the world. The other is that the laws implicitly define an ontological and causal order of priority between the concepts, which, in the case classical mechanics, goes from masses to forces to motions.

The theory I intend to develop has two parts. First, that the idea of conceptual closure is a governing principle for theory change and the development of science. Secondly, that when theories change, the first choice is to add or amend laws of the second kind (for example to admit new kinds of forces, or that forces occur in new situations) whereas the first type is shielded as long as possible. In the language of ontological priority it means that priorities are not reversed and that amendments are preferably made early in the chain rather than later.

7.3 Historical Examples

To substantiate my outline by careful case studies is a book-size project. Here I must limit myself to some brief examples.

It is natural to start with the simplest case, namely how to predict a time series from its previous behaviour. The prime example is the motion of the heavenly bodies. Astronomy, from Babylonian time all the way up to Copernicus, was basically a mathematical exercise, designing numerical models, expressed as epicycles, to fit a vast amount of observations without much regard to the physical reality behind. The enterprise had remarkable success in terms of accuracy – in fact, even today calendars are calculated by the same mathematical methods – which no doubt created a belief that the heavenly bodies formed a system closed in itself, exhaustively described in terms of positions at various times.

It is significant that the man who broke with this tradition and began to think of physical driving forces, namely Kepler, did so only after a tremendous effort to remedy the old system, and even then he expressed his results as purely numerical and geometrical relations between for example speed and position in the orbit, concealing the intuition that had led him to his discoveries, namely that the sun exerted some kind of pulling force on the planets.

The idea of a time series containing all the necessary information in itself is still living. The so-called Box-Jenkins analysis of time series, popular in the 1980s, relied on the calculation of autocorrelations in the series' past to predict its future, and the activity known as stock charting has a similar structure. In the latter case, one often hears a justifying argument to the effect that, since all the relevant information is already discounted by the market, there is nothing else that matters than the market data themselves. This seems to be the error of believing that the existence of the functions m_w and d in the diagram above automatically guarantee that the theoretical tracking function m_t also exists.

The further development of classical mechanics gives additional support to the idea of conceptual closure. Potential and kinetic energy were separately definable in the basic concepts, and the conservation of their sum could be established in many useful situations, and could, for example, be used to calculate the speed of a falling object when it hits the ground by regarding the fall as a gradual transformation of potential energy into kinetic. But after the fall both types of energy seem to have vanished, and their sum is zero. The problem was solved by the introduction of heat as new kind of energy, restoring closure in this respect.

Electrostatics is another case in point. Charged particles move and accelerate in ways that cannot be accounted for by gravitational forces. If the force equation, a structural law, is to be retained, there must exist new types of forces. Therefore the concept of charge is introduced together with a new existential law, Coulomb's law, parallel to Newton's law of gravitation, thus preserving closure.

When charges and their associated fields move they produce magnetic phenomena, and when magnetic fields move they produce electric currents. The theory of electromagnetism is often seen as a paradigmatic case of unification of these two phenomena. But it can just as well, or perhaps even better, be seen as a move to achieve closure. Without magnetism, electric phenomena were not all derivable within the system, and without electrodynamics the same held for magnetic phenomena. Many cases of unification can, in a similar fashion, be seen as special cases of closure.

From these sketches there emerges a pattern for the normal development of a science as it grows more mature and expands to cover wider fields of application.

First some phenomenon appears that threatens the closure of the received theory. If it is a stable and recurring phenomenon, the preferred strategy is to introduce new concepts and thereby making the theoretical description of states more fine-grained. It means that the homomorphism function d in the diagram will have more complex values, and the crucial question becomes whether the added structure at one point will suffice to determine the value of the new concept at a later stage, i.e. whether there will still exist a tracking function m_t.

If the phenomenon itself, for example a movement due to no known forces, can be described in the language of the old theory, as is usually the case, then the new theory will be a non-conservative extension of the old one.

When a new concept, for example charge, is added to the theory, it usually means that some other concept, like force, gets its range extended. While this may be a trivial matter in itself, being merely extensional and not affecting meaning, it may be vital for the formulation of for example conservation laws, like the conservation of total energy.

Sometimes, however, more radical revisions are necessary. This is a matter that needs a thorough discussion with varied examples and which therefore cannot be pursued in this note. About the most natural continuation of the above series of examples, namely the severely over-discussed transition from Newton's to Einstein's mechanics, I will simply note that, while the concept of rest mass can still be said to have ontological priority, the relativistic mass cannot, because it depends on velocity. Yet, it is the relativistic mass that is used in the force equation and which therefore determines acceleration. Closure is bought at a higher price, which may be the reason why this transition is regarded as particularly fundamental.

7.4 Revolutions and Reductions

Kuhnian revolutions are triggered by anomalies. Yet Kuhn has difficulties in explaining what exactly it is that makes an anomaly an anomaly rather than an unusually hard nut yet to be cracked. In short his answer is that it has resisted so many attempted solutions that scientists in general have given up the hope to solve it within the current theory. It is all in the eye of the researcher.

In the present framework an anomaly can be seen as a phenomenon that threatens the conceptual closure of a theory, suggesting that something essential is lacking in that theory. It need not lead to a revolution if the theory can be remedied by comparatively simple means, like in the case of charge and Coulomb's law. But if structural laws cannot be upheld, or the ontological priority of basic concepts needs revision, then the required revision challenges the received way of thinking in a much more fundamental way, and might well be called a revolution.

Another Kuhnian notion is that of incommensurable concepts. Incommensurability is sometimes presented as a necessary ingredient in scientific revolutions and sometimes as a kind of gradual conceptual drift. In both cases incommensurability is supposed to prevent mutual understanding. In the present framework it is possible to discuss such matters in a more precise way. For example, there seems to be little reason to assume that any conceptual change is involved in the addition of new kinds of forces. Rather, it is adherence to the original idea of a force as that which "causes motion or change of motion" that necessitates the addition of new kinds of forces, for else there would be uncaused motions. But if structural laws or ontological priorities have to be adjusted, this is a much more significant change in our perception of the clockwork of the world.

The present framework may also help in understanding why reduction of one science or one theory to another has often been considered desirable, at least as a matter of principle. The typical case is when it is felt that one science is a bit shaky in its foundations, perhaps because there is no conceptual closure or otherwise unclear relations between its basic concepts, and that a reduction to a more solid science would give greater stability.

If by reduction is meant both that all concepts of the reduced theory are defined in terms of the reducing theory and that the laws of the reduced theory then follow analytically from the laws of the reducing theory, then closure of the reducing theory automatically extends to the reduced theory. It does not follow, however, that the derived laws immediately suffice for an explicit way to express the tracking function m_t. If, however, reduction is thought if in a weaker way, where the laws of the reduced theory are merely *expressed* in terms of the reducing theory but justified by other means, for example by empirical induction, then closure remains to be established independently, so reduction will not be an automatic gain in this respect.

7.5 Application to Belief Revision Theory

One could argue that there are some structural similarities between belief revision theory and classical mechanics. The "objects" of belief revision theory are (complete) epistemic states, which may be subject to various types of impacts or "acting forces", like perception, reflection or linguistic messages, which push them into new states.

In the most general terms, the task for a theory of belief revision would therefore be to find, first, a conceptual representation of epistemic states and, secondly, an exhaustive list of types of impact and ways to represent them as functions from states to states, preferably satisfying closure.

The following brief analysis is intended to show that the problems with belief revision theory are more fundamental than is usually assumed. It is not merely a question of making minor adjustments to the conditions in the customary framework, but it is necessary to address the basic ontological question of what the essential features of an epistemic state are and how these determine how the state is transformed when exposed to certain impacts. This is necessary before the states can be modelled in sufficient detail to allow conceptual closure.

The most natural way for many philosophers would be to try to represent epistemic states simply as sets of propositions. Then the various types of impacts would be represented by functions from sets of propositions to sets of propositions. Examples would be the functions usually known as expansions, contractions, and revisions, only that they are often not fully specified as functions but only as limiting conditions.

It is rather obvious, however, that closure fails, at least for contractions and revisions. There are two ways to react to this. Either one tries more richly structured representations than sets of propositions, reflecting more fully the concrete manifestation of an epistemic state, or one keeps propositions as a surface structure but adds some other conceptual component, like a deep structure that does not show in the actual state but governs conditional statements and revisions.

A small step in the first direction would be to use Quine's metaphor of a man-made fabric, reflecting pedigree and net connections, perhaps using the concept of coherence or some other holistic concept. It would, however, take quite an effort to raise this approach above the metaphorical level.

There are two commonly used approaches in the other direction, to add a deep structure, namely orderings which reflect proximity of some kind, and gradings which reflect degrees of belief or disbelief.

Grading belief is tantamount to a kind of probability, but ordinary probability measures do not satisfy closure.[4] So-called rank functions, as introduced by

[4] See my "Infallibility and incorrigibility", in *Knowledge and inquiry*, ed. by Erik J Olsson, Cambridge University Press 2007 for an argument to this effect.

Wolfgang Spohn,[5] are a better variant in this respect since they satisfies closure. But it can be argued, with reasons similar to those for common probability measures in the previous foot-note, that it suffers from too rigid an approach to conditionalisation.

I introduced proximity orderings as a means to solve the problem about conditionals in deontic logic in the late 1960s.[6] There is no reason why the same basic idea should not work for other types of conditionals too, and the idea was picked up in the theory of non-monotonic logics in the 1970s,[7] and again in belief revision theory under the name of "entrenchment" in the late 1970s.[8] The problem with this approach is that there is no built-in guarantee that it satisfies closure.

The idea behind proximity ordering as a means to deal with conditionals is that if a condition is not satisfied in one world, then you can go to the closest one where it is satisfied and look how things are there. It is therefore tantamount to the idea of minimal change, also discussed by David Lewis as systems of nested spheres.[9]

Why, then, minimal change? Despite my previous involvement I have grown sceptical about the idea. In fact, minimal change seems to me to have a close relationship with ad hoc reasoning. "Find a fixing for the immediate problem and don't touch the rest!" is very akin to "Change as little as possible!". Deductive closure can easily be achieved, but not conceptual closure. Rather, the immediate reaction to an anomaly should be to *explain* it. Depending on one's favourite theory of explanation, this may mean different things, but on my own account[10] it would mean spelling out conceptual connexions with the main body of one's knowledge. And certainly the best explanation would be the one that fitted the anomaly into a conceptually closed system!

The problems encountered by adding deep structure to propositional representations therefore suggest that the most promising approach to further development of cognitive dynamics is to develop a richer type of models. The propositional content is only the visible surface of an epistemic state, and there lurk far more complex things below than mere orderings or their equivalents.

[5] Gradually developed in many articles, beginning with "Ordinal conditional functions. A dynamic theory of epistemic states" in *Causation in decision, belief change, and statistics*, ed. by W.L. Harper and Brian Skyrms, Dordrecht 1988.

[6] See my "An analysis of some deontic logics". *Noûs* vol. 3 (1969), pp. 373–398. Reprinted in *Deontic logic: introductory and systematic readings* (ed. by Risto Hilpinen), pp. 121–147. Dordrecht 1971.

[7] For an overview, see e.g. several articles in *Defeasible deontic logic* (ed. by Donald Nute), Boston 1997.

[8] Published rather late in Peter Gärdenfors' *Knowledge in flux*, Cambridge (MA) 1988, although the original ideas were developed in the late 1970s in a frequently meeting discussion group with Gärdenfors, Nils-Eric Sahlin and myself as the regular members.

[9] See e.g. Lewis' *Counterfactuals*, Blackwell 1973.

[10] See my "Why explanations? Fundamental, and less fundamental ways of understanding the world". Theoria vol. 72, part 1 (2006) or "Explanations are about concepts and concept formation", in *Rethinking explanation*, ed. by Petri Ylikoski and Johannes Persson, Springer 2007.

Chapter 8
Social Norms, Rational Choice and Belief Change

Horacio Arló-Costa and Arthur Paul Pedersen

8.1 Introduction

The classical theory of rational choice as developed by mathematical economists such as Kenneth Arrow (1951, 1959), Marcel K. Richter (1966, 1971), Paul Samuelson (1938, 1947), and Amartya Sen (1970, 1971), has occupied a central role in the philosophy of the social sciences for almost a century now. Jon Elster has provided a convincing argument for the applicability of the theory to the social sciences in various books and articles (see, for example, Elster 1989a). The theory also occupies a central foundational role in the contemporary theory of belief change.

In fact, recently, Hans Rott (2001) has shown how so-called rationality postulates of belief change and non-monotonic reasoning correspond in a one-to-one fashion to so-called coherence constraints of classical theories of rational choice. In particular, Rott provides a connection between classical coherence constraints on selection functions of rational choice and rationality constraints on operators of belief change and non-monotonic reasoning. A recurrent feature of formal theories of belief change and non-monotonic reasoning is that some extralogical selection function is employed to do the dirty work, and Rott's study rigorously exhibits how this feature affords a nexus between formal theories of belief change and non-monotonic reasoning.[1] But more importantly, Rott's work shows that belief

H. Arló-Costa (✉)
Department of Philosophy, Carnegie Mellon University, Baker Hall 135, Pittsburgh,
PA 15213, USA
e-mail: hcosta@andrew.cmu.edu

We wish to thank Isaac Levi, David Makinson, and Teddy Seidenfeld for valuable comments on earlier versions of this chapter.

[1] The relationship between formal theories of belief change and non-monotonic reasoning has been well examined in Gärdenfors and Makinson (1994) and Makinson and Gärdenfors (1991). For arguments calling into question the claim that the AGM theory of belief change and the Rational Logic of Kraus et al. 1990 are two sides of the same coin, see Arló-Costa (1995).

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic,
Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_8,

change and non-monotonic reasoning can be investigated in the formal framework of rational choice.[2]

Amartya Sen, who contributed crucially to the development of the received view of rational choice, has also been one of its main critics. In fact, in a series of articles Sen (1993, 1996, 1997) has argued convincingly against the *a priori* imposition of requirements of internal consistency of choice, such as the weak and the strong axioms of revealed preference, Arrow's axiom of choice consistency, and Sen's Property α. The following example, presented in Sen (1993), can give the reader an idea of the difficulties that Sen has in mind:

> Suppose...[a] person faces a choice at a dinner table between having the last remaining apple in the fruit basket (y) and having nothing instead (x), forgoing the nice-looking apple. She decides to behave decently and picks nothing (x), rather than the one apple (y). If, instead, the basket had contained two apples, and she had encountered the choice between having nothing (x), having one nice apple (y) and having another nice one (z), she could reasonably enough choose one (y), without violating any rule of good behavior. The presence of another apple (z) makes one of the two apples decently choosable, but this combination of choices would violate the standard consistency conditions, including Property α, even though there is nothing particularly 'inconsistent' in this pair of choices . . . (p. 501).

This example, as with many others offered by Sen, involves a *social norm*, in this case a rule of politeness that seems to make the option y unavailable for the person when she is faced with the first choice. As Sen indicates, the combination of choices the person takes in this example violates standard consistency conditions (also known as *coherence constraints*) such as condition α, which demands that whatever is rejected for choice must remain rejected if the set of alternatives available for choice is expanded. Examples like this one have led many to believe that the standard rationalizability approach to the theory of choice faces serious difficulties coping with the existence of external social norms.[3]

[2] Occasionally Rott claims in addition that his formal results should be interpreted as a formal reduction of theoretical rationality to practical rationality or, less ambitiously, as a way of utilizing the theory of choice functions as a more primitive (and secure) theory to which the theory of belief revision has been reduced (e.g., see Rott (2001, pp. 5–6, 142, 214)). It is unclear whether these claims hold independently of the formal results presented by Rott. Erik Olsson (2003) offers criticisms along these lines. Isaac Levi (2004b) argues that the reduction is not a reduction to a theory of choice *per se*. In addition, Levi (2004b) offers an analysis of belief change where the act of changing view is constructed as an epistemic decision. The main technical tools Levi uses are not taken from the theory of choice functions but from other areas of decision theory.

We will appeal to some of the techniques Rott (2001) uses, but we do not claim that we are offering a reduction of belief change to the theory of choice or a reduction of theoretical rationality to practical rationality. As the reader will see, nevertheless, the mathematical techniques Rott (2001) exploits have a heuristic value to discover interesting postulates regulating belief change when social norms are relevant.

[3] There are many possible responses to Sen's examples. One option could be to redefine the space of options in such a way as to tag y as the option of 'taking the last apple from the plate' (see Levi 2004a, for an analysis along these lines). Sen himself seems ambivalent regarding the analysis of his own examples. On some occasions he seems to think that redefinitions of this sort are feasible, yet on other occasions he has argued that these type of redefinitions make the principles of rational

The above example, where social norms influence choice, is an illustration of a phenomenon that Sen calls *menu dependence*. Roughly speaking, menu dependence arises in rational choice when the *evaluation* of alternatives for choice or the *mode of selection* guiding choice varies parametrically with what collection of alternatives is available for choice. For instance, menu dependence may arise when a set of alternatives—often called a *menu* in rational choice—from which an agent is to choose carries information directly relevant to what the agent has reasons to choose. In particular, an agent faced with choosing among some set of alternatives may learn something about the underlying situation, thereby influencing the agent's decision over the alternatives available for choice.

To take an example, a chooser may learn something about a person offering a choice on the basis of what the person is offering: Given the choice between having a beer after work at a distant acquaintance's home (b) and not spending time with the acquaintance (s), a person who chooses to have a beer instead of going home may nevertheless choose to not spend time with the acquaintance if the acquaintance instead offers a choice among having a beer (b), not spending time with the acquaintance (s), and smoking some crack (c). The appearance of the third alternative has altered the person's evaluation of the other two alternatives. In particular, the person has *learned* something about the acquaintance when offered the third alternative, and so the expanded menu has triggered additional inferences about the acquaintance; accordingly, the person chooses to not spend time with the acquaintance. In this example, the set of alternatives $\{b, s, c\}$ has epistemic relevance for the person's decision, and the person's choices are *menu dependent*. The foregoing example is also an illustration of a violation of condition α, since s was rejected for choice from $\{b, s\}$ yet was chosen from $\{b, s, c\}$.

Examples in which social norms influence choice (like the first example involving the fruit basket) can also be seen to constitute cases of menu dependence of choice. In the case of the fruit basket, the menu $\{x, y\}$ occasions a particular mode of choice, i.e., a particular means by which a choice is made. In this case there is a concrete mechanism that explains why this is so. Option y is rendered unfeasible for choice from the menu $\{x, y\}$ in virtue of an operative social norm, whereupon the person chooses x according to her preferences. When z is added to the menu, the norm is no longer applicable, whereby y becomes available for choice for the person and she chooses y according to her preferences. We find especially interesting and more tractable cases in which social norms place constraints upon maximizing according to fixed preferences. We will accordingly focus on cases of this sort in this article.

In light of Rott's correspondence results, it is natural to inquire whether violations of rationality postulates of belief change and non-monotonic reasoning can be understood as a result of menu dependence and in particular operative social norms.

choice empty. In general, maneuvers of this kind tend to be blind with respect to the role of social norms in reasoning. We prefer here to take norms at face value as Sen does in many of his writings.

In other words, it is natural to ask whether menu dependence, which can wreak havoc on almost all coherence constraints on selection functions in rational choice, can—in epistemic form—also undercut rationality postulates of belief change and non-monotonic reasoning. We will see that some of the existing counterexamples against well-known principles of belief change (of the sort proposed by Rott (2004)) are indeed explicable in terms of the phenomenon of menu dependence. In addition, we present counterexamples for which menu dependence is explainable in terms of social norms.

The methodological difficulty that the possibility of menu dependence and in particular social norms pose for belief change and non-monotonic reasoning is by no means trivial. Indeed, as we will see, the possibility of menu dependence threatens an all but universal presumption in the literature of belief change and non-monotonic reasoning—that selection functions are *relational*.[4] Roughly, what this presumption amounts to is the requirement that there is some underlying binary relation—whether over all sentences, sets of sentences, worlds, sets of worlds, models, or sets of models, and so on—according to which a selection function picks the 'best' or 'unsurpassed' elements from its arguments. As discussed by Sen (1997), social norms (and menu dependence in general) threatens essentially the same presumption in the context of rational choice—that selections function are *rationalizable*. This presumption requires that there is some underlying binary preference relation over all alternatives for choice according to which a selection function picks *optimal* or *maximal* elements from its arguments. In fact, relationality in belief change and non-monotonic reasoning is formally equivalent to rationalizability in rational choice.[5]

Some of the recent literature Bossert and Suzumura (2007) has argued that it is possible to accommodate social norms in the theory of rational choice by adopting suitably modified axioms of revealed preference. In this article we offer an argument along these lines that improves on the existing arguments presented in Bossert and Suzumura (2007). We will use this result to provide a solution to some of the problems previously diagnosed in the theory of belief change, offering a novel axiomatization of belief change that, we claim, helps to resolve problems recently raised against standard axiomatizations of belief revision *à la* Alchourrón, Gärdenfors, and Makinson (AGM) (1985). Surprisingly perhaps (taking into account that we arrive at our conditions by way of a completely different route), the central condition deployed in our improved argument is a variant of a condition entertained by AGM in 1985. So in a way the main new axioms we propose have precursors in the literature.

[4] A non-exhaustive list of influential articles in belief change and non-monotonic reasoning which presume relationality: Alchourrón et al. (1985), Alchourrón and Makinson (1985), Rott (1993), Rott and Pagnucco (2000), Hansson (1999), Arló-Costa (2006), Makinson (1989), Kraus et al. (1990), Lehmann and Magidor (1992), Gärdenfors and Makinson (1994).

[5] In fact, the influence of social norms and more generally menu dependence even threatens the presumption that selection functions are *pseudo-rationalizable* (see Moulin, 1985, for a discussion of this notion).

Our article intends to show that the debate regarding the foundations of the theory of rational choice is directly relevant to the understanding of apparent counterexamples against principles of belief formation used in the theory of belief change (for example, the counterexamples presented in Rott (2004)). At the same time we intend to contribute to the contemporary debate in the foundations of the social sciences itself by showing the robustness of the program of rationalizability both in rational choice and in belief change. Thus, we intend to offer a common solution to problems that many have identified as fatal for the foundational program defended by Rott (see, for example, Olsson, 2003) and for the rationalizability program in the foundations of rational choice.

The article is divided into two halves, each of which has two parts. Part I is structured as follows. In Section 8.2.1, we present the technical machinery used both in the half devoted to rational choice and in the half devoted to belief revision. Section 8.2.2 offers an introduction to the theory of choice functions in rational choice. We review the traditional notion of rationalizability in terms of optimization and several coherence constraints on choices across varying menus that have been important in the literature. Then we propose a condition in terms of which it is possible to formulate a functional characterization of rationalizability with respect to *general domains*. We will see that this condition is a natural generalization of a condition first entertained in the seminal paper by AGM (1985) on belief change.

In Part II, we first focus in Section 8.3.1 on attempts to extend the notion of rationalizability to cope with social norms. We discuss the work of Walter Bossert and Kotaro Suzumura (2007) and present their extension of the theory of rationalizability. Section 8.3.2 presents an alternative extension of rationalizability capable of accommodating social norms that we call *norm-conditional choice*. We compare our extension with Bossert and Suzumura's extension. We argue that our proposal offers some advantages that will play a crucial role in the sections on epistemological applications that follow.

Part III initiates the second half of the article. In Section 8.4.1 we present basic background on the literature of belief revision. Then in Section 8.4.2 the use of selection functions in belief revision is connected formally with the use of choice functions in rational choice. Section 8.4.3 is a self-contained review of some of the central results offered by Rott in Rott's 2001 linking postulates of belief revision to coherence constraints in the theory of rational choice.

Section 8.4.4 presents various counterexamples to principles of belief formation necessary for rationalizability, such as postulate ($*7$). The initial counterexample is due to Rott (2004). Then we present new examples which illustrate the role of social norms in belief revision.

In Part IV, Section 8.5.1 presents a new theory of belief revision called *norm-inclusive belief revision*. Like the theory of norm-conditional choice we discuss, norm-inclusive belief revision is intended to take into account the role social norms play in belief change. In Section 8.6 we discuss our theory and illustrate how our theory works. Section 8.7 closes the article with a conclusion and a discussion of future work.

8.2 Part I: New Foundations for Rational Choice

8.2.1 Technical Preliminaries

In the following we presuppose a propositional language $\mathcal{L}$ with the connectives $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$. We let For($\mathcal{L}$) denote the set of formulae of $\mathcal{L}$, $a, b, c, \ldots p, q, r, \ldots$ denote propositional variables of $\mathcal{L}$, and $\alpha, \beta, \delta, \ldots, \varphi, \psi, \chi, \ldots$ denote arbitrary formulae of $\mathcal{L}$. Sometimes we assume that the underlying language $\mathcal{L}$ is *finite*. By this we mean that $\mathcal{L}$ has only finitely many propositional variables.

As is customary, we assume that $\mathcal{L}$ is governed by a consequence operation $\text{Cn} : \mathcal{P}(\text{For}(\mathcal{L})) \rightarrow \mathcal{P}(\text{For}(\mathcal{L}))$ such that for all $A, B \subseteq \text{For}(\mathcal{L})$,

(i) $A \subseteq \text{Cn}(A)$.
(ii) If $A \subseteq B$, then $\text{Cn}(A) \subseteq \text{Cn}(B)$.
(iii) $\text{Cn}(\text{Cn}(A)) \subseteq \text{Cn}(A)$.
(iv) $\text{Cn}_0(A) \subseteq \text{Cn}(A)$, where Cn_0 is classical tautological implication.
(v) If $\varphi \in \text{Cn}(A)$, then there is some finite $A_0 \subseteq A$ such that $\varphi \in \text{Cn}(A_0)$.
(vi) If $\varphi \in \text{Cn}(A \cup \{\psi\})$, then $\psi \rightarrow \varphi \in \text{Cn}(A)$.

Conditions (i)–(vi) are respectively called *Inclusion, Monotony, Idempotence, Supraclassicality, Compactness*, and *Deduction* (Hansson, 1999, p. 26). As usual, A is called *logically closed* if $\text{Cn}(A) = A$, and $A \vdash \varphi$ is an abbreviation for $\varphi \in \text{Cn}(A)$. We let $\mathbb{K}$ denote the collection of logically closed sets in $\mathcal{L}$.

We let $\mathcal{W}_{\mathcal{L}}$ denote the collection of all maximal consistent sets of $\mathcal{L}$ with respect to Cn. Members of $\mathcal{W}_{\mathcal{L}}$ are often called *states*, *possible worlds* or just *worlds*. For a non-empty collection of worlds W of $\mathcal{W}_{\mathcal{L}}$, let $\widehat{W}$ denote the set of sentences of $\mathcal{L}$ which are members of all worlds in W (briefly, $\widehat{W} := \bigcap_{w \in W} w$). If A is a set of sentences of $\mathcal{L}$, we let $[\![A]\!] := \{w \in \mathcal{W}_{\mathcal{L}} : A \subseteq w\}$. If φ is a sentence of $\mathcal{L}$, we write $[\![\varphi]\!]$ instead of $[\![\{\varphi\}]\!]$. Observe that for every set of sentences A of $\mathcal{L}$, $\text{Cn}(A) = \widehat{[\![A]\!]}$. A member of $\mathcal{P}(\mathcal{W}_{\mathcal{L}})$ is often called a proposition, and $[\![\varphi]\!]$ is often called the *proposition expressed by* φ. Intuitively, $[\![A]\!]$ consists of those worlds in which all sentences in A hold. Finally, let $\mathcal{E}_{\mathcal{L}}$ be the set of all elementary subsets of $\mathcal{W}_{\mathcal{L}}$, i.e., $\mathcal{E}_{\mathcal{L}} := \{W \in \mathcal{P}(\mathcal{W}_{\mathcal{L}}) : W = [\![\varphi]\!] \text{ for some } \varphi \in \mathit{For}\ (\mathcal{L})\}$.

We now briefly turn to selection functions. For a non-empty set X and a non-empty collection $\mathcal{S}$ of subsets of X, we call the pair $(X, \mathcal{S})$ a *choice space*. A *selection function* (or *choice function*) on a choice space $(X, \mathcal{S})$ is a function $\gamma : \mathcal{S} \rightarrow \mathcal{P}(X)$ such that $\gamma(S) \subseteq S$ for every $S \in \mathcal{S}$.[6] Intuitively, a selection function $\gamma : \mathcal{S} \rightarrow \mathcal{P}(X)$ chooses the 'best' elements of each S in $\mathcal{S}$.

Now let $(X, \mathcal{S})$ be a choice space. We say that $\mathcal{S}$ is *closed under finite unions* if for every $n < \omega$, if $\{S_i : i < n\} \subseteq \mathcal{S}$, then $\bigcup_{i<n} S_i \in \mathcal{S}$; we say that $\mathcal{S}$ is *closed under relative complements* if whenever $S, T \in \mathcal{S}$, $S \backslash T \in \mathcal{S}$; and we call $\mathcal{S}$ *compact*

[6] Following Rott (2001), we do not require that $\mathcal{S}$ or $\gamma(\mathcal{S})$ consists solely of nonempty sets. This approach allows for more generality.

if for every $S \in \mathcal{S}$ and $I \subseteq \mathcal{S}$, if $S \subseteq \bigcup_{T \in I} T$, then there is some finite $I_0 \subseteq I$ such that $S \subseteq \bigcup_{T \in I_0} T$. We also say that $(X, \mathcal{S})$ is *closed under finite unions* (*closed under relative complements, compact*) if $\mathcal{S}$ is closed under finite unions (closed under relative complements, compact).

8.2.2 Choice Functions in Rational Choice

The idea that a selection function $\gamma : \mathcal{S} \to \mathcal{P}(X)$ takes the 'best' elements of each set in $\mathcal{S}$ has been made more precise by assuming that there is some ordering over the elements of X according to which $\gamma(S)$ distinguishes the best elements of each $S \in \mathcal{S}$. Two formalizations of this idea have been widely utilized in the literature. The first formalization is based on a non-strict (reflexive) relation $\geq$ on X. It demands that for all $S \in \mathcal{S}$,

$$(\mathrm{Eq}_{\geq}) \qquad \gamma(S) = \{x \in S : x \geq y \text{ for all } y \in S\}.$$

This formalization is called *optimization* in Sen (1997 p. 763) and *G-rationality* in Suzumura (1983 p. 21). The second formalization of the idea of picking the 'best' elements is based on a strict (asymmetric) relation $>$ on X. It demands that for all $S \in \mathcal{S}$,

$$(\mathrm{Eq}_{>}) \qquad \gamma(S) = \{x \in S : y > x \text{ for no } y \in S\}.$$

This is called *maximization* in Sen (1997, p. 763) and *M-rationality* in Suzumura (1983, p. 21). This formalization captures a somewhat weaker notion than that of picking the 'best' elements of each $S \in \mathcal{S}$. Indeed, it would be more accurate to say $\mathrm{Eq}_{>}$ is a formalization of the idea of picking elements which are 'no worse' than any other elements. In this article, much of our discussion concerning rational choice will be framed with respect to optimization.

In rational choice, the ideas represented in $\mathrm{Eq}_{\geq}$ and $\mathrm{Eq}_{>}$ have been quite generally exploited, the methodology of rational choice characterized quite well by the slogan, 'Rational choice is relational choice' (Rott, 1993, p. 1429). There the elements of X from a choice space $(X, \mathcal{S})$ represent *alternatives of choice* under the potential control of an agent, the members of $\mathcal{S}$ represent possible *decision problems* for the agent, and a selection function γ on $(X, \mathcal{S})$ represents the *values* of the agent. Then for an agent confronted with a decision problem S (also often called a *menu*), the set $\gamma(S)$ consists of those elements that the agent regards as *equally adequate or satisfactory* alternatives from S (Herzberger, 1973, p. 189).[7] Accordingly,

[7] Herzberger also writes, 'Under the natural interpretation of $\gamma(S)$ as a set of solutions to the problem S, the value $\gamma(S) = \emptyset$ earmarks a decision problem that is unsolvable by the function γ; and so the domain $\mathcal{S}$ bears interpretation as the class of all decision problems that are solvable under the given choice function' (Herzberger 1973, p. 189, notation adapted).

choice is often construed as that which is based upon an underlying preference relation, where $\gamma(S)$, often called a *choice set*, represents those alternatives an agent takes to be ‘best’ or ‘no worse’ with respect to his underlying preferences.

Before we continue, let us recapitulate the forgoing discussion with a few defintions.

Definition 2.1 Let γ be a selection function on a choice space $(X, \mathcal{S})$. We say that a binary relation R on X *rationalizes* γ if for every $S \in \mathcal{S}$,

$$\gamma(S) = \{x \in S : xRy \text{ for all } y \in S\}.$$

We thereby call γ *rational* (or *rationalizable*) if there is a binary relation R on X that rationalizes γ.[8]

Definition 2.2 Let γ be a selection function on a choice space $(X, \mathcal{S})$.

(i) We say that γ is *G-rational* (or *G-rationalizable*) if there is a reflexive binary relation $\geq$ on X that rationalizes γ.

(ii) We say that γ is *M-rational* (or *M-rationalizable*) if there is an asymmetric binary relation $>$ on X such that $((X \times X) \backslash >)^{-1}$ rationalizes γ.[9]

Clearly a selection function γ on a choice space $(X, \mathcal{S})$ is G-rational just in case there is a reflexive binary relation $\geq$ on X such that γ satisfies $\mathrm{Eq}_{\geq}$ for every $S \in \mathcal{S}$. Moreover, a selection function γ on $(X, \mathcal{S})$ is M-rational if and only if there is a asymmetric binary relation $>$ on X such that γ satisfies $\mathrm{Eq}_{>}$ for every $S \in \mathcal{S}$. Thus, G-rational and M-rational selection functions correspond to the notions discussed above. In addition, every M-rational selection function is G-rational, but the converse does not in general hold.[10] It is in this way that maximization is a weaker notion than optimization.

[8] There are alternative notions of rationalizablity one might find appealing. For example, we might instead say that a binary relation R on X *rationalizes* γ if for every $S \in \mathcal{S}$ such that $\gamma(S) \neq \emptyset$, $\gamma(S) = \{x \in S : xRy \text{ for all } y \in S\}$.

[9] For a binary relation R on X, $R^{-1} := \{(x, y) \in X \times X : (y, x) \in R\}$.

[10] However, every selection function rationalized by a complete binary relation is M-rational. (A binary relation R on X is *complete* if for every $x, y \in X$, either xRy or yRx.) To see why not every G-rational selection function is M-rational, we present an example from Suzumura (1976, pp. 151–152).

Consider a selection function γ on a choice space $(X, \mathcal{S})$, where $X := \{x, y, z\}$, $\mathcal{S} := \{\{x, y\}, \{x, z\}, X\}$, $\gamma(\{x, y\}) := \{x, y\}, \gamma(\{x, z\}) := \{x, z\}$, and $\gamma(\{x, y, z\}) := \{x\}$. Then γ is rationalized by a reflexive binary relation $\geq$ defined by $\geq := \{(x, x), (y, y), (z, z), (x, y), (x, z), (y, x), (z, x)\}$. However, γ is not M-rationalizable, for otherwise, if $>$ is an asymmetric binary relation for which $(X \times X)^{-1}$ rationalizes γ, then since $\gamma(\{x, y\}) = \{x, y\}$ and $\gamma(\{x, z\}) = \{x, z\}$, it follows that $x \not> y$ and $x \not> z$, but since $\gamma(\{x, y, z\}) = \{x\}$, it follows that $y > z$ and $z > y$, contradicting that $>$ is asymmetric.

(α) For every $S, T \in \mathcal{S}$, if $S \subseteq T$, then $S \cap \gamma(T) \subseteq \gamma(S)$. (*Sen's Property* α)

(β^+) For every $S, T \in \mathcal{S}$, if $S \subseteq T$ and $S \cap \gamma(T) \neq \emptyset$, then $\gamma(S) \subseteq \gamma(T)$. (*Sen's Property* β^+)

(γ) For every nonempty $I \subseteq \mathcal{S}$ such that $\bigcup_{S \in I} S \in \mathcal{S}$, $\bigcap_{S \in I} \gamma(S) \subseteq \gamma(\bigcup_{S \in I} S)$. (*Sen's Property* γ)

(Aiz) For every $S, T \in \mathcal{S}$, if $S \subseteq T$ and $\gamma(T) \subseteq S$, then $\gamma(S) \subseteq \gamma(T)$. (*Aizerman's Axiom*)

In the study of rational choice, *coherence constraints* have been imposed on the form relationships may take among choices across varying menus. In other words, these requirements specify how choices must be made across different decision problems. Some predominant coherence constraints are the following:[11]

We mentioned condition α in the introduction of this article. Recall that this condition demands that whatever is rejected for choice from a menu must remain rejected if the menu is expanded. More formally, this means that for any menu S, if x is an alternative in S and x is not in $\gamma(S)$—that is, x is not chosen, i.e., is rejected, from S—then if S is expanded to a menu S'—that is, if S' is such that S is a subset of S'—then x is not in $\gamma(S')$. Equivalently, this condition demands that whatever is chosen from a menu must also be chosen from any smaller menu for which this choice is still available.[12] Condition α entails the following coherence constraint:

(α^*) For every $S, T \in \mathcal{S}$ such that $S \cup T \in \mathcal{S}$, $\gamma(S \cup T) \subseteq \gamma(S) \cup \gamma(T)$. (*Sen's Property* α^*)

Furthermore, if $\mathcal{S}$ is closed under finite unions and relative complements, then condition α is equivalent to condition α^*.

Similarly, condition γ entails the following coherence constraint:

(γ^*) For every $S, T \in \mathcal{S}$ such that $S \cup T \in \mathcal{S}$, $\gamma(S) \cap \gamma(T) \subseteq \gamma(S \cup T)$. (*Sen's Property* γ^*)

If $\mathcal{S}$ is closed under finite unions and is compact, then condition γ is equivalent to condition γ^*.

We can combine α^* and γ^* into one condition:

(γR) For every $S, T \in \mathcal{S}$ such that $S \cup T \in \mathcal{S}$,
$\gamma(S) \cap \gamma(T) \subseteq \gamma(S \cup T) \subseteq \gamma(S) \cup \gamma(T)$.

Now observe that if $\mathcal{S}$ is closed under finite unions and is compact, the following coherence constraints are pairwise equivalent:

[11] Actually, Sen's Property β is more pervasive than Sen's Property β^+ (Sen, 1977, p. 66). Condition β demands that if $S \subseteq S'$ and $\gamma(S') \cap \gamma(S) \neq \emptyset$, then $\gamma(S) \subseteq \gamma(S')$ (Sen 1971, p. 313). Condition β^+ entails condition β, and in the presence of condition α, condition β and condition β^+ are logically equivalent.

[12] Condition α, also known as Chernoff's Axiom, should not be confused with another important condition, the so-called *Independence of Irrelevant Alternatives* (Arrow, 1951, p. 27). See Sen (1977 pp. 78–80) for a vivid discussion of the difference between these two conditions. See also Ray (1973) for another clear discussion of this sort.

($\gamma\mathrm{R}_\infty$) For every nonempty $I \subseteq \mathcal{S}$ and $S \in \mathcal{S}$, if $S \subseteq \bigcup_{T\in I} T$, then $S \cap (\bigcap_{T\in I} \gamma(T)) \subseteq \gamma(S)$.

($\gamma\mathrm{R}_{<\omega}$) For every $n < \omega$, if $T_i \in \mathcal{S}$ for each $i < n$ and $S \in \mathcal{S}$, then if $S \subseteq \bigcup_{i<n} T_i$, then $S \cap (\bigcap_{i<n} \gamma(T)) \subseteq \gamma(S)$.

($\gamma\mathrm{R}_1$) For every $S, T_0, T_1 \in \mathcal{S}$, if $S \subseteq T_0 \cup T_1$, then $S \cap \gamma(T_0) \cap \gamma(T_1) \subseteq \gamma(S)$.

If in addition $\mathcal{S}$ is closed under relative complements, then condition $\gamma\mathrm{R}$ is equivalent to each of the aforementioned coherence constraints. We record this fact in a proposition, leaving its proof to the reader.

Proposition 2.3 *Let γ be a selection function on a choice space $(X, \mathcal{S})$ that is closed under finite unions and is compact. Then the following are equivalent:*

(i) *γ satisfies γR_∞.*
(ii) *γ satisfies $\gamma R_{<\omega}$.*
(iii) *γ satisfies γR_1.*
(iv) *γ satisfies α and γ.*
(v) *γ satisfies α and γ^*.*

If in addition $(X, \mathcal{S})$ is closed under relative complements, then the following are pairwise equivalent and equivalent to γR_∞:

(vi) *γ satisfies γR.*
(vii) *γ satisfies $\mathrm{\alpha}^*$ and γ.*
(viii) *γ satisfies $\mathrm{\alpha}^*$ and γ^*.*

We invite the reader to consider the implications among the above coherence constraints (in particular, observe that condition $\gamma\mathrm{R}_\infty$ entails conditions α and γ).

It is well-known that conditions α, α^*, γ, and γ^* are each necessary for a selection function to be rationalizable. Indeed, it can be shown that conditions $\gamma\mathrm{R}$, $\gamma\mathrm{R}_\infty$, $\gamma\mathrm{R}_{<\omega}$, and $\gamma\mathrm{R}_1$ are also each necessary for a selection function to be rationalizable.

Proposition 2.4 *A rational selection function γ satisfies α, α^*, γ, γ^*, $\gamma\mathrm{R}$, $\gamma\mathrm{R}_\infty$, $\gamma\mathrm{R}_{<\omega}$, and $\gamma\mathrm{R}_1$.*

Yet only select subsets of these conditions are sufficient for a selection function to be rationalizable. To be sure, conditions α and γ are jointly sufficient for rationalizability, as are $\mathrm{\alpha}^*$ and γ, α and γ^*, and $\mathrm{\alpha}^*$ and γ^*. Furthermore, conditions $\gamma\mathrm{R}$, $\gamma\mathrm{R}_{<\omega}$, $\gamma\mathrm{R}_1$ are each sufficient to rationalize a choice function. However, only under certain constraints on domains of selection functions are any of the aforementioned conditions sufficient. Thus, in the spirit of generality, we offer a general characterization of rationalizability which does not depend on any domain restrictions.

Theorem 2.5 *A selection function γ is rational if and only if it satisfies condition $\gamma\mathrm{R}_\infty$.*[13]

For the sake of brevity, we omit a proof this theorem, referring the reader to Theorem 3.6 to see how to assemble a proof of the above theorem. The reader may consult the proof in Richter (1971, Theorem 2, p. 33) to gather some of the elements required for a proof of the above theorem.[14]

In, as in most literature of kin, selection functions are assumed to exclude the empty set from their domains and are moreover assumed to satisfy the following condition:

$$(\gamma_{>\emptyset}) \qquad \text{For every } S \in \mathcal{S}, \text{ if } S \neq \emptyset, \text{ then } \gamma(S) \neq \emptyset. \qquad (\textit{Regularity})$$

Rott (2001, p. 150) calls this condition *success*. We will call a selection function that satisfies condition $\gamma_{>\emptyset}$ *regular*. In the presence of condition $\gamma_{>\emptyset}$, a selection function satisfying condition $\gamma\mathrm{R}_\infty$ is G-rational, i.e., rationalized by a reflexive binary relation. But one need not presuppose condition $\gamma_{>\emptyset}$ in a theorem. Here we offer a condition which in conjunction with condition $\gamma\mathrm{R}_\infty$ is sufficient for G-rationality:

$$(\gamma_{1>\emptyset}) \qquad \text{For every } x \in X \text{ such that } \{x\} \in \mathcal{S},\ \gamma(\{x\}) \neq \emptyset. \qquad (\textit{Singleton Regularity})$$

We thereby have the following theorem, which we offer without proof.[15]

Theorem 2.6 *A selection function γ is G-rational if and only if it satisfies condition $\gamma\mathrm{R}_\infty$ and condition $\gamma_{1>\emptyset}$.*

We immediately have the following corollary:

Corollary 2.7 *Let γ be a selection function satisfying condition $\gamma_{>\emptyset}$. Then γ is G-rationalizable if and only if it satisfies condition $\gamma\mathrm{R}_\infty$.*

[13] Recall the notion of rationalizability briefly discussed in footnote 8. It can be shown that a selection function is rationalizable in the sense of footnote 8 just in case it satisfies the following condition: For every nonempty $I \subseteq \mathcal{S}$ and $S \in \mathcal{S}$, if $S \subseteq \bigcup_{T \in I} T$ and $\gamma(S) \neq \emptyset$, then $S \cap (\bigcap_{T \in I} \gamma(T)) \subseteq \gamma(S)$. This illustrates how one can modify coherence constraints for other notions of rationalizablity.

[14] Theorem 3.6 is stated within a more general framework that we introduce in Section 8.3. A direct proof of Theorem 2.5 proceeds in a way unlike the proof in Richter (1971). This is primarily because the results in Richter (1971) concern what is called the *V-Axiom*. Condition $\gamma\mathrm{R}_\infty$ is not discussed in Richter (1971).

[15] The minimal conditions needed for a proof can be gathered from the proof in Richter (1971, Theorem 3, p. 34). Although the proof in Richter (1971) does not itself establish Theorem 2.6, a careful inspection of the proof in Richter (1971) should make it clear that some assumptions of the theorem associated with this proof can be weakened. Indeed, condition $\gamma_{1>\emptyset}$ is not discussed in Richter (1971) or to our knowledge anywhere else in the literature on choice functions. As we have indicated, selection functions are assumed to be regular in Richter (1971). See footnote 14.

The astute reader will have observed that conditions γR_{∞}, $\gamma R_{<\omega}$, and γR_1 bear a striking resemblance to conditions $\gamma 7{:}\infty$, $\gamma 7{:}N$, and $\gamma 7{:}2$ of Alchourrón et al. (1985). Indeed, γR_{∞}, $\gamma R_{<\omega}$, and γR_1 are generalized forms of these conditions. As Rott (1993, p. 1432) points out, Alchourrón, Gärdenfors, and Makinson realized that conditions $\gamma 7{:}2$ and $\gamma 7{:}\infty$ are each alone sufficient for rationalizability in the context of their theory of belief change (1985, pp. 521–522, 529–530) (in the context of the AGM theory of contraction, the domain of each selection function is closed under finite unions, closed under relative complements, and compact).[16] They posed as an open question whether condition $\gamma 7{:}2$ (which is a special case of our γR_1) can be expressed as a rationality postulate of belief contraction. Rott (1993) provides an affirmative answer to this question, showing first that $\gamma 7{:}2$ can be decomposed into conditions α^* and γ^*, whereupon he demonstrates that condition α^* corresponds to rationality postulate ($\dot{-}7$), while under certain assumptions condition γ^* corresponds to rationality postulate ($\dot{-}8$). [17]

In the next section, we will discuss recent work in rational choice aimed at accommodating the possibility of menu dependence in the standard theory of choice. We will focus on a theory of choice that takes into account the influence of social norms in choice. Thereafter we will adopt some of the central ideas of this theory to develop a theory that we take to be an improvement upon the old one.

8.3 Part II: Social Norms and Rational Choice

8.3.1 Norm-Conditional Rationalizability

Amartya Sen (1993, 1996, 1997) considers one aspect of the general phenomenon of menu dependence—namely, the problems external social norms pose for the theory of choice. Sen's examples seem to show that the standard rationalizability approach to the theory of choice—as exercised by Arrow (1959), Hansson (1968), Richter (1966, 1971), Sen (1971), Suzumura (1976), as well as many others—has serious difficulties dealing with social norms. The example we offered in the introduction of this article according to which a person must choose among fruits from a basket and contrary to her preferences does not choose the last apple from the basket is clearly an illustration in which a social norm of politeness is in operation.

[16] Interestingly, it seems that neither Rott nor AGM noticed that condition γR_{∞} is necessary and sufficient for rationalizability over *general* domains (i.e., domains for which no restrictions are imposed, such as closure under finite unions or compactness). Even more interesting is that to our knowledge, condition γR_{∞} has not appeared anywhere in the literature on choice functions. In particular, it appears that no one has explicitly pointed to a connection between condition γR_{∞} and rationalizability.

[17] Postulates ($\dot{-}7$) and ($\dot{-}8r$) are supplementary postulates of belief contraction (see [AGM85] and [Rott93]). For a fixed belief set K and contraction function $\dot{-}$, postulate ($\dot{-}7$) demands that $K\dot{-}\varphi \cap K\dot{-}\psi \subseteq K\dot{-}(\varphi \wedge \psi)$, while postulate ($\dot{-}8r$) requires that $K\dot{-}(\varphi \wedge \psi) \subseteq \mathrm{Cn}(K\dot{-}\varphi \cup K\dot{-}\psi)$.

Sen (1997) considers various ways in which norms can be explicitly represented formally but ultimately does not manage to accommodate them systematically within the standard framework of the theory of rational choice. Nevertheless, recent work by Walter Bossert and Kotaro Suzumura (2007) attempts to resolve the difficulties Sen has brought to the forefront. In this section we will present the main ideas behind Bossert and Suzumura's theory. We will also offer some criticisms. Nonetheless, we will later adopt some of the basic ideas upon which Bossert and Suzumura's framework is based.

Bossert and Suzumura (2007) seek to demonstrate that the rationalizability approach to the theory of choice remains robust in spite of the difficulties Sen raises concerning the influence of social norms. Introducing a notion of rationalizability they call *norm-conditional rationalizability*, Bossert and Suzumura extend the standard framework to cover some areas where it cannot be straightforwardly applied. We are sympathetic with this general idea.

In essence, Bossert and Suzumura develop an idea Sen (1997) briefly explores to incorporate into standard rationalizability theory the influence of social norms in choice. According to Sen, some options from a menu of alternatives are excluded from permissible conduct through what he calls *self-imposed choice constraints* (Sen 1997, p. 769). A person faced with a menu S may first exclude some alternatives from S by taking a *permissible* subset $K(S)$, which represents the person's self-imposed choice constraints, thereupon taking those alternatives from $K(S)$ which are 'best' or 'no worse' than the other alternatives from $K(S)$. Thus, Sen's self-styled *permissibility function* K is such that for each menu S, $K(S)$ identifies a permissible subset of S. Bossert and Suzumura (2007, p. 10) attempt to develop a formal framework to "bridge the idea of norm-induced constraints and the theory of rationalizability," a bridge for which Sen only originated its concept.

Let us for a moment return to the example from the introduction of this article to acquire an understanding of the rudiments of Bossert and Suzumura's approach. In this example, a person is faced with a choice between having the last remaining apple in the fruit basket (y) and having nothing instead (x). Faced with the decision problem represented by $\{x, y\}$, she decides to behave decently and picks nothing (x). Yet if instead she had been confronted with the choice among having nothing (x), having one nice apple (y) and having another nice one (z), she could reasonably enough choose an apple (y), without violating any rule of good behavior. Thus, faced with the decision problem represented by $\{x, y, z\}$, she could reasonably choose y. Now if γ is a selection function that characterizes these choices, we have that $\gamma(\{x, y\}) = \{x\}$ while $\gamma(\{x, y, z\}) = \{y\}$, which is clearly in violation of condition α, thereby precluding rationalizability (see Proposition 2.4).

Bossert and Suzumura's framework explicitly represents the influence of norms in choice. Thus, in the foregoing example, the social norm enjoining that one should not choose the last available apple is represented simply by specifying that the choice of y from $\{x, y\}$ is prohibited, whereas the choice of y (or z) from $\{x, y, z\}$ is permissible. The norm so represented thereby takes into account the consequences of its application. Formally, a social norm is represented as a collection $\mathcal{N}$ of pairs of the form (S,x), where S is a menu and $x \in S$ is prohibited from being chosen from the set S.

Before we continue we should lay down the central components of Bossert and Suzmura's framework. We will do this using our notation and terminology as necessary. According to Bossert and Suzumura, a choice space $(X,\mathcal{S})$ must be such that $\emptyset \notin \mathcal{S}$, and a choice function is a regular selection function $\gamma : \mathcal{S} \to \mathcal{P}(X)$ such that $\gamma(S) \subseteq S \setminus \{x \in S : (S,x) \in \mathcal{N}\}$ for all $S \in \mathcal{S}$. To ensure that the regularity requirement does not conflict with the constraints imposed by the norm $\mathcal{N}$, Bossert and Suzumura stipulate that $\mathcal{N}$ is such that for all $S \in \mathcal{S}$, $\{S\} \times S \nsubseteq \mathcal{N}$, i.e., there exists $x \in S$ such that $(S,x) \notin \mathcal{N}$. Let us call a selection function on a choice space in the sense of Bossert and Suzumura a *normative choice function*.

We may now represent the foregoing example as follows. For $X := \{x,y,z\}$ and $\mathcal{S} := \{\{x,y,z\},\{x,y\}\}$, the operative social norm may be expressed by the set $\mathcal{N} := \{(\{x,y\},y)\}$. Accordingly, since it is required that $\gamma(\{x,y\}) \subseteq \{x,y\} \setminus \{v \in \{x,y\} : (\{x,y\},v) \in \mathcal{N}\}$, the social norm demands that $y \notin \gamma(\{x,y\})$.

As indicated above, the primary goal in Bossert and Suzumura (2007) is to develop a new concept of rationalizability called *norm-conditional rationalizability* in a effort to show that the standard rationalizability approach can accommodate the existence of social norms. Roughly, for a normative choice function γ on a choice space $(X,\mathcal{S})$, norm-conditional rationalizability requires the existence of a preference relation such that for each menu $S \in \mathcal{S}$, $\gamma(S)$ consists of those alternatives which are at least as good as all alternatives from S, except for those alternatives prohibited by the social norm $\mathcal{N}$.

To make this precise, we need some notation and terminology, but the main idea is rather simple and direct. Let $\mathcal{N}$ be a social norm, let $(X,\mathcal{S})$ be a normative choice space, and let $S \in \mathcal{S}$. Bossert and Suzumura define an $\mathcal{N}$*-admissible set* for $(\mathcal{N},S)$, $A^{\mathcal{N}}(S)$, by setting

$$A^{\mathcal{N}}(S) := \{x \in S : (S,x) \notin \mathcal{N}\}.$$

Observe that according to Bossert and Suzumura's framework, for each $S \in \mathcal{S}$, $A^{\mathcal{N}}(S) \neq \emptyset$ and $A^{\mathcal{N}}(S) \subseteq S$. Also observe that $A^{\mathcal{N}}(S)$ is a permissibility function in the sense of Sen.

Bossert and Suzumura then define norm-conditional rationalizability as follows (again adopting our notation and terminology as needed). A normative choice function γ on a choice space $(X,\mathcal{S})$ is $\mathcal{N}$*-rationalizable* if there exists a binary relation $R^{\mathcal{N}}$ on X such that for all $S \in \mathcal{S}$,

$$\gamma(S) := \{x \in A^{\mathcal{N}}(S) : xR^{\mathcal{N}}y \text{ for all } y \in A^{\mathcal{N}}(S)\}.$$

Thus the central idea is simple: for each menu $S \in \mathcal{S}$, $\gamma(S)$ selects the best alternatives from the $\mathcal{N}$-admissible set $A^{\mathcal{N}}(S)$.

In order to facilitate their analysis of norm-conditional rationalizability, Bossert and Suzumura utilize a generalization of the notion of so-called *Samuelson preferences*. Given a social norm $\mathcal{N}$ and a normative choice function γ on a choice space $(X, \mathcal{S})$, Bossert and Suzumura define a binary relation on X, R_γ, called *direct revealed preference*, by setting

$$R_\gamma := \bigcup\nolimits_{S \in \mathcal{S}} (\gamma(S) \times A^{\mathcal{N}}(S)).$$

Bossert and Suzumura then propose a generalization of a coherence constraint due to Richter (1971, p. 33):

($\mathcal{N}$drc) For all $S \in \mathcal{S}$ and $x \in A^{\mathcal{N}}(S)$, ($\mathcal{N}$*-conditional direct-revelation coherence*)
if for every $y \in A^{\mathcal{N}}(S)$ $xR_\gamma y$, then $x \in \gamma(S)$.

In one of the central results of their article, Bossert and Suzumura prove that condition $\mathcal{N}$drc is indeed necessary and sufficient for $\mathcal{N}$*-rationalizability*. The result is direct, and it is rather clear that this property is mathematically required in order to establish the result that Bossert and Suzumura desire. There are, nevertheless, some aspects of this result that we find unsatisfactory.

First, the condition in question, as many coherence constraints of this type used in the theory of rational choice (such as Suzumura's so-called Generalized Condorcet property (1983, p. 32) establishes a constraint on a selection function γ only indirectly by way of a constraint on R_γ. This type of condition not only incorporates a non-primitive notion but also is difficult to use in order to obtain mappings between selection functions and belief revision functions. Second—a somewhat related point—-the condition proposed by Bossert and Suzumura offers little insight into the notion of $\mathcal{N}$-rationalizability in terms of the behavior of selection functions. Thus, we find it better to have a 'pure' constraint on γ in the spirt of coherence constraints such as conditions α and γ.

Nonetheless, we find that Bossert and Suzumura have made a step in the right direction. First, they have developed a modified approach to rationalizability which presupposes no constraints on the domains of choice functions (except that domains cannot include the empty set). Second, because they have presupposed no restrictions on how norms come about, Bossert and Suzumura's approach is very general. Indeed, Bossert and Suzumura's theory of norm-conditional rationalizability is an extension of the classical theory of rationalizability, which is included as a special case.

In the following section, we will offer a refinement of Bossert and Suzumura's theory of norm-conditional rationalizability. We will also offer a coherence constraint that is 'pure.' Ultimately, we will see how to apply our theory to belief change in an effort to accommodate social norms in belief change.

8.3.2 Norm-Conditional Choice Models

In this section, we introduce what we call *norm-conditional choice models*. These models, inspired by both the ideas of Sen and the work of Bossert and Suzumura we discussed in the previous section, are intended to accommodate the role social norms play in choice. We reformulate Bossert and Suzumura's framework to improve upon the formal foundations upon which their theory of norm-condtional rationalizability is built. Among other things, our reformulation squares better with Sen's original conception of what he calls *permissibility functions*. In particular, we take permissibility functions as primitive. We offer several coherence constraints analogous to those discussed in Section 8.2.2 . We also offer a 'pure' coherence constraint, which in the context of our framework, is both necessary and sufficient for norm-conditional rationalizability.

The reader will notice that we have borrowed several central concepts from Bossert and Suzumura's framework. Nonetheless, to avoid confusion, we have adopted a somewhat different notation and terminology, couching these concepts within our framework of norm-conditional choice models.

Definition 3.1 Let γ and π be selection functions on a choice space $(X, \mathcal{S})$. We call the pair (γ, π) a *norm-conditional choice model* on $(X, \mathcal{S})$ if for every $S \in \mathcal{S}$, $\gamma(S) \subseteq \pi(S)$. We call γ a *π-conditional choice function*, and we say that π is a *permissiblity function* for γ.

Thus, in contrast with Bossert and Suzumura's framework, we take permissibility functions as primitive. We also do not prohibit $\emptyset \in \mathcal{S}$, $\gamma(S) = \emptyset$, or $\pi(S) = \emptyset$. We now directly borrow Bossert and Suzumura's notion of norm-conditional rationalizability.

Definition 3.2 Let (γ, π) be norm-conditional choice model on $(X, \mathcal{S})$. We say that a binary relation R on X *π-rationalizes* γ if for every $S \in \mathcal{S}$,

$$\gamma(S) = \{x \in \pi(S) : xRy \text{ for all } y \in \pi(S)\}.$$

We also say that γ is *π-rationalizable* if there is a binary relation R on X that π-rationalizes γ.

As with the standard theory of rationalizability, we can articulate notions of optimization and maximization for norm-conditional choice models.

Definition 3.3 Let (γ, π) be a norm-conditional choice model on $(X, \mathcal{S})$.

(i) We say that γ is *$G\pi$-rational* (or *$G\pi$-rationalizable*) if there is a reflexive binary relation $\geq$ on X that π-rationalizes γ.
(ii) We say that γ is *$M\pi$-rational* (or *$M\pi$-rationalizable*) if there is an asymmetric binary relation $>$ on X such that $((X \times X) \quad >)^{-1}$ π-rationalizes γ.

The following coherence constraints impose conditions upon the interaction of the components of a norm-conditional choice model.

As its name suggests, condition $\gamma[\pi]_{>\emptyset}$ is a hybrid analogue of condition $\gamma_{>\emptyset}$ from Section 8.2.2. A π-conditional choice function γ satisfies condition $\gamma_{>\emptyset}$ just in case the norm-conditional choice model (γ, π) satisfies condition $\gamma[\pi]_{>\emptyset}$ and the

($\gamma[\pi]_{>\emptyset}$) For each $S \in \mathcal{S}$, if $\pi(S) \neq \emptyset$, then $\gamma(S) \neq \emptyset$. (*π-Conditional Regularity*)
($\gamma[\pi]_{1>\emptyset}$) For each $S \in \mathcal{S}$, if $|\pi(S)| = 1$, then $\gamma(S) \neq \emptyset$. (*Singleton π-Conditional Regularity*)
($\gamma[\pi]\mathrm{R}_\infty$) For every nonempty $I \subseteq \mathcal{S}$ and $S \in \mathcal{S}$, (*π-Conditional Coherence*)
if $\pi(S) \subseteq \bigcup_{T\in I} \pi(T)$, then
$\pi(S) \cap (\bigcap_{T\in I} \gamma(T)) \subseteq \gamma(S)$.

permissibility function π for γ satisfies condition $\gamma_{>\emptyset}$. As with condition $\gamma[\pi]_{>\emptyset}$, condition $\gamma[\pi]_{1>\emptyset}$ is a hybrid analogue of condition $\gamma_{1>\emptyset}$ from Section 8.2.2. Clearly condition $\gamma[\pi]_{>\emptyset}$ entails condition $\gamma[\pi]_{1>\emptyset}$. If π is a permissibility function for γ that satisfies condition $\gamma_{1>\emptyset}$, then γ also satisfies condition $\gamma_{1>\emptyset}$ provided the norm-conditional choice model (γ, π) satisfies condition $\gamma[\pi]_{1>\emptyset}$.

In the following, we say that a norm-conditional choice model (γ, π) is *π-conditional regular* if (γ, π) satisfies condition $\gamma[\pi]_{>\emptyset}$. We also call (γ, π) *regular* if γ is regular.

Observe that condition $\gamma[\pi]\mathrm{R}_\infty$ is an analogue of condition $\gamma\mathrm{R}_\infty$ from Section 8.2.2. Among other things, it entails hybrid versions of conditions α and γ.

($\gamma[\pi]\alpha$) For every $S, T \in \mathcal{S}$, if $\pi(S) \subseteq \pi(T)$, then (*Norm-Conditional*
$\pi(S) \cap \gamma(T) \subseteq \gamma(S)$. *Contraction*)
($\gamma[\pi]\gamma$) For every nonempty $I \subseteq \mathcal{S}$ such that $\bigcup_{S\in I} S \in \mathcal{S}$, (*Norm-Conditional Expansion*)
if $\pi(\bigcup_{S\in I} S) = \bigcup_{S\in I} \pi(S)$, then
$\bigcap_{S\in I} \gamma(S) \subseteq \gamma(\bigcup_{S\in I} S)$.

In this context, condition $\gamma[\pi]\alpha$ demands that any permissible alternative rejected for choice from a decision problem must remain rejected in any other menu for which the permissible options from the decision problem are permissible in the other menu.

Condition $\gamma[\pi]\mathrm{R}_\infty$ has a somewhat complicated formulation. The reason for this is that we have not imposed any substantial constraints on the behavior of permissibility functions. Even if we assume that the underlying domain of a norm-conditional choice model satisfies constraints such as closure under finite unions and compactness, condition $\gamma[\pi]\mathrm{R}_\infty$ need not be reducible to simpler coherence constraints as condition $\gamma\mathrm{R}_\infty$ is reducible to conditions α^* and γ^*. Thus, a reduction of this sort would in part depend on conditions imposed on permissibility functions.

We have seen that condition $\gamma\mathrm{R}_\infty$ characterizes rationalizability. In particular, we have seen that this result holds without imposing any conditions on the domains of selection functions. Shortly we will witness a similar result with respect to norm-conditional models.

We consider one more coherence constraint that represents an interesting limit case.

(π_ι) For each $S \in \mathcal{S}$, $S \subseteq \pi(S)$. (*Norm in Absentia*)

A permissibility function satisfying condition π_ι is devoid of any real influence on choice. So as one should expect, rationalizablity and norm-conditional rationalizability collapse.

Proposition 3.4 *Let (γ, π) be a norm-conditional choice model on $(X, \mathcal{S})$. Suppose γ satisfies condition π_ι. Then γ is rational if and only if it is π-rational.*

For our purposes we now borrow Bossert and Suzumura's notion of *direct revealed preference*.

Definition 3.5 Let (γ, π) be a norm-conditional choice model on $(X, \mathcal{S})$. We define a binary relation R_γ on X by setting

$$R_\gamma := \bigcup_{S \in \mathcal{S}} (\gamma(S) \times \pi(S)).$$

We now present the central result of this section.

Theorem 3.6 *Let (γ, π) be a norm-conditional choice model on $(X, \mathcal{S})$. Then γ is π-rational if and only if (γ, π) satisfies condition $\gamma[\pi]R_\infty$.*

Proof In the following, for a binary relation R and $S \in \mathcal{S}$, let

$$G(\pi(S), R) := \{x \in \pi(S) : xRy \text{ for all } y \in \pi(S)\}.$$

($\Rightarrow$) Suppose that γ is π-rationalizable, and let R π-rationalize γ. We show that (γ, π) satisfies $\gamma[\pi]\mathrm{R}_\infty$. Let $I \subseteq \mathcal{S}$ be such that $I \neq \emptyset$, let $S \in \mathcal{S}$, and suppose that $\pi(S) \subseteq \bigcup_{T \in I} \pi(T)$. Assume that $x \in \pi(S) \cap (\bigcap_{T \in I} \gamma(T))$. We show that $x \in G(\pi(S), R)$. We must only establish that for every $y \in \pi(S)$, xRy. So let $y \in \pi(S)$. Then $y \in \bigcup_{T \in I} \pi(T)$, whereby $y \in \pi(T)$ for some $T \in I$. Observe that since $x \in \bigcap_{T \in I} \gamma(T)$, we have that xRz for each $T \in I$ and $z \in \pi(T)$, whence xRy. It follows that $x \in \gamma(S)$.

($\Leftarrow$) Suppose that (γ, π) satisfies condition $\gamma[\pi]\mathrm{R}_\infty$. We show that R_γ π-rationalizes γ. Let $S \in \mathcal{S}$.
$\gamma(S) \subseteq G(\pi(S), R_\gamma)$: If $x \in \gamma(S)$, then $x \in \pi(S)$ and for every $y \in \pi(S)$, by definition $xR_\gamma y$; thus, $x \in G(\pi(S), R_\gamma)$.
$G(\pi(S), R_\gamma) \subseteq \gamma(S)$: Let $x \in G(\pi(S), R_\gamma)$. Then $x \in \pi(S)$ and for every $y \in \pi(S)$, $xR_\gamma y$, so for every $y \in \pi(S)$, there is $T_y \in \mathcal{S}$ such that $x \in \gamma(T_y)$ and $y \in \pi(T_y)$. Observe that since $\pi(S) \subseteq \bigcup_{y \in \pi(S)} \pi(T_y)$, by condition $\gamma[\pi]\mathrm{R}_\infty$ we have that $\pi(S) \cap (\bigcap_{y \in \pi(S)} \gamma(T_y)) \subseteq \gamma(S)$. Hence, since $x \in \pi(S) \cap (\bigcap_{y \in \pi(S)} \gamma(T_y))$, it follows that $x \in \gamma(S)$, as desired.

□

In Section 8.2.2 we showed that condition $\gamma_{1>\emptyset}$ is necessary for G-rationality. We have a similar result for condition $\gamma[\pi]_{1>\emptyset}$.

Theorem 3.7 *Let (γ, π) be a norm-conditional choice model on $(X, \mathcal{S})$. Then γ is $G\pi$-rational if and only if (γ, π) satisfies condition $\gamma[\pi]R_\infty$ and condition $\gamma[\pi]_{1>\emptyset}$.*

Proof

($\Rightarrow$) Suppose that γ is $G\pi$-rationalizable, and let $\geq$ be a reflexive binary relation on X that π-rationalizes γ. By Theorem 3.6, (γ, π) satisfies condition $\gamma[\pi]\mathrm{R}_\infty$. We show that (γ, π) satisfies condition $\gamma[\pi]_{1>\emptyset}$. For *reductio ad absurdum*, assume that there is $S \in \mathcal{S}$ such that $|\pi(S)| = 1$, but $\gamma(S) = \emptyset$.

Then there is $x \in S$ such that $\pi(S) = \{x\}$, but since $\geq$ π-rationalizes γ, it follows that $x \not\geq x$, contradicting that $\geq$ is reflexive.

($\Leftarrow$) Suppose that (γ, π) satisfies condition $\gamma[\pi]\mathrm{R}_\infty$ and condition $\gamma[\pi]_{1>\emptyset}$. By the proof of Theorem 3.6, R_γ π-rationalizes γ. We show that there is a reflexive binary relation on X that π-rationalizes γ. Let Δ be the diagonal of X (i.e., $\Delta := (x,x) : x \in X$). Define a binary relation $\geq$ on X by setting

$$\geq := R_\gamma \cup \Delta$$

Clearly $\geq$ is reflexive. We show that $\geq$ π-rationalizes γ. Let $S \in \mathcal{S}$. In the following, let

$$G(\pi(S), \geq) := \{x \in \pi(S) : x \geq y \text{ for all } y \in \pi(S)\}.$$

$\gamma(S) \subseteq G(\pi(S), \geq)$: Suppose that $x \in \gamma(S)$. Then since R_γ π-rationalizes γ, we have that $x \in \pi(S)$ and for every $y \in \pi(S)$, $xR_\gamma y$, whence for every $y \in \pi(S)$, $x \geq y$. Thus, $x \in G(\pi(S), \geq)$.

$G(\pi(S), \geq) \subseteq \gamma(S)$: Suppose that $x \in G(\pi(S), \geq)$. Then $x \in \pi(S)$ and for every $y \in \pi(S)$, $x \geq y$, so for every $y \in \pi(S)$ for which $x \neq y$, $xR_\gamma y$. Now if $|\pi(S)| = 1$, by condition $\gamma[\pi]_{1>\emptyset}$ we have that $\gamma(S) \neq \emptyset$, whereby since γ is π-rational, it follows that $x \in \gamma(S)$, and we are done. So suppose $|\pi(S)| > 1$. Then there is $y \in \pi(S)$ such that $x \neq y$ and $x \geq y$ and so $xR_\gamma y$, whereupon it follows that for some $T \in \mathcal{S}$, $x \in \gamma(T)$. Therefore, since R_γ π-rationalizes γ, we have that $xR_\gamma x$. Hence, for every $y \in \pi(S)$, $xR_\gamma y$, so since R_γ π-rationalizes γ, we have that $x \in \gamma(S)$, as desired.

□

We thereby have the following corollary.

Corollary 3.8 *Let (γ, π) be a norm-conditional choice model on $(X, \mathcal{S})$. Suppose that γ satisfies condition $\gamma_{>\emptyset}$ or condition $\gamma[\pi]_{>\emptyset}$. Then γ is $G\pi$-rational if and only if (γ, π) satisfies condition $\gamma[\pi]R_\infty$.*

We have seen in this section that condition $\gamma[\pi]\mathrm{R}_\infty$ characterizes norm-conditional rationalizability. This result holds without having presupposed conditions on the domains underlying norm-conditional choice models or on the components of norm-conditional choice models. Our results are therefore comparable to those of Bossert and Suzumura. Yet we have also refined Bossert and Suzumura's framework, and in particular, we have offered, as promised, a coherence constraint that is 'pure' in the spirit of conditions such as α and γ.

This completes our proposal regarding social norms in the realm of rational choice. Now we turn to epistemology and the notion of belief revision, where we will soon see how the foregoing results can be fruitfully applied. In the next section it will be become apparent that conditions $\gamma\mathrm{R}_\infty$, $\gamma\mathrm{R}_{<\omega}$, and $\gamma\mathrm{R}_1$ have natural translations into postulates we label ($*\mathrm{R}_\infty$), ($*\mathrm{R}_{<\omega}$), and ($*\mathrm{R}_1$). Later we will see that just as these conditions have natural translations into postulates of belief revision, condition $\gamma[\pi]\mathrm{R}_\infty$ corresponds to a central postulate of what we will call *norm-inclusive belief revision*.

8.4 Part III: Belief Revision

Belief change has been formalized in several frameworks. In this article, the general framework of belief change under discussion is based on the work of Alchourrón, Gärdenfors, and Makinson (AGM) (1985). We will presume familiarity with the AGM framework, but here we will review some of the basic ideas.[18]

In the AGM framework, an agent's belief state is represented by a logically closed set of sentences K, called a *belief set*. The sentences of K are intended to represent the *beliefs* held by the agent. Belief change then comes in three flavors: *expansion, revision,* and *contraction*.

In expansion, a sentence φ is added to a belief set K to obtain an expanded belief set $K+\varphi$. This expanded belief set $K+\varphi$ might be logically inconsistent. In revision, by contrast, a sentence φ is added to a belief set K to obtain a revised belief set $K*\varphi$ in a way that preserves logical consistency. To ensure that $K*\varphi$ is consistent, some sentences from K might be removed. In contraction, a sentence φ is removed from K to obtain a contracted belief set $K\dot{-}\varphi$ that does not include φ. In this article we will be primarily concerned with *belief revision*.

8.4.1 Postulates for Belief Revision

For a fixed belief set K, the following are the six *basic postulates* of belief revision (Alchourrón et al., 1985, p. 513; Hansson, 1999, p. 212):

$(*1)$	$K*\varphi$ is a belief set.	(*Closure*)
$(*2)$	$\varphi \in K*\varphi$.	(*Success*)
$(*3)$	$K*\varphi \subseteq \mathrm{Cn}(K\cup\{\varphi\})$.	(*Inclusion*)
$(*4)$	If $\neg\varphi \notin K$, then $\mathrm{Cn}(K\cup\{\varphi\}) \subseteq K*\varphi$.	(*Vacuity*)
$(*5)$	If $\mathrm{Cn}(\{\varphi\}) \neq \mathrm{For}(\mathcal{L})$, then $K*\varphi \neq \mathrm{For}(\mathcal{L})$.	(*Consistency*)
$(*6)$	If $\mathrm{Cn}(\{\varphi\}) = \mathrm{Cn}(\{\psi\})$, then $K*\varphi = K*\psi$.	(*Extensionality*)

Let us henceforth call a function $*_K : \mathrm{For}(\mathcal{L}) \to \mathbb{K}$ a *revision function* over K if it satisfies postulates $(*1)$, $(*2)$, and $(*6)$. Of course, we write $K*\varphi$ instead of $*_K\varphi$.

The six basic postulates are elementary requirements of belief revision and taken by themselves are much too permissive. Invariably, several postulates are added to the basic postulates to rein in this permissiveness and to add structure to belief change. Such postulates are called *supplementary postulates*. Among the various postulates added to the mix, the following postulate—or some equivalent or stronger version of it—never fails to be set forth (Gärdenfors, 1979, p. 393):

$$(*7g)\ K*\varphi \cap K*\psi \subseteq K*(\varphi\vee\psi).$$

[18] A comprehensive introduction to theories of belief change is Hansson (1999). A brief introduction to belief change may be found in Gärdenfors (1992).

In Hansson (1999, p. 217), postulate ($*7g$) is called *Disjunctive Overlap*.[19] It encodes the intuitive idea that if an agent believes δ whether it revises its beliefs K by φ or by ψ, then the agent ought to believe δ if the agent revises its beliefs K by $\varphi \vee \psi$. Peter Gärdenfors (1988, pp. 211–212) shows that in the presence of postulates ($*1$)–($*6$), postulate ($*7g$) is equivalent to the following postulate:

$$(*7)\ K * (\varphi \wedge \psi) \subseteq \mathrm{Cn}((K * \varphi) \cup \{\psi\}).$$

In fact, an examination of the proof in Gärdenfors (1988) reveals that this equivalence holds even in the presence of only postulates ($*1$), ($*2$), and ($*6$), i.e., if $*$ is revision function over K.

Often another postulate—or some postulate at least as strong as it—is added to the mix:

$$(*8r)\ K * (\varphi \vee \psi) \subseteq \mathrm{Cn}(K * \varphi \cup K * \psi).$$

This postulate is called *Disjunction* in Gärdenfors and Rott (1995, p. 54).[20]

As with conditions α^* and γ^*, we can combine postulates ($*7g$) and ($*8r$) into one postulate:

$$(*\mathrm{R})\ K * \varphi \cap K * \psi \subseteq K * (\varphi \vee \psi) \subseteq \mathrm{Cn}(K * \varphi \cup K * \psi).$$

If $*$ is a revision function over K, postulate ($*$R) is equivalent to each of the following postulates:

$$(*\mathrm{R}_{<\omega})\ \text{For every } n < \omega,$$
$$\text{if } \bigcap_{i<n} \mathrm{Cn}(\{\psi_i\} \subseteq \mathrm{Cn}(\{\varphi\})), \text{ then } K * \varphi \subseteq \mathrm{Cn}((\bigcup_{i<n} K * \psi_i) \cup \{\varphi\}).$$
$$(*\mathrm{R}_1)\ \text{If } \psi_0 \vee \psi_1 \in \mathrm{Cn}(\{\varphi\}), \text{ then } K * \varphi \subseteq \mathrm{Cn}(K * \psi_0 \cup K * \psi_1 \cup \{\varphi\}).$$

Furthermore, if $\mathcal{L}$ is finite and $*$ is a revision function over K, postulate ($*$R) is equivalent to the following postulate:

$$(*\mathrm{R}_\infty)\ \text{For every nonempty } I \subseteq \mathrm{For}(\mathcal{L}),$$
$$\text{if } \bigcap_{\psi \in I} \mathrm{Cn}(\{\psi\}) \subseteq \mathrm{Cn}(\{\varphi\}), \text{ then } K * \varphi \subseteq \mathrm{Cn}((\bigcup_{\psi \in I} K * \psi) \cup \{\varphi\}).$$

We record these observations in a proposition, leaving the proof to the reader:

Proposition 4.1 *Let $*$ be a revision function over* K. *Then* ($*$R), ($*R_{<\omega}$), *and* ($*R_1$) *are pairwise equivalent. If in addition $\mathcal{L}$ is finite, then* ($*R_\infty$) *is pairwise equivalent to each of the aforementioned postulates.*

[19] The 'g' in ($*7g$) is for 'Gärdenfors' (Rott, 2001, p. 110).

[20] Rott labels this postulate ($*8vwd$) in Rott (2001, p. 110).

8.4.2 Selection Functions in Belief Revision

The major innovation in Alchourrón et al. (1985) is the employment of selection functions to define operators of belief change. In Alchourrón et al. (1985), selection functions take *remainder sets* as arguments.[21] In this chapter we utilize selection functions which take *propositions* expressed by formulae as arguments, i.e., selection functions on the choice space $(\mathcal{W}_{\mathcal{L}}, \mathcal{E}_{\mathcal{L}})$ (see Section 8.2.1; see also Grove, 1988). Such selection functions are called *semantic selection functions*. Rott (2001) has shown that this approach is a fruitful generalization of the AGM approach.

Optimization, called *strong maximization* in Gärdenfors and Rott (1995, p. 65), is put to use in the classical AGM theory of belief change Alchourrön et al. (1985). There a selection function chooses the remainders of a remainder set that are 'best' in the sense that they are most worth retaining according to some non-strict ordering (the so-called 'marking-off' relation in Hansson, 1999, p. 82).[22]

It is also possible to apply maximization to study belief change. This notion, called *weak maximization* in Gärdenfors and Rott (1995, p. 65), is explored at length by the first author in Arló-Costa (2006), and Rott advocates using this notion in Rott (1993, p. 1430) and Rott (2001, p. 156). Indeed, there are good reasons to believe that this formalization is superior to the aforementioned formalization. Here we take a neutral position with respect to this issue, and most of our discussion about belief change will be framed with respect to optimization.

We point to a simple formal connection between rational choice on the one hand and belief change and non-monotonic reasoning on the other. In rational choice, G-rational and M-rational selection functions are often called *rationalizable*. However, in the study of belief change and non-monotonic reasoning, G-rational (i.e., strongly rationalizable) and M-rational (i.e., weakly rationalizable) selection functions are often called *relational*. Thus, formally speaking, rationalizablity in rational choice is equivalent to relationality in belief change and non-monotonic reasoning.

8.4.3 Rott's Correspondence Results

In this section we will review Rott's correspondence results linking conditions of belief revision and coherence postulates in the theory of rational choice. We will

[21] For a belief set K and a sentence φ, a *remainder set* $K\perp\varphi$ is the set of maximal consistent subsets of K that do not imply φ. Members of $K\perp\varphi$ are called *remainders*. Thus, in the AGM framework, a belief set K is fixed, and for every sentence φ such that $\varphi \notin \mathrm{Cn}(\emptyset)$, $\gamma(K\perp\varphi)$ selects a set of remainders of $K\perp\varphi$. The situation in which $\varphi \in \mathrm{Cn}(\emptyset)$ can be handled as a limiting case at the level of the selection function (Alchourrón et al., 1985) or at the level of the revision operator (Rott, 1993).

[22] In Alchourrón et al. (1985, pp. 517–518), a relation $\geq$ is defined over remainder sets for a fixed belief set K, and $\mathrm{Eq}_{\geq}$ is called the *marking off identity*:

$$\gamma(K\perp\varphi) = \{B \in K\perp\varphi : B \geq B' \text{ for all } B' \in K\perp\varphi\}.$$

present his results in a way that brings out their bearing upon rationalizabillty in belief change. We begin with several definitions (at this point the reader may wish to review Section 8.2.1).

Definition 4.2 A *semantic selection function* is a selection function on choice space $(\mathcal{W}_{\mathcal{L}}, \mathcal{E}_{\mathcal{L}})$.

Recall that $\mathcal{W}_{\mathcal{L}}$ denotes the collection of all maximal consistent sets of $\mathcal{L}$ with respect to Cn, while $\mathcal{E}_{\mathcal{L}}$ denotes the set of all elementary subsets of $\mathcal{W}_{\mathcal{L}}$.

Definition 4.3 Let γ be a semantic selection function.

(i) We define a semantic selection function $\overline{\gamma}$ by setting for all $S \in \mathcal{E}_{\mathcal{L}}$,

$$\overline{\gamma}(S) := \begin{cases} [\![\widehat{\gamma(S)}]\!] & \text{if } \gamma(S) \neq \emptyset \\ \emptyset & \text{otherwise.} \end{cases}$$

We call $\overline{\gamma}$ the *completion* of γ.

(ii) We say that γ is *complete* if $\gamma = \overline{\gamma}$.

Observe that for every $S \in \mathcal{E}_{\mathcal{L}}$, $\overline{\gamma}(S) \subseteq S$, so $\overline{\gamma}$ is a selection function. Also observe that for all $S \in \mathcal{E}_{\mathcal{L}}$, $\gamma(S) \subseteq \overline{\gamma}(S)$. Finally, observe that if $\mathcal{L}$ is finite, then every semantic selection function is complete.

We now define choice-based revision functions.

Definition 4.4 Let K be a belief set, and let γ be a semantic selection function. The *semantic choice-based revision function* $*$ *over* K *generated by* γ is defined by setting for every $\varphi \in \text{For}(\mathcal{L})$,

$$K * \varphi := \begin{cases} \widehat{\gamma([\![\varphi]\!])} & \text{if } \gamma([\![\varphi]\!]) \neq \emptyset \\ \text{For}(\mathcal{L}) & \text{otherwise.} \end{cases}$$

We say that γ *generates* $*$ or that $*$ is *generated by* γ.

To bring the ideas concerning rationalizablity to the foreground, we offer the following definition.

Definition 4.5 Let K be a belief set. We call a function $*$ a (*regular, rational, G-rational, complete*) *choice-based revision function* over K if there is a (regular, rational, G-rational, complete) semantic selection function γ that generates $*$.

Observe that every semantic choice-based revision function over a belief set K satisfies postulates $(*1)$, $(*2)$, and $(*6)$ and so is indeed a revision function over K. It is an easy matter to check that the converse holds as well: If $*$ is a revision function over a belief set K, then $*$ is a semantic choice-based revision function over K.

Also observe that $*$ is a semantic choice-based function over K generated by γ if and only if for every sentence ψ of $\mathcal{L}$,

$$\psi \in K * \varphi \text{ if and only if } \gamma([\![\varphi]\!]) \subseteq [\![\psi]\!] \,.$$

Intuitively, an agent believes a sentence ψ in the revision of K by φ just in case ψ is true in all the most 'plausible' worlds in which φ is true. Of course, the role of a semantic selection function—or any selection function—can be interpreted in various ways in different contexts.

Rott (2001) discusses a handful of coherence constraints for selection functions, some of which are well-known and others of which he debuts. We present two conditions of the latter sort without offering motivation (see Rott 2001, pp. 147–149) for such motivation):

(F1$_B$) For every $S \in \mathcal{S}$, if $S \cap B \neq \emptyset$, then $\gamma(S) \subseteq B$. (*Faith 1 respect to B*)
(F2$_B$) For every $S \in \mathcal{S}$, $S \cap B \subseteq \gamma(S)$. (*Faith 2 respect to B*)

Let us now see how some of the coherence constraints—especially condition α—are intimately connected with the presumption that selection functions are rationalizable in the study of belief change. Here we turn to Rott's recent correspondence results. Among other things, Rott's recent results establish a connection between condition α and postulate $(*7)$ of belief revision.[23] Presented in a form suitable for this article, the following theorem provides one part of this connection (Rott 2001, p. 197).

Theorem 4.6 (Rott, 2001) *Let K be a belief set. For every semantic selection function γ which satisfies*

[23] Rott's results (2001) show much more. For example, Rott shows that condition α corresponds not only to posutlate $(*7)$, but also to postulate $(\dot{-}7)$ of belief contraction (which requires that $K \dot{-} \varphi \cap K \dot{-} \psi \subseteq K \dot{-} (\varphi \wedge \psi)$) [Rot01, pp. 193-196] and to rule (*Or*) of non-monotonic reasoning (which demands observance of the following: From $\varphi \mid\sim \chi$ and $\psi \mid\sim \chi$, infer $\varphi \vee \psi \mid\sim \chi$) (Rott 2001, pp. 201–204).

$$\left\{\begin{matrix} \mathrm{F1}_{[\![K]\!]} \\ \mathrm{F2}_{[\![K]\!]} \\ \gamma > \emptyset \\ \alpha \\ \gamma^{*} \text{ and is complete} \\ \gamma\mathrm{R} \text{ and is complete} \end{matrix}\right\}, \textit{the semantic choice-based revision function } * \textit{ over } K$$

$$\textit{generated by } \gamma \textit{ satisfies } \left\{\begin{matrix} - \\ (*4) \\ (*3) \\ (*5) \\ (*7) \\ (*8r) \\ (*\mathrm{R}) \end{matrix}\right\}, \textit{respectively.}$$

Theorem 4.6 is a ‘soundness’ result. (The reader should observe the modular character of Theorem 4.6 as well as Theorem 4.7. Theorem 4.6, for example, says that for every belief set K and semantic selection function γ, *if* γ satisfies condition $\mathrm{F1}_{[\![K]\!]}$, *then* the semantic choice-based revision function $*$ over K generated by γ satisfies postulate $(*4)$; Theorem 4.6 *also* says that for every belief set K and semantic selection function γ, *if* γ is complete and satisfies condition α, *then* the semantic choice-based revision function $*$ over K generated by γ satisfies postulate $(*7)$. Rott also establishes a number of ‘completeness’ results. Also presented in a form suitable for this chapter, the following completeness result is the other part of the connection between coherence constraints and rationality postulates of belief revision (Rott 2001, p. 198).

Theorem 4.7 (Rott 2001) *Every revision function $*$ over a belief set K which satisfies*

$$\left\{\begin{matrix} - \\ (*3) \\ (*4) \\ (*5) \\ (*7) \\ (*8r) \\ (*\mathrm{R}) \end{matrix}\right\} \textit{can be represented as the semantic choice-based revision function}$$

$$\textit{over } K \textit{ generated by a semantic selection function } \gamma \textit{ which satisfies } \left\{\begin{matrix} - \\ \mathrm{F2}_{[\![K]\!]} \\ \mathrm{F1}_{[\![K]\!]} \\ \gamma_{>\emptyset} \\ \alpha \\ \gamma^{*} \\ \gamma\mathrm{R} \end{matrix}\right\},$$

respectively.

Observe that the preceding theorems do not presuppose any basic postulates other than $(*1)$, $(*2)$, and $(*6)$. Since $(\mathcal{W}_{\mathcal{L}}, \mathcal{E}_{\mathcal{L}})$ is closed under finite unions and relative complements and is compact, we can apply the results from the previous section to obtain the following corollaries which are of particular relevance for the purposes of this chapter.[24]

Corollary 4.8 *Let K be a belief set.*

(i) *Every rational choice-based revision function* $*$ *over* K *is a revision function satisfying* $(*7)$*, and every rational complete choice-based revision function* $*$ *over* K *satisfies* $(*7)$ *and* $(*8\mathrm{r})$.
(ii) *Every revision function* $*$ *over* K *satisfying* $(*7)$ *and* $(*8\mathrm{r})$ *is a rational (complete) choice-based revision function over* K.

Corollary 4.9 *Let K be a belief set.*

(i) *Every regular G-rational choice-based revision function* $*$ *over* K *is a revision function satisfying* $(*5)$ *and* $(*7)$*, and every regular G-rational complete choice-based revision function* $*$ *over* K *satisfies* $(*5)$, $(*7)$, *and* $(*8\mathrm{r})$*.*
(ii) *Every revision function* $*$ *over* K *satisfying* $(*5)$, $(*7)$, *and* $(*8\mathrm{r})$ *is a regular (complete) G-rational choice-based revision function over* K.

We remark that if $\mathcal{L}$ is infinite, then for no $w \in \mathcal{W}_{\mathcal{L}}$ is it the case that $\{w\} \in \mathcal{E}_{\mathcal{L}}$. Thus, every selection function on $(\mathcal{W}_{\mathcal{L}}, \mathcal{E}_{\mathcal{L}})$ satisfies $\gamma_{1>\emptyset}$. It follows that if $\mathcal{L}$ is infinite, then one can drop the requirement of regularity in Corollary 4.9. Yet if $\mathcal{L}$ is finite, one must impose regularity to guarantee G-rationality.

Corollary 4.10 *Let* $\mathcal{L}$ *be finite, and let K be a belief set.*

(i) *A function* $*$ *is a rational choice-based revision function over K if and only if it is a revision function satisfying* $(*7)$ *and* $(*8\mathrm{r})$*.*
(ii) *A function* $*$ *is a regular G-rational choice-based revision function over* K *if and only if it is a revision function satisfying* $(*5)$, $(*7)$, *and* $(*8\mathrm{r})$*.*

With respect to rationalizability, the foregoing results are quite general. As indicated above, the only basic postulates presumed to hold in these results are $(*1)$, $(*2)$, and $(*6)$ (rationality postulates and coherence constraints can be added and subtracted modularly). According to the results, if there is a rational selection function γ that generates a belief revision function $*$, then $*$ must satisfy

[24] Sen's Property β^+ was mentioned in footnote 11 because it has been given special attention in connection to rationalizablity in belief change and non-monotonic reasoning (e.g., in Rott 1993, 2001, and Arló-Costa 2006). It is known that condition β^+ corresponds to belief revision's rationality postulate $(*8)$, which requires that $K*\varphi \subseteq K*(\varphi\wedge\psi)$ whenever $\neg\psi \notin K*\varphi$ (see Rott 2001, p. 198). But since condition β *is* more pervasive in the study of rational choice, one might ask the following question: What rationality postulate corresponds to condition β? The second author has shown elsewhere (Pedersen 2008) that condition β corresponds to postulate $(*8\beta)$, which demands that if $\mathrm{Cn}(K * \varphi \cup K * (\varphi \wedge \psi)) \neq \mathrm{For}(\mathcal{L})$, then $K * \varphi \subseteq K * (\varphi \wedge \psi)$.

postulate $(*7)$. In the next section, we will offer plausible examples in which postulate $(*7)$ is violated. And this should raise eyebrows: For such violations imply that *no rational selection function exists that models the agent's belief change,* and *a fortiori*, no G-rational or M-rational selection function exists that models the belief change. Thus, the universal presumption in the study of belief change and non-monotonic reasoning—that selection functions are rationalizable—must be called into question.

But simply abandoning this presumption is much too quick and much too damaging. The utilization of rationalizable selection functions in the study of belief change and non-monotonic reasoning has proved to be quite useful and indeed indispensable, so it would be valuable to see what can be salvaged from the theoretical wreck.

We claim that the phenomenon of menu dependence—in epistemic form—is to blame for these violations of $(*7)$. The counterexamples we offer in this article illustrate the role of menu dependence in the context of belief change and non-monotonic reasoning. We will indicate what formal measures may be taken to anticipate menu dependence, which nonetheless admit a restricted form of optimization and maximization.

8.4.4 Counterexamples

In this section we will offer counterexamples to postulate $(*7)$. Actually, these counterexamples violate many postulates of belief revision, just as counterexamples *à la* Sen violate many coherence constraints of rational choice, including condition α.

Each counterexample involves three hypothetical scenarios in which an agent accepts belief-contravening information. Each scenario describes a potential unfolding of events. The scenarios in the counterexamples are *not* consecutive stages of a single chain of events. Rather, each scenario describes one way things could turn out. Moreover, only one of these scenarios will be realized.

Following each example, we will indicate how postulate $(*7)$ fails. The first example is essentially a reproduction of an example presented by Rott (2004). We will see that menu dependence plays an essential role in this example. Then we will introduce variants of Rott's example where menu dependence can be explained in terms of the intervention of social norms.

8.4.4.1 Example

A philosophy department has announced an open position in metaphysics. Tom, an interested bystander, happens to know a few of the applicants: Amanda Andrews, Bernice Becker, Carlos Cortez, and Don Doyle. Tom, just like everyone else, knows that Andrews is an outstanding specialist in metaphysics, whereas Becker, who is also a very good metaphysician, is not quite as excellent as Andrews. However, Becker has done some substantial work in logic. Cortez has a comparatively slim record in metaphysics, yet he is widely recognized as one of the most brilliant logicians of his generation. By contrast, Doyle is a star metaphysician, while Andrews has done close to no work in logic.

Now suppose Tom initially believes that neither Andrews, Becker, nor Cortez will be offered the position because he, like everyone else, believes that Doyle is the obvious candidate to be offered the position. Tom is well-aware that only one of the applicants will be offered the position. Let a, b, c, and d stand for the following sentences:

a: Andrews will be offered the position.
b: Becker will be offered the position.
c: Cortez will be offered the position.
d: Doyle will be offered the position.

Tom is having lunch with the dean. The dean is a very competent, serious, and honest man. He is also the chairman of the selection committee.

Scenario 1. The dean informs Tom that either Andrews or Becker will be offered the position. That is, the dean informs Tom that $a \vee b$. Because Tom presumes that expertise in metaphysics is the decisive criterion for the selection committee's decision, Tom concludes that Andrews will be offered the position (and of course that all other applicants will not be offered the position).

Scenario 2. The dean confides to Tom that either Andrews, Becker, or Cortez will be offered the position, thereby supplying him with $a \vee b \vee c$. Because Cortez is a brilliant logician, Tom realizes that he cannot sustain his presumption that metaphysics is *the* decisive criterion for the selection committee's decision. From Tom's perspective, logic *also* appears to be regarded as a considerable asset by the selection committee. Nonetheless, because Cortez has such a slim record in metaphysics, Tom believes that Cortez will not be offered the position. But Tom sees that logic contributes to an applicant's chances of being offered a position. Tom thereby concludes that Becker will be offered the position (and so no other applicant will be offered the position).

Scenario 3. The dean tells Tom that Cortez will be offered the position, thereby supplying him with c. Tom is certainly surprised, yet he believes what the dean tells him.

□

Let us take stock of Tom's beliefs in these scenarios. Initially, Tom believes d, $\neg a$, $\neg b$, and $\neg c$. Thus, letting K denote Tom's initial belief set, $d, \neg a, \neg b$ and $\neg c$ are in K. In Scenario 1, Tom's revises his belief set K by $a \vee b$, and his revised belief set $K * (a \vee b)$ contains a and $\neg b$, as well as $\neg c$ and $\neg d$. In Scenario 2, Tom revises his belief set K by $a \vee b \vee c$. His revised belief set $K * (a \vee b \vee c)$ includes b, $\neg a$, $\neg c$, and $\neg d$. Finally, in Scenario 3, Tom revises his belief set K by c, whereby his revised belief set $K * c$ contains c, $\neg a$, $\neg b$, and $\neg d$.

We are now in a position to see that Example 3.4.1 constitutes a violation of postulate ($*7$). First, observe that $\neg b \in K*(a \vee b) \cap K*c$ and $\neg b \notin K*(a \vee b \vee c)$. Hence,

$$K*(a \vee b) \cap K*c \nsubseteq K*(a \vee b \vee c).$$

So postulate ($*7g$) and therefore postulate ($*7$) is violated.

In light of Theorem 4.6, we should be unsurprised to see that condition α* is also violated. And it is. Let γ be *any* semantic selection function that generates $*$. Then it must be the case that $\gamma([\![a \vee b \vee c]\!]) \subseteq [\![\neg a \wedge b \wedge \neg c \wedge d]\!]$, $\gamma([\![a \vee b]\!]) \subseteq [\![a \wedge \neg b \wedge \neg c \wedge d]\!]$, and $\gamma([\![c]\!]) \subseteq [\![\neg a \wedge \neg b \wedge c \wedge \neg d]\!]$. It must further be the case that $\gamma([\![a \vee b \vee c]\!]) \nsubseteq [\![\neg b]\!]$, whereby $\gamma([\![a \vee b \vee c]\!]) \cap [\![b]\!] \neq \emptyset$. It follows that

$$\gamma([\![a \vee b \vee c]\!]) \nsubseteq \gamma([\![a \vee b]\!]) \cup \gamma([\![c]\!]).$$

Thus, condition α^* is violated.[25]

The phenomenon of menu dependence seems to explain the choices made in this case. When Tom faces the menu represented by $a \vee b$, he does it under the presumption that metaphysics is the decisive criterion for the selection committee's decision. Therefore, when he has to judge the relative merits of Andrews and Becker as candidates, Tom concludes that Andrews will be offered the position. But the disclosure of certain facts about Cortez in Scenario 2 alters Tom's evaluation of the relative merits of Andrews and Becker as candidates and as a consequence Tom concludes that Becker will be offered the position instead. Since the information Tom receives includes certain facts about Cortez, and since this information has been acquired from a reliable source (viz., the dean), Tom *learns* something important about the selection criterion used by the selection committee (viz., that expertise in metaphysics is not the only decisive criterion used by the selection committee). So we can say that Tom's epistemic choice from the menu represented by $a \vee b \vee c$ has *epistemic relevance* for Tom's epistemic decision and that Tom's epistemic choices are *menu dependent*.

Notice that the *mere* inclusion of facts about Cortez in the extended menu does not trigger the phenomenon of menu dependence. One needs to know in addition that the extra information has been acquired from a reliable source. The dean satisfies this requirement, but the example does not depend on the *identity* of the dean (as happens in many of the examples proposed by Sen). If, for example, the information were provided by a member of the selection committee the example would be equally effective in triggering a case of menu dependence.[26]

[25] In fact, as Rott (2004) points out, Aizerman's Axiom is also violated (see Section 8.2, for a statement of this condition). This coherence constraint corresponds to postulate ($*8c$) via Rott's correspondence results (2001, pp. 197–198). (Postulate ($*8c$) demands that if $\psi \in K*\varphi$, then $K*\varphi \subseteq K*(\varphi \wedge \psi)$.) This means that *pseudo-rationalization* is precluded (see footnote 5).

[26] David Makinson suggested to us in a private communication that there might be cases of *pure* menu dependence, where the choice depends on the content of the menu, irrespective of the context

Here we want to consider two objections that one might raise to Example 3.4.1. The first objection is that we have not accurately represented the example. That is, one might object that we have not accurately represented the information that Tom receives from the dean. Allow us to illustrate this objection by way of example. On the one hand, in Scenario 2, when the dean informs Tom that either Andrews or Becker or Cortez will be offered the position ($a \vee b \vee c$), one might contend that the dean's information at least leaves it open for Cortez to be offered the position. On the other hand, in Scenario 1, when then dean informs Tom that either Andrews or Becker will be offered the position ($a \vee b$), the possibility that Cortez will be offered the position is seemingly excluded. So, one might say, we should have illustrated this difference in Scenario 1 by representing the dean's information by, say, $(a \vee b) \wedge \neg c$. Rott (2004) addresses this sort of objection. Readers who are sympathetic with this objection should consult Rott's article.

The second objection is that we have conflated the notion of belief with expectation.[27] One might contend that in Scenario 2, for example, Tom does not come to *fully believe* that Becker will get the position. Rather, Tom comes to fully believe, say, only that precisely one of Andrews, Becker, and Cortez will be offered the position and that Doyle will not be offered the position. Importantly, Tom only *strongly expects* that Becker will be offered the position. After all, if Tom *were* to fully believe that Becker will be offered the position, Tom would apparently be jumping to conclusions.

Indeed, expectations may be different from beliefs. And expectations may guide our beliefs without quite being part of them (Gärdenfors and Makinson 1994, p. 2). Be this as it may, a principled distinction between expectations and beliefs has been quite elusive. Rott (2001) writes:

> We have not found a sharp boundary between beliefs and expectations. Any potential standard for separating beliefs from expectations may be contextually shifted, according to the situation one is facing. It is very doubtful whether an agent makes use of the same set of propositions in different situations—even if his doxastic state does not change at all. A loose conjecture may count as a full belief in party chat, but in the courtroom one ought to be firmly convinced of the truth of a proposition in order to affirm that one believes it to be true. Pragmatic considerations are needed to determine what qualifies as a belief (p. 29).

Gärdenfors and Makinson (1994) express a similar view:

> Epistemologically, the difference between belief sets and expectations lies only in our attitude to them, i.e., what we are willing to do with them. For so long as we are *using* a belief

in which it is offered. It is unclear whether there are pure cases of this sort. It seems difficult to find examples that do not depend on the reliability of the information source used to extend the menu. In any case, the distinction between pure and impure cases of menu dependence seems worthy of further analysis.

[27] The notion of expectation here should not be confused with the notion of expected utility in rational choice. Whereas expected utility concerns expectations of the values of various outcomes, here expectation concerns beliefs about the world (see Gärdenfors and Makinson 1994, p. 5). Perhaps the only dissenting voice regarding this point is Levi who in [Lev96] treats expectations as cognitive expected value. We use expectations here in the first epistemic sense of the word.

> set K, its elements function as full beliefs. But as soon as we seek to *revise* K, thus putting its elements into question, they lose the status of full belief and become merely expectations... (p. 35).

In this article, we will not attempt to explicate the distinction between beliefs and expectations. But we fully agree that Example 3.4.1 is most convincing if what is revised are *expectations* rather than full beliefs. Thus, we may take K in Example 3.4.1 to be the agent's expectations, without defeat, for this is a presumption often made in the literature of belief change (e.g., in Spohn, to appear; Pearl and Goldszmidt 1996; Gärdenfors and Makinson 1994), and in particular, by Rott (2001). This issue notwithstanding, the example still shows that postulate ($*7$) (and so condition α^*) is violated,[28] and the formal implications of these violations are what are at issue in this chapter. [29]

So much for the first example. We now present an example that is structurally similar to the first example but where social norms play a crucial role.

8.4.4.2 Example

The candidates for a position in epistemology in a philosophy department are Anita Adams, John Becker, Peter Collins, Don Doyle, and Sasha Earl. Don Doyle is a star epistemologist, while John Becker is close in running. He is only slightly surpassed by Doyle with respect to objective merits. To be sure, Becker has a very good record of publications, and he is a candidate that gave one of the the best talks. Anita Adams is also a very good candidate. She happens to be African-American. Although she does excellent work, her work is not quite as good as Becker's work. Yet she is younger than Becker, and she is quite promising. Another candidate, Sasha Earl, is surpassed by Anita Adams in terms of objective merits. She is nonetheless a good candidate for the position. As it turns out, she also happens to be African-American.

Finally we have Peter Collins. Collins has done work of comparable quality to the work of Adams. But he is well-known for his political support of groups that promote white supremacy, and he is still involved in an unsettled case of harassment at a different university.

[28] Other supplementary postulates are violated as well, such as postulate ($*8c$) (see footnote 25). No basic postulates are violated in this example.

[29] Again, Levi traces a sharp distinction separating full beliefs and expectations. In Levi (1996) he distinguishes between the ordinary versions of the postulates of belief change, like postulate ($*7$), and inductively extended versions of these postulates. And he has pointed out that the inductively extended version of postulate ($*7$) fails to hold. So, we assume that this would be his preferred explanation of examples like the one offered by Rott. Perhaps the distinction holds even if one does not buy Levi's theory of induction, and if one uses a different theory of induction instead. So the issue of what counts as a counterexample to well-known principles of belief formation sanctioned by AGM depends on a previous understanding of the notion of expectation as opposed to full belief. While Levi's strategy might work (assuming that one buys his notion of expectation) in the cases where menu dependence is the main mechanism, it is unclear whether this strategy applies to cases where the main underlying mechanism is determined by the use of social norms. More about this will be discussed below.

Now suppose Tom initially believes that neither Adams, Becker, Earl, nor Collins will be offered the position because he, like everyone else, believes that Doyle is the obvious candidate to be offered the position. Thus, Tom believes

$\neg a$: Adams will not be offered the position.
$\neg b$: Becker will not be offered the position.
$\neg c$: Collins will not be offered the position.
d: Doyle will be offered the position.
$\neg e$: Earl will not be offered the position.

Tom ran into the dean while crossing campus, and they have decided to find a place to sit and chat. Tom takes the dean to be a very competent, serious, and honest man. Tom is also aware that the dean is the chairman of the selection committee.

Now consider the following scenarios:

Scenario 1. Tom learns from the dean that Adams, Becker, or Earl will be offered the position in epistemology ($a \vee b \vee e$). Although Adams' work is surpassed by the work of Becker, she belongs to two demographic groups and is a better candidate for the position than Earl. Accordingly, Tom concludes that Adams will be offered the position.

Scenario 2. The dean informs Tom that Collins will be offered the position (c). Tom knows that the dean is an upright individual, so he does not doubt what the dean has told him.

Scenario 3. Tom learns from the dean that Adams, Becker, Collins, or Earl will be offered the position ($a \vee b \vee c \vee e$). In this case the presence of Collins signals to Tom that the department might not take into account affirmative action. He concludes that the position will be offered to Becker.

□

The analysis of Tom's beliefs is similar to the previous example (among other things, we have $\neg b \in K * (a \vee b \vee e) \cap K * c$ and $\neg b \notin K * (a \vee b \vee c \vee e)$, so postulate ($*7$) is violated). But here we have the influence of a social norm in Scenario 1. Clearly in terms of considerations of objective merits alone Becker is expected to surpass Adams, Collins, and Earl. Nevertheless, when faced with the information given by $a \vee b \vee e$, that Becker will be offered the position (b) is rendered unfeasible by a norm according to which, all things considered, candidates belonging to disadvantaged groups should be selected. Tom thereby concludes that Adams will be offered the position (a). But when Tom is faced with the information given by $a \vee b \vee c \vee e$, we witness the phenomenon of menu dependence at work. Tom sees that the above social norm is not taken into account, and in this case Tom's belief change is guided by considerations of objective merits alone and so Tom concludes that Becker will be offered the position (b).

Let us consider an analysis of this example in terms of possible worlds. We have five relevant worlds: $w_1 = (\neg a, \neg b, \neg c, d, \neg e)$, $w_2 = (\neg a, b, \neg c, \neg d, \neg e)$, $w_3 = (a, \neg b, \neg c, \neg d, \neg e)$, $w_4 = (\neg a, \neg b, c, \neg d, \neg e)$, and $w_5 = (\neg a, \neg b, \neg c, \neg d, e)$.[30] Given the background considerations of objective merits of the candidates, it is reasonable to suppose that we have an ordering $\geq$ over possible worlds which works as follows: w_1 dominates w_2, w_2 dominates w_3 and w_4, and w_3 is tied with w_4, while the latter tied worlds dominate w_5. Thus, where, $\sim$ and $>$ are defined in the usual way ($w \sim v$:iff $w \geq v$ and $v \geq w$; $w > v$:iff $w \geq v$ and $v \not\geq w$), $w_1 > w_2$, $w_2 > w_3$ and $w_2 > w_4$, $w_3 \sim w_4$, and $w_3 > w_5$ and $w_4 > w_5$ (all other worlds are dominated by w_1, w_2, w_3, w_4, and w_5, and $w_i \geq w_j$ for $i \leq j$).

Now when $a \vee b \vee e$ is learned the underlying norm related to the promotion of underrepresented demographic groups makes the world w_2 unfeasible, while two worlds remain feasible, w_3 and w_5. Since Adams dominates Earl in terms of objective merits, Tom concludes that Adams will be offered the position (a), whereby w_3 becomes admissible. Thus, if γ is a semantic selection function that generates the revision function $*$, we have $\gamma([\![a \vee b \vee e]\!]) = \{w_3\}$. But when the quadruple disjunction $a \vee b \vee c \vee e$ is learned the norm is seen to be playing no role. Considerations of objective merits alone lead Tom to conclude that Becker will be offered the position (b). We accordingly have $\gamma([\![a \vee b \vee c \vee e]\!]) = \{w_2\}$. Of course, since this example is an illustration of a violation of postulate ($*7$), there is no rational semantic selection function that models Tom's belief revision. Yet it seems reasonable to assume that the above ordering is relevant in this example.

Indeed, we can maintain the ordering $\geq$ over possible worlds, viewing the semantic selection function γ that models Tom's belief revision as conditional upon a permissibility function π. When Tom learns $a \vee b \vee e$ in Scenario 1, the underlying social norm renders w_2 unfeasible while permitting w_3 and w_5 to remain feasible. We can thereby view the role of the permissibility function in such a way that $w_2 \notin \pi([\![a \vee b \vee e]\!])$ and $\{w_3, w_5\} \subseteq \pi([\![a \vee b \vee e]\!])$. If $\gamma([\![a \vee b \vee e]\!])$ is then understood to select those worlds optimal among those from $\pi([\![a \vee b \vee e]\!])$, then we have that $\gamma([\![a \vee b \vee e]\!]) = \{w \in \pi([\![a \vee b \vee e]\!]) : w \geq w'$ for all $w' \in \pi([\![a \vee b \vee e]\!])\} = \{w_3\}$. By contrast, when Tom learns $a \vee b \vee c \vee e$ in Scenario 3, the norm is seen to not be taken into account, so it is reasonable to take $\pi([\![a \vee b \vee c \vee e]\!]) = [\![a \vee b \vee c \vee e]\!]$, whence $\gamma([\![a \vee b \vee c \vee e]\!]) = \{w \in \pi([\![a \vee b \vee c \vee e]\!]) : w \geq w'$ for all $w' \in \pi([\![a \vee b \vee c \vee e]\!])\} = \{w_2\}$. In the next section we will return to this sort of analysis of belief change, introducing what we call *norm-inclusive belief revision*.

It is clear that there is an intimate connection between the underlying social norm relevant to the example and the expectations generated by it. Notice in the above example that although expectations are involved we do not have an inductive machinery (of the sort used by Levi, 1996) to generate them. All epistemic

[30] Here we adopt a notational convention: the expression $(a, \neg b, \neg c, \neg d, \neg e)$ denotes the possible world for which a is true and the rest of the items – b, c, e, d – are false.

choices are explained in terms of a unique ordering of epistemic options plus considerations of feasibility, which, in turn, are the consequence of the operation of underlying norms.[31]

Let us now consider another hiring example with a different structure. The example is simpler than the previous two examples considered above.

8.4.4.3 Example

Jeff Johns and Mara Lee Pearl are two outstanding candidates for a job search a philosophy department has been conducting. They are married, and Johns is a better candidate than Pearl all things considered. But Pearl is a decent candidate who could be a good addition to the department as a teacher. Johns has already another offer from an university in a different town, but, all things considered he would prefer an offer from this department. Tom, just like everybody else, believes that due to budget cuts the department will not be able to offer a position to either of them ($\neg j \wedge \neg p$). Consider the following three scenarios:

Scenario 1. Tom is informed that Johns will be hired (j). Under the point of view of merit and the convenience of the department the two states ($j \wedge \neg p$) and ($j \wedge p$) are tied, but Tom applies a norm in this case, whereby all things considered the unity of the family should be preserved. He concludes that both will be hired ($j \wedge p$).

Scenario 2. Tom learns that Johns will not be hired but that Pearl will be hired ($\neg j \wedge p$).

Scenario 3. Tom learns that either Johns will be hired or that Johns will not be hired but his wife will be hired. In this case Tom receives information that is compatible with a situation where the couple will have jobs in different towns ($\neg j \wedge p$). The fact that this situation (which is the worst option for the merit ranking) is considered possible convinces Tom that the aforementioned norm does not apply here and by considerations of merit alone concludes that Johns will be hired (j).

□

So, we have that p belong to the intersection of the first two revisions, but p does not belong to the revision with the disjunction of the first two items.

An analysis in terms of possible worlds is possible here as well. The relevant worlds are $w_1 = (j, p)$, $w_2 = (\neg j, \neg p)$, $w_3 = (j, \neg p)$ and $w_4 = (\neg j, p)$. Under the point of view of merit and the convenience of the department worlds w_1 and w_3 are

[31] A possible solution of compromise between Levi's inductive approach and the use of norms in examples of this sort could be to say that we are dealing with norms that induce or generate expectations. But in this case, these expectations do not seem to be those from a theory like Levi's but rather the purely epistemic expectations usually used in theories of belief change. Still, we can see a norm-sensitive operator of belief change as the composition of two operators: one classical AGM operator plus an inductive operator sanctioning an inductive jump made possible by the underlying norms. Most of what follows can be seen as a positive theory about this operator—where we provide new axioms that the operator should obey.

tied and optimal. The worst option in this ordering is w_4. When j is learned the world w_3 is made unfeasible by an underlying norm promoting the unity of families whenever possible ($\pi(\{w_1, w_3\}) = \{w_1\}$). When the disjunction is learned in the third scenario the norm is deactivated and the agent settles in a theory corresponding to these three relevant worlds ($\pi(\{w_1, w_2, w_3\}) = \{w_1, w_2, w_3\}$). As with the previous examples, this represents a violation of postulates ($*7$) and condition α.

At this point we expect to have convinced the reader that social norms are as ubiquitous in epistemology as they are in rational choice. It seems that there is a robust connection between the structure of our expectations and social norms. Social norms seem to justify many of our expectations and they seem crucial in the way we change them in the presence of new information.

To focus on a different type of social norm, let's consider social norms related to undergraduate students' beer drinking habits on the main campuses of American universities. Say that Tom is convinced that John is studying at home for an exam. Say as well that in this situation Tom learns that either John is at the local bar drinking beer with his friends or he is at the local bar drinking tea with his friends. Given this information, Tom would probably conclude that John is drinking beer with his friends. This expectation is formed by taking into account the norm that sanctions beer drinking habits in bars of this type among students. The norm plays a crucial role in the way one might form and change one's expectations.

Here we wish to emphasize that we are considering only social norms, not norms of another type (like legal norms). Social norms are the type of norms that Sen considers relevant in the realm of social choice. Our point here is that these norms are equally relevant in epistemology. Their main role is related to the process of forming and changing our expectations.

In this section we have offered examples in which postulate ($*7$) (among many other postulates) is violated. Yet the belief changes involved in these examples seem perfectly reasonable. It is indeed evident that some violations of postulate ($*7$) can be explained in terms of the influence of social norms in belief formation. Our goal in the remainder of this article is to develop a theory of belief revision which takes into account the role social norms play in belief formation.

8.5 Part IV: Norm-Inclusive Belief Revision

In Section 8.4.3, we reviewed Hans Rott's correspondence results for belief revision, which show how postulates of belief revision correspond in a one-to-one fashion to coherence constraints of rational choice. We saw how these results, furnished within a general framework, bear upon rationalizability in belief revision.

In Section 8.4.4, we encountered several counterexamples to postulates of belief revision—in particular, to postulate ($*7$). In virtue of Rott's correspondence results, these counterexamples threatened the all but universal presumption that belief revision is relational. We argued that many counterexamples which threaten this presumption are driven by the influence social norms have in belief formation. We have reviewed recent work concerned with accommodating the role social norms play in choice. We have also improved upon the results of this work, developing a theory of norm-conditional choice.

The primary purpose of this section is to introduce a new theory of belief revision, called *norm-inclusive belief revision*. Like the theory of norm-conditional choice we discussed earlier, norm-inclusive belief revision is intended to take into account the role social norms play in belief change. Also like the theory of norm-conditional choice, norm-inclusive belief revision is an extension of the classical belief revision theory investigated by many researchers in formal epistemology.

In this section we will introduce postulates of norm-inclusive belief revision. We will then state and prove correspondence theorems for norm-inclusive belief revision, providing a direct connection between conditions imposed upon norm-conditional choice models and conditions placed upon what we call *norm-inclusive revision models*. We will conclude with discussion and an example that illustrates our theory at work.

8.5.1 Postulates

Roughly, a norm-inclusive belief revision model for a belief set K is a pair of the form $(*, \boxast)$, where $*$ is what we call a $\boxast$-*inclusive revision function* over K and $\boxast$ is what we call a *norm representation function* for $*$. Intuitively, $K * \varphi$ is the revision of K that is compatible with those beliefs $K \boxast \varphi$ an underlying social norm warrants for acceptance when φ is learned.[32]

As with standard theories of belief revision, we presume that $*$ satisfies postulates $(*1)$, $(*2)$, and $(*6)$. We also presume $\boxast$ satisfies the following postulates:

$(\boxast 1)$	$K \boxast \varphi = \mathrm{Cn}(K \boxast \varphi)$.	(*Normative Closure*)
$(\boxast 2)$	$\varphi \in K \boxast \varphi$.	(*Normative Success*)
$(\boxast 6)$	If $\mathrm{Cn}(\{\varphi\}) = \mathrm{Cn}(\{\psi\})$, then $K \boxast \varphi = K \boxast \psi$.	(*Normative Extensionality*)

Postulate $(\boxast\ 1)$ simply requires that the set of beliefs an underlying social norm warrants for acceptance is closed under logical consequence. According to postulate $(\boxast\ 2)$, even when a social norm is operative in a revision, a norm representation function must give the incoming information priority. Postulate $(\boxast\ 6)$ demands that a norm representation function treats logically equivalent information in the same way. In particular, a norm representation function is not sensitive to the linguistic formulation of incoming information.

With this we can offer a precise formulation of what we mean by a norm-conditional belief revision model.

[32] A norm warrants a belief ψ for acceptance if the norm makes the acceptance of ψ permissible. The warranted beliefs are the beliefs permitted by the norm.

Definition 5.1 Let K be a belief set, let $*$ be a function satisfying $(*1)$, $(*2)$, and $(*6)$, and let $\boxast$ be a function satisfying $(\boxast\ 1)$, $(\boxast\ 2)$, and $(\boxast\ 6)$. We call the pair $(*, \boxast)$ a *norm-inclusive revision model* over K if for every $\varphi \in \mathrm{For}(\mathcal{L})$, $K \boxast \varphi \subseteq K * \varphi$.[33]

Thus, we require that a norm-inclusive belief revision model satisfies the following postulate:

$$(*[\boxast]) \qquad K \boxast \varphi \subseteq K * \varphi. \qquad (\textit{Norm-Inclusive Revision})$$

As indicated above, the intuition here is that for an agent revising its beliefs K by a sentence φ, $K \boxast \varphi$ represents those beliefs that an underlying social norm warrants for acceptance when φ is learned. Thus, postulate $(*[\boxast])$ signifies that an agent ought to believe every sentence ψ an underlying norm warrants for acceptance in the revision of K by φ.

The following postulates are analogues of postulates $(*3)$, $(*4)$, and $(*5)$:

$(\boxast\ 3)$ $\quad K \boxast \varphi \subseteq \mathrm{Cn}(K \cup \{\varphi\})$.
$(\boxast\ 4)$ $\quad$ If $\neg\varphi \notin K$, then $\mathrm{Cn}(K \cup \{\varphi)\} \subseteq K \boxast \varphi$.

Clearly, a $\boxast$-inclusive revision function $*$ over K satisfies postulate $(*3)$ only if $\boxast$ satisfies postulate $(\boxast 3)$, and $*$ satisfies postulate $(*4)$ provided satisfies postulate $(\boxast\ 4)$. We stress that a norm representation function $\boxast$ need not represent a *belief change* operation, so one might find it undesirable to impose postulate $(\boxast\ 3)$ and especially postulate $(\boxast 4)$. Yet if $*$ is intended to represent an AGM-style belief revision operation, the following postulate seems to be a reasonable constraint for norm representation functions:

$(\boxast_{\iota,\, K})$ $\quad$ If $\neg\varphi \notin K$, then $K \boxast \varphi \subseteq \mathrm{Cn}(\{\varphi\})$. $\quad$ (*Norm in Absentia with respect to K*)

According to postulate $(\boxast_{\iota,K})$, a norm representation function may modulate incoming information only when the information is incompatible with an agent's beliefs. Thus, if a sentence φ is consistent with a belief set K, then the beliefs warranted for acceptance by the underlying norm should not go beyond the information from $\mathrm{Cn}(\{\varphi\})$. In conjunction with postulates $(*3)$ and $(*4)$, this means that revision by a sentence compatible with an agent's beliefs proceeds purely by expansion.

The following postulate corresponds to postulate $(*5)$.

$(\boxast\ 5)$ $\quad$ If $\mathrm{Cn}(\{\varphi\}) \neq \mathrm{For}(\mathcal{L})$, then $K \boxast \varphi \neq \mathrm{For}(\mathcal{L})$.

A relative of postulate $(\boxast 5)$, the next mixed postulate demands that if the set of beliefs an underlying norm warrants for acceptance in the revision of a belief set K

[33] Of course, it would have been sufficient to specify that $*$ satisfies only postulates $(*2)$ and $(*6)$, without specifying that $*$ satisfies $(*1)$.

by sentence φ is consistent, then the revision of K by φ ought to be consistent as well.

($*[\boxast]5$) If $K\boxast\varphi \neq \text{For}(\mathcal{L})$, then $K * \varphi \neq \text{For}(\mathcal{L})$. (*Norm-Inclusive Consistency*)

It is an easy matter to check that if a -inclusive revision function $*$ over K satisfies postulate ($*5$), then the norm representation function $\boxast$ satisfies postulate ($\boxast$ 5) while the norm-inclusive revision model $(*, \boxast)$ satisfies postulate ($*[\boxast]5$). Indeed, if $(*, \boxast)$ is a norm-inclusive revision model satisfying postulate ($*[\boxast]5$) such that the norm representation function $\boxast$ satisfies postulate ($\boxast$ 5), then the norm-inclusive revision function $*$ satisfies postulate ($*5$).

As with condition π_ι, the next postulate represents an interesting limit case.

($\boxast_\iota$) $K\boxast\varphi \subseteq \text{Cn}(\{\varphi\})$. (*Norm in Absentia*)

Postulate ($\boxast_\iota$) signifies that the underlying norm in question does not sanction the acceptance of beliefs beyond the input sentence. Thus, in effect, postulate ($\boxast_\iota$) expresses the fact that the underlying norm, even if it is operative, does not have any real influence on belief formation.

We finally turn to the central postulate of this section.

($*[\boxast]\text{R}_\infty$) For every non-empty $I \subseteq \text{For}(\mathcal{L})$., (*Norm-Inclusive Revision Coherence*)
if $\bigcap_{\psi\in I} K\boxast\psi \subseteq K\boxast\varphi$, then $K * \varphi \subseteq \text{Cn}((\bigcup_{\psi\in I} K * \psi) \cup K\boxast\varphi)$.

This postulate, as it presently stands, appears to be quite complicated. As with condition $\gamma[\pi]\text{R}_\infty$, the reason for this is that we have not placed any substantial conditions upon norm representation functions. Nonetheless, the proof of the pudding is in the eating. Our goal here is to show how it is possible to accommodate the influence of social norms in belief revision. As we will see, postulate ($*[\boxast]\text{R}_\infty$) embodies the minimal commitments one must undertake for rational norm-inclusive belief revision.

8.5.2 Correspondence Theorems

In this section we will present correspondence theorems connecting postulates of norm-inclusive belief revision to coherence constraints of norm-conditional choice. As in Section 8.4.3, we begin with several definitions.

Definition 5.2 A *semantic norm-conditional choice model* is a norm-conditional choice model on choice space $(\mathcal{W}_\mathcal{L}, \mathcal{E}_\mathcal{L})$.

In Section 8.4.3, we reviewed the notion of a choice-based revision function. We now introduce the notion of a *norm-inclusive choice-based revision model*.

Definition 5.3 Let K be a belief set, and let (γ, π) be a semantic norm-conditional choice model. The *semantic norm-inclusive choice-based revision model* $(*, \boxast)$ *over* K *generated by* (γ, π) is defined by setting for every $\varphi \in \mathrm{For}(\mathcal{L})$,

$$K * \varphi := \begin{cases} \widehat{\gamma(\llbracket\varphi\rrbracket)} & \text{if } \gamma(\llbracket\varphi\rrbracket) \neq \emptyset \\ \mathrm{For}(\mathcal{L}) & \text{otherwise,} \end{cases}$$

and

$$K \boxast \varphi := \begin{cases} \widehat{\pi(\llbracket\varphi\rrbracket)} & \text{if } \pi(\llbracket\varphi\rrbracket) \neq \emptyset \\ \mathrm{For}(\mathcal{L}) & \text{otherwise.} \end{cases}$$

We say that (γ, π) *generates* $(*, \boxast)$ or that $(*, \boxast)$ is *generated by* (γ, π).

Observe that every norm-inclusive choice-based revision model $(*, \boxast)$ over K is such that $*$ satisfies postulates $(*1)$, $(*2)$, and $(*6)$, $\boxast$ satisfies postulates $(\boxast\, 1)$, $(\boxast\, 2)$, and $(\boxast\, 6)$, and $(*, \boxast)$ satisfies postulate $(*\boxast)$. Thus, a norm-inclusive choice-based revision model is indeed a norm-inclusive revision model. As we will see below, the converse holds as well.

The intuitive interpretation of $K * \varphi$ in this context is similar to that of the usual interpretation: An agent believes a sentence ψ in the revision of K by φ just in case ψ is true in all the most 'plausible' π-permissible worlds in which φ is true.

We now introduce a final coherence constraint for norm-conditional choice models. As with conditions $(\mathrm{F1}_B)$ and $(\mathrm{F2}_B)$, this coherence constraint is relative to a set B of options.

$(\pi_{\iota,B})$ For every $S \in \mathcal{S}$, if $S \cap B \neq \emptyset$, then $S \subseteq \pi(S)$. (*Norm in Absentia with respect to B*)

Shortly we will see that this condition corresponds to postulate $(\boxast_{\iota,B})$ of norm-inclusive belief revision.

We offer a final definition.

Definition 5.4 Let K be a belief set. We call $(*, \boxast)$ a (*regular, norm-conditional regular, rational, G-rational*) *norm-inclusive choice-based revision model* over K if there is a (regular, π-conditional regular, π-rational, Gπ-rational) semantic norm-conditional choice model (γ, π) that generates $(*, \boxast)$.

We are now in a position to state and prove the long-awaited correspondence theorems.

Theorem 5.5 *Let K be a belief set. For every semantic norm-conditional choice model (γ, π) which satisfies*

$$\left\{\begin{array}{c} - \\ \mathrm{F1}_{[\![K]\!]}, - \\ \mathrm{F2}_{[\![K]\!]}, - \\ -, \mathrm{F1}_{[\![K]\!]} \\ -, \mathrm{F2}_{[\![K]\!]} \\ -, \pi_{\iota,[\![K]\!]} \\ \gamma_{>\emptyset}, - \\ -, \gamma_{>\emptyset} \\ \gamma[\pi]_{>\emptyset} \\ -, \pi_{\iota} \\ \alpha, - \\ \gamma^{*}, - \\ \gamma[\pi]\mathrm{R}_{\infty} \text{ and } \mathcal{L} \text{ is finite} \end{array}\right\}$$

the semantic norm-inclusive choice-based revision model $(*, \boxed{*})$ over K generated by (γ, π) satisfies

$$\left\{\begin{array}{c} - \\ (*4), - \\ (*3), - \\ -, (\boxed{*}4) \\ -, (\boxed{*}3) \\ -, (\boxed{*}_{\iota,K}) \\ (*5), - \\ -, (\boxed{*}5) \\ (*[\boxed{*}]5) \\ -, (\boxed{*}_{\iota}) \\ (*7), - \\ (*8r), - \\ (*[\boxed{*}]R_{\infty}) \end{array}\right\}$$

, respectively.

Proof Let (γ, π) be a semantic norm-conditional choice model, and let $(*, \boxed{*})$ be the semantic norm-inclusive choice-based revision model over K generated by (γ, π). We have seen above that $(*, \boxed{*})$ is indeed a norm-inclusive revision model over K. We only prove the implications for a subset of the conditions of the theorem, leaving the remaining implications to the reader.

$\gamma[\pi]_{>\emptyset} \Rightarrow (*[\boxed{*}]5)$: Suppose that (γ, π) satisfies condition $\gamma[\pi]_{>\emptyset}$. If $K\boxed{*}\varphi \neq \mathit{For}(\mathcal{L})$, then by definition $\pi([\![\varphi]\!]) \neq \emptyset$, so by condition $\gamma[\pi]_{>\emptyset}$, we have that $\gamma([\![\varphi]\!]) \neq \emptyset$, whence again by definition $K * \varphi \neq \mathrm{For}(\mathcal{L})$.

$\pi_{\iota} \Rightarrow \boxed{*}_{\iota}$: Suppose that π satisfies condition π_{ι}. Then if $\alpha \in K\boxed{*}\varphi$, $\pi([\![\varphi]\!]) \subseteq [\![\alpha]\!]$, so by condition π_{ι} it follows that $[\![\mathrm{Cn}(\{\varphi\})]\!] = [\![\varphi]\!] \subseteq [\![\alpha]\!]$, whereby $\alpha \in \mathrm{Cn}(\{\varphi\})$, as desired.

$\gamma[\pi]\mathrm{R}_{\infty} \Rightarrow (*[\boxed{*}]\mathrm{R}_{\infty})$: Suppose that (γ, π) satisfies condition $\gamma[\pi]\mathrm{R}_{\infty}$ and $\mathcal{L}$ is finite. Let $I \subseteq \mathrm{For}(\mathcal{L})$ be such that $I \neq \emptyset$. Suppose $\bigcap_{\psi\in I} K\boxed{*}\psi \subseteq K\boxed{*}\varphi$. Then since $\mathcal{L}$ is finite, it follows that $[\![K\boxed{*}\varphi]\!] \subseteq \left[\!\!\left[\bigcap_{\psi\in I} K\boxed{*}\psi\right]\!\!\right] = \bigcup_{\psi\in I} [\![K\boxed{*}\psi]\!]$. Again, since $\mathcal{L}$ is finite, π is complete, so $\pi([\![\varphi]\!]) \subseteq \bigcup_{\psi\in I} \pi([\![\psi]\!])$, whence it follows by condition $\gamma[\pi]\mathrm{R}_{\infty}$

that $\pi([\![\varphi]\!]) \cap (\bigcap_{\psi\in I}\gamma([\![\psi]\!])) \subseteq \gamma([\![\varphi]\!])$. Now because γ and π are complete,

$$\begin{aligned}
\left[\!\!\left[\mathrm{Cn}\left(K\boxast\varphi \cup \left(\bigcup_{\psi\in I} K*\psi\right)\right)\right]\!\!\right] &= \left[\!\!\left[K\boxast\varphi \cup \left(\bigcup_{\psi\in I} K*\psi\right)\right]\!\!\right] \\
&= [\![K\boxast\varphi]\!] \cap \left[\!\!\left[\bigcup_{\psi\in I} K*\psi\right]\!\!\right] \\
&= [\![K\boxast\varphi]\!] \cap \left(\bigcap_{\psi\in I} [\![K*\psi]\!]\right) \\
&=\pi([\![\varphi]\!]) \cap \left(\bigcap_{\psi\in I} \gamma\ ([\![\psi]\!])\right) \subseteq \gamma([\![\varphi]\!]) \\
&= [\![K*\varphi]\!]\,.
\end{aligned}$$

It follows that $K*\varphi \subseteq \mathrm{Cn}((\bigcup_{\psi\in I} K*\psi)\cup K\boxast\varphi)$, as desired.

□

Theorem 5.6 *Every norm-inclusive revision model* $(*,\boxast)$ *over a belief set K which satisfies* $\left\{\begin{array}{c} - \\ (*3),- \\ (*4),- \\ -,(\boxast 3) \\ -,(\boxast 4) \\ -,(\boxast_{\iota,K}) \\ (*5),- \\ -,(\boxast 5) \\ (*[\boxast]5) \\ -,(\boxast_\iota) \\ (*7),- \\ (*8r),- \\ (*[\boxast]R_\infty) \end{array}\right\}$ *can be represented as the semantic norm-inclusive choice-based revision model over K generated by a semantic norm-conditional choice model* (γ,π) *which satisfies* $\left\{\begin{array}{c} - \\ \mathrm{F2}_{[\![K]\!]},- \\ \mathrm{F1}_{[\![K]\!]},- \\ -,\mathrm{F2}_{[\![K]\!]} \\ -,\mathrm{F1}_{[\![K]\!]} \\ -,\pi_{\iota,[\![K]\!]} \\ \gamma_{>\emptyset},- \\ -,\gamma_{>\emptyset} \\ \gamma[\pi]_{>\emptyset} \\ -,\pi_\iota \\ \alpha,- \\ \gamma^*,- \\ \gamma[\pi]R_\infty \end{array}\right\}$, *respectively.*

Proof Let $(*,\boxast)$ be a norm-inclusive revision model over a belief set K. We define a semantic norm-conditional choice model (γ,π) by setting for every $\varphi \in \mathrm{For}(\mathcal{L})$, $\gamma([\![\varphi]\!]) := [\![K*\varphi]\!]$ and $\pi([\![\varphi]\!]) := [\![K\boxast\varphi]\!]$. Because by definition $*$ satisfies postulate $(*1)$, $(*2)$, and $(*6)$, $\boxast$ satisfies postulates $(\boxast\ 1)$, $(\boxast\ 2)$, and $(\boxast\ 6)$, and $(*, \boxast)$ satisfies postulate $(*[\boxast])$, clearly (γ, π) is a semantic norm-conditional choice model.

We must first show that (γ, π) generates $(*, \boxast)$. Let $\varphi \in \mathrm{For}(\mathcal{L})$. On the one hand, if $\gamma([\![\varphi]\!]) = \emptyset$, then $[\![K * \varphi]\!] = \emptyset$, so $K * \varphi = \mathrm{For}(\mathcal{L})$. On the other hand, if $\gamma([\![\varphi]\!]) \neq \emptyset$, then $K * \varphi = \widehat{[\![K * \varphi]\!]} = \widehat{\gamma([\![\varphi]\!])}$. We have thereby shown that

$$K * \varphi = \begin{cases} \widehat{\gamma([\![\varphi]\!])} & \text{if } \gamma([\![\varphi]\!]) \neq \emptyset \\ \mathrm{For}(\mathcal{L}) & \text{otherwise.} \end{cases}$$

A similar argument shows that

$$K \boxast \varphi = \begin{cases} \widehat{\pi([\![\varphi]\!])} & \text{if } \pi([\![\varphi]\!]) \neq \emptyset \\ \mathrm{For}(\mathcal{L}) & \text{otherwise.} \end{cases}$$

Hence, (γ, π) generates $(*, \boxast)$. We now turn to prove the implications of the theorem. As before, we only prove the implications for a subset of the postulates of the theorem.

$(*[\boxast]5) \Rightarrow \gamma[\pi]_{>\emptyset}$: Suppose that $(*, \boxast)$ satisfies postulate $(*[\boxast]5)$. If $\pi([\![\varphi]\!]) \neq \emptyset$, then $[\![K \boxast \varphi]\!] \neq \emptyset$ and so $K \boxast \varphi \neq \mathrm{For}(\mathcal{L})$, whence by postulate $(*[\boxast]5)$ it follows that $K * \varphi \neq \mathrm{For}(\mathcal{L})$ and therefore $\gamma([\![\varphi]\!]) = [\![K * \varphi]\!] \neq \emptyset$, as desired.

$(\boxast_\iota) \Rightarrow \pi_\iota$: Suppose that $\boxast$ satisfies postulate $\boxast_\iota$. Then for every $\varphi \in \mathrm{For}(\mathcal{L})$, $K \boxast \varphi \subseteq \mathrm{Cn}(\{\varphi\})$ and so $[\![\varphi]\!] = [\![\mathrm{Cn}(\{\varphi\})]\!] \subseteq [\![K \boxast \varphi]\!] = \pi([\![\varphi]\!])$.

$(*[\boxast]\mathrm{R}_\infty) \Rightarrow \gamma[\pi]\mathrm{R}_\infty$: Suppose that $(*, \boxast)$ satisfies postulate $(*[\boxast]\mathrm{R}_\infty)$. Let $I \subseteq \mathcal{E}_{\mathcal{L}}$ be such that $I \neq \emptyset$. Suppose that $\pi([\![\varphi]\!]) \subseteq \bigcup_{[\![\psi]\!] \in I} \pi([\![\psi]\!])$. Then $[\![K \boxast \varphi]\!] \subseteq \bigcup_{[\![\psi]\!] \in I} [\![K \boxast \psi]\!] \subseteq \left[\!\!\left[\bigcap_{[\![\psi]\!] \in I} K \boxast \psi\right]\!\!\right]$, whereby $\bigcap_{[\![\psi]\!] \in I} K \boxast \psi \subseteq K \boxast \varphi$. It follows by postulate $(*[\boxast]\mathrm{R}_\infty)$ that $K * \varphi \subseteq \mathrm{Cn}((\bigcup_{[\![\psi]\!] \in I} K * \psi \cup K \boxast \varphi)$. Then

$$\begin{aligned} \pi([\![\varphi]\!]) \cap \left(\bigcap_{[\![\psi]\!] \in I} \gamma\left([\![\psi]\!]\right) \right) &= [\![K \boxast \varphi]\!] \cap \left(\bigcap_{[\![\psi]\!] \in I} [\![K * \psi]\!] \right) \\ &= [\![K \boxast \varphi]\!] \cap \left[\!\!\left[\bigcup_{[\![\psi]\!] \in I} K * \psi \right]\!\!\right] \end{aligned}$$

$$= \left[\!\!\left[K\boxast\varphi \cup \left(\bigcup_{[\![\psi]\!]\in I} K * \psi \right) \right]\!\!\right]$$

$$= \left[\!\!\left[\mathrm{Cn}\left(K\boxast\varphi \cup \left(\bigcup_{[\![\psi]\!]\in I} K * \psi \right) \right) \right]\!\!\right] \subseteq [\![K * \varphi]\!] = \gamma([\![\varphi]\!]).$$

That is, $\pi([\![\varphi]\!]) \cap (\bigcap_{[\![\psi]\!]\in I} \gamma([\![\psi]\!])) \subseteq \gamma([\![\varphi]\!])$, as desired. □

In light of the results of Section 8.3.2, we have the following corollary.

Corollary 5.7 *Let $\mathcal{L}$ be finite, and let K be a belief set.*

(i) A pair $(*, \boxast)$ is a rational norm-inclusive choice-based revision model over K *if and only if it is a norm-inclusive revision model satisfying postulate $(*[\boxast]R_\infty)$.*
(ii) A pair $(*, \boxast)$ is a regular G-rational norm-inclusive choice-based revision model over K *if and only if it is a norm-inclusive revision model satisfying postulates $(*5)$ and $(*[\boxast]R_\infty)$.*
(iii) A pair $(*, \boxast)$ is a norm-conditional regular G-rational norm-inclusive choice-based revision model over K *if and only if it is a norm-inclusive revision model satisfying postulates $(*[\boxast]5)$ and $(*[\boxast]R_\infty)$.*

8.6 Examples and Discussion

The modular nature of the correspondence results of the previous section afford applicability to a variety of theories of belief revision. A theorist of belief revision can utilize our semantic representation of norm-inclusive belief revision to accommodate the influence of social norms in belief formation. Indeed, because our framework extends the classical framework in which belief revision is studied, a theorist can utilize our framework even in cases for which social norms do not have any real influence on belief formation.[34] Of course, a significant benefit of adopting our framework is that it offers a way to study belief revision with an eye toward the possibility that social norms sway belief formation.

The minimal conditions required for *rational* norm-inclusive belief revision are embodied in Corollary 5.7. We accordingly take the postulates comprising this corollary to represent the central conditions for norm-inclusive belief revision. As we mentioned earlier, we have not imposed any substantial conditions on the behavior of norm representation functions. We invite belief revision theorists to investigate what conditions can be plausibly imposed on norm representation functions.

As the reader may suspect, the above correspondence results translate naturally to correspondence results for what might be called *norm-inclusive non-monotonic*

[34] When social norms do not have any real on influence belief formation, postulate $\boxast_\iota$ is satisfied.

reasoning. Thus, our correspondence results also afford applicability to a variety of theories of non-monotonic reasoning. We take this to be a virtue of the modular nature of our results.

We now illustrate our theory at work. We continue with Example 8.4.4.2 of Section 8.4.4.

8.6.1 Example

Recall the analysis of Example 8.4.4.2 in terms of possible worlds. We have five relevant worlds:[35]

$$\begin{aligned} w_1 &= (\neg a, \neg b, \neg c, d, \neg e)\\ w_2 &= (\neg a, b, \neg c, \neg d, \neg e)\\ w_3 &= (a, \neg b, \neg c, \neg d, \neg e)\\ w_4 &= (\neg a, \neg b, c, \neg d, \neg e)\\ w_5 &= (\neg a, \neg b, \neg c, \neg d, e). \end{aligned}$$

The ordering $\geq$ over possible worlds is such that $w_1 > w_2$, $w_2 > w_3$ and $w_2 > w_4$, $w_3 \sim w_4$, and $w_3 > w_5$ and $w_4 > w_5$ (all other worlds are dominated by w_1, w_2, w_3, w_4, and w_5, and $w_i \geq w_j$ for $i \geq j$).

We may first inquire about the contents of $K \boxast (a \vee b \vee e)$, where K is the theory that corresponds to w_1. Here we have a norm related to the promotion of underrepresented demographic groups. According to the norm, w_2 is unfeasible, while two worlds will remain feasible, w_3 and w_5. So in terms of a norm representation function, if we take $\pi([\![a \vee b \vee e]\!]) = \{w_3, w_5\}$, we will have that $K \boxast (a \vee b \vee e)$ is the intersection of w_3 and w_5—namely, $\mathrm{Cn}(\{a \vee e, \neg(a \wedge e), \neg b, \neg c, \neg d\})$.

We now inquire about the contents of $K * (a \vee b \vee e)$. In this case the idea is to pick the worlds that are optimal among the worlds that are normatively feasible. In terms of a π-conditional semantic selection function γ, we have $\gamma([\![a \vee b \vee e]\!]) = \{w \in \pi([\![a \vee b \vee e]\!]) : w \geq w' \text{ for all } w' \in \pi([\![a \vee b \vee e]\!])\} = \{w_3\}$. So the theory that corresponds to w_3 gives us $K * (a \vee b \vee e)$.

We conclude this section with an intriguing variant of the example we presented above.

8.6.2 Example

Consider Example 8.4.4.2 of Section 8.4.4, but suppose instead that Tom is initially in *suspense* about who will be hired. Now rerun the scenarios considered in Example 8.4.4.2. The example so constructed preserves the format of Example 8.4.4.2, but now the information given by the Dean is *compatible* with Tom's initial beliefs. As

[35] As before, we adopt the notational convention that $(a, \neg b, \neg c, \neg d, \neg e)$ denotes the possible world for which a is true and the rest of the items – b, c, e, d – are false.

with the previous examples, this example illustrates a violation of postulate ($*7$) (among other things). But here another postulate is violated. In the first scenario, for example, postulate ($*3$) is violated. Although the information Tom learns is compatible with his view, he does not change his view by expansion. Yet in the previous examples the changes in each scenario are AGM-permissible changes insofar as they satisfy postulate ($*3$).

One possible reaction to this variant of Example 8.4.4.2 is to say that Rott's examples and some of its possible variations do not *really* reveal violations of the AGM postulates but violations of inductively extended versions of these postulates (this seems to be Levi's position regarding the original examples—see Levi, 1996, for a precise definition of inductive expansion). So in the case of the change with ($a \vee b \vee e$), the agent first accepts the disjunction in a way compatible with AGM and then jumps to conclusions using some inductive method.

Our theory also offers an explanation of these examples. In the first scenario, for example, a norm related to the promotion of unrepresented minorities is seen to be active and so two salient worlds remain feasible (the world where Adams is selected and nobody else is selected and the world where Earl is selected and nobody else is selected). One then optimizes over these feasible worlds, ultimately selecting Adam's world (w_3). So as long as we take the previous example seriously, perhaps we have no reason to endorse postulate ($*3$) in a theory of norm-inclusive belief change. We do not, nevertheless, exclude the possibility that there could be different applications of our formalism where $*$ obeys all the basic AGM postulates.

The main issue analyzed in this paper (how to extend rationalizability in order to deal with cases where social norms are relevant) is independent of the considerations about the role of postulate ($*3$). Our representation results and our correspondence results are modular, permitting a formal separation of these issues.

8.7 Conclusion and Future Work

The classical theory of rational choice deriving from the work of mathematical economists such as Arrow, Richter, Samuelson, and Sen has important connections with the theory of belief change initiated by the seminal work of Alchourrón, et al. (1985). Rott has articulated the formal and conceptual consequences of this connection in his recent (Rott, 2001). Usually when this connection is studied the theory of choice is taken as the more primitive and secure theory, from which several consequences for the theory of belief revision are drawn. But rarely are results in the theory of belief revision applied to produce novel results in the theory of choice.

One of the first theorems offered in our article is nevertheless of this type (see Theorem 2.6). Indeed, in Section 8.2.2 we undertook the traditional problem of stating necessary and sufficient conditions for G-rationalizability, whereupon we furnished a complete characterization of G-rationalizability in terms of a coherence constraint (viz., condition $\gamma\mathrm{R}_\infty$)[36] that is a generalization of a condition considered

[36] Condition $\gamma_{1>\emptyset}$ should be imposed as well, but the central condition in the result is $\gamma\mathrm{R}_\infty$.

by AGM in 1985. This condition permits a characterization of G-rationality that stands independently of constraints on the underlying domain. Consequently, this condition permits an extension of the interesting program concerned with the foundations of rational choice, as investigated by mathematical economists such as Richter (1966, 1971).

The usual restrictions on domains required by traditional results for rationalizability (of the sort Sen 1971, offers) are incompatible with the empirical applicability of the theory of rational choice, something many economists nevertheless consider desirable. Thus, only a theory of rational choice formulated in terms of what Suzumura calls *general domains* (see Suzumura 1983, p. 17) should be considered acceptable for an empirically viable theory of choice. Our characterization of G-rationalizability offers a result of this type, with $\gamma\mathrm{R}_\infty$ playing an essential role.

The first half of this article focused on a problem for the standard theory of rational choice, a problem Amartya Sen has given careful attention in a series of important articles. The problem is associated with the role of social norms in choice. Sen has presented various examples which illustrate that the influence of social norms in choice threatens a central idea in the theory of rational choice—that choice functions are *rationalizable*. In particular, Sen has offered several examples which represent violations of coherence constraints such as condition α in cases where norms are operative. The central challenge prompted by Sen's examples (which he left unresolved in his articles on this issue) is whether it is possible to extend the theory of choice to accommodate the role of social norms in rational choice.

The first step needed to extend the existing theory of choice is to develop a formal framework that gives footing to social norms. There is a rich literature focusing on the role of social norms in choice from the seminal paper by Elster (1989b) to the more recent and ambitious work by Cristina Bicchieri (2006). Much of this work focuses on providing *rational reconstructions* of social norms (this is so especially in the philosophical literature). As an illustration, in the case of Bicchieri the idea is to provide epistemic conditions (in terms of expectations and conditional expectations) which are necessary and sufficient for the very existence of social norms. To take another example, Elster considers whether norms exist to promote self-interest (or common interest or genetic fitness).

The work on social norms that derives from the investigations of Sen circumvents some of the aforementioned philosophical problems by taking social norms for granted and representing them in terms of constraints on feasibility. We followed Sen here, utilizing a *permissibility* operator π that when applied to a menu of options returns those alternatives which are permitted for choice by the underlying norm. We have not imposed substantial constraints on the permissibility operator. Some of the questions which philosophers ask about norms can be reconstructed in terms of additional constraints on π. We leave this for future work, and we follow the tradition inaugurated by Sen and continued by Bossert and Suzumura, which uses a fairly unconstrained permissibility operator.

We considered the natural idea about how to extend the notion of rationalizability to cases where underlying norms influence choice. In this context, we called a choice function γ π-rationalizable if there is a binary relation R such that:

$$\gamma(S) = G(\pi(S), R) := \{x \in \pi(S) : xRy \text{ for all } y \in \pi(S)\}.$$

The challenge then was to characterize functionally this extended notion of rationalizabillity. The work done on rationalizability in Section 8.2.2 is helpful here. In fact, there is a natural variant of condition $\gamma\mathrm{R}_\infty$ that does the job (viz., condition $\gamma[\pi]\mathrm{R}_\infty$). The functional characterization holds with respect to general domains.

Bossert and Suzumura (2007) also offer a characterization of this extended notion of rationalizability. In fact, their characterization also holds for general domains, but the condition that they propose implements a notion of revealed preference in its formulation. This type of formulation seems less direct than a condition formulated purely in terms of the choice operator, such as our $\gamma[\pi]\mathrm{R}_\infty$. Indeed, our condition has proven to be useful for the epistemological application we discussed in the second part of this chapter (while Bossert and Suzumura's condition seems difficult to apply in this context).

The second half of this chapter applied the results of the first part to epistemology. First we offered some examples showing that social norms also play a significant role in undermining crucial principles of belief revision which guarantee that belief revision functions are *relational*. Rott (2004) has offered examples of this kind, although these examples do not turn on the role of social norms but on a more general phenomenon that Sen calls *menu dependence*. In any case, Rott offer examples but he does propose a solution. To be sure, he writes, 'We have identified a formidable problem, but we havenOt been able to offer an acceptable solution for it. But problems there are, and creating awareness of problems is one of the important tasks of philosophy' (Rott 2004, p. 238).

One of the main goals of the second half of this chapter was to offer a solution to the sort of counterexamples we presented. Our solution can be extended to cover counterexamples of the sort offered by Rott; we intend to deal with the general issue of menu dependence in a companion article. The idea behind our solution is to use the generalized notion of rationalizability in the area of belief change. Thus, we have here as well a permissibility function π and the usual selection function γ familiar to students of belief change since its use in Alchourrón et al. (1985). Yet corresponding to these mechanisms are two components which comprise belief change, $\boxast$ and $*$, which can be defined in terms of the corresponding selection functions:

$$K * \varphi := \begin{cases} \widehat{\gamma(\llbracket\varphi\rrbracket)} & \text{if } \gamma(\llbracket\varphi\rrbracket) \neq \emptyset \\ \mathrm{For}(\mathcal{L}) & \text{otherwise,} \end{cases}$$

and

$$K \boxast \varphi := \begin{cases} \widehat{\pi(\llbracket\varphi\rrbracket)} & \text{if } \pi(\llbracket\varphi\rrbracket) \neq \emptyset \\ \mathrm{For}(\mathcal{L}) & \text{otherwise.} \end{cases}$$

We verify that when the language is finite, condition $\gamma[\pi]\mathrm{R}_\infty$ is mappable to a novel mixed postulate combining the $*$ and $\boxast$ operators (viz., postulate $*[\boxast]\mathrm{R}_\infty$).

This yields an extension of the constraint that belief revision is relational to the case where social norms are operative. Finally we offer a complete characterization for the two change operators.

The mappings linking selection functions and belief revision operators provide heuristic insight about the shape of desired postulates in the area of belief change. It seems that the same method can be used to deal with the general issue of menu dependence, although this phenomenon remains a bit more elusive. Social norms offer a mechanism that explains shifts in feasibility, yet other cases of menu dependence involve shifts which are purely epistemic and more difficult to explain. Nonetheless, perhaps by interpreting the underlying formalism in a different way, the theory offered here can be lent a broader interpretation that covers other cases of menu dependence.

The general moral to be taken from this article and from recent work by Bossert and Suzumura is that the theory of rational choice can be extended fruitfully in order to cope with counterexamples involving social norms. By the same token, the origin of the counterexamples to principles of belief formation offered by Rott (2004) resides in the same phenomenon that motivated Sen's counterexamples against fundamental conditions in the theory of choice. As we have seen, social norms can be a source of counterexamples to principles of belief formation. A solution to problems in the theory of choice inspired a solution in the realm of belief change.

References

Alchourrón, Carlos E., and David Makinson. 1985. On the logic of theory change: safe contraction. *Studia Logica* 44(4):405–422.

Alchourrón, Carlos E., Peter Gärdenfors, and David Makinson. 1985. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic* 50(2): 510–530.

Arló-Costa, Horacio. 1995. Epistemic conditionals, snakes and stars, conditionals: From philosophy to computer science. In *Studies in logic and computation*, eds. A. Herzig, L. Farinas del Cerro, and G. Crocco, vol. 5, Oxford: Oxford University Press, 193–239.

Arló-Costa, Horacio. 2006. Rationality and value: The epistemological role of interdeterminate and agent-dependent values. *Philosophical Studies* 128(1):7–48.

Arrow, Kenneth J. 1951. *Social choice and individual values*. 1 edn. New York, NY: Wiley.

Arrow, Kenneth J. 1959. Rational choice functions and orderings. *Economica* 26:121–127.

Bicchieri, Cristina. 2006. *The grammar of society: The nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Bossert, W., and K. Suzumura. 2007. Social norms and rationality of choice, Preprint, Départment de Sciences Economiques, Université de Montréal, August 2007.

Elster, Jon. 1989a. *Nuts and bolts for the social sciences*. Cambridge: Cambridge University Press.

Elster, Jon. 1989b. Social norms and economic theory. *The Journal of Economic Perspectives* 3(4):99–117.

Gärdenfors, Peter. 1979. Conditionals and changes of belief. In *Logic and epistemology of scientific change* eds. I. Niiniluoto and R. Tuomela, 381–404. *Acta Philosophica Fennica*, vol. 30. Amsterdam: North-Holland Publishing, 1979.

Gärdenfors, Peter. 1988. *Knowledge in flux. modeling the dynamics of epistemic states*. Cambridge, MA: MIT Press.

Gärdenfors, Peter. (ed.) 1992. *Belief revision: An introduction, belief revision*. Cambridge: Cambridge University Press, 1–28.

Gärdenfors, Peter., and David Makinson. 1994. Nonmonotonic inference based on expectations. *Artificial Intelligence* 65(2):197–245.

Gärdenfors, Peter., and Hans Rott. 1995. Belief revision. In *Handbook of logic in artificial intelligence and logic programming*, eds. Dov M. Gabbay, C.J. Hogger, and J.A. Robinson, vol. 4, 35–132. Oxford: Oxford University Press.

Grove, Adam. 1988. Two modellings for theory change. *Journal of Philosophical Logic* 17(2):157–170.

Hansson, B., 1968. Choice structures and preference relations. *Synthese* 18(4):443–458.

Hansson, Sven Ove., 1999. *A textbook of belief dynamics: Theory change and database updating*. Dordrecht: Kluwer.

Herzberger, Hans G. 1973. Ordinal preference and rational choice. *Econometrica* 41(2):187–237.

Kraus, Sarit, Daniel Lehmann, and Menachem Magidor. (1990) Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence* 44(1–2):167–207.

Lehmann, Daniel, and Menachem Magidor. 1992. What does a conditional knowledge base entail? *Artificial Intelligence* 55(1):1–60.

Levi, Isaac. 1996. *For the sake of the argument*. Cambridge: Cambridge University Press.

Levi, Isaac. 2004a. Inclusive rationality: A review of Amartya Sen: rationality and freedom. *The Journal of Philosophy* 101(5):255–276.

Levi, Isaac. 2004b. *Mild contraction*. Oxford: Oxford University Press.

Makinson, David. 1989. *General theory of cumulative inference*, Proceedings of the 2nd International Workshop on Non-Monotonic Reasoning, 1–18. Lecture Notes in Computer Science, vol. 346, Berlin: Springer.

Makinson, David., and Peter Gärdenfors. 1991. Relations between the logic of theory change and nonmonotonic logic. In *The logic of theory change*, eds. André Fuhrmann and Michael Morreau, 185–205. Lecture Notes in Computer Science, Berlin: Springer.

Moulin, H. 1985. Choice functions over a finite set: A summary. *Social Choice and Welfare* 2: 147–160.

Olsson, Erik J. 2003. Belief revision, rational choice and the unity of reason. *Studia Logica* 73(2):219–240.

Pedersen, Arthur Paul. 2008. Rational choice and belief revision: An essay in formal epistemology, Master's thesis, Carnegie Mellon University, Department of Philosophy.

Pearl, Judea., and M. Goldszmidt. 1996. Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence* 84(1–2):57–112.

Ray, Paramesh. 1973. Independence of irrelevant alternatives. *Econometrica* 41(5):987–991.

Richter, M.K. 1966. Revealed preference theory. *Econometrica* 34(3):635–645.

Richter, M.K. 1971. Rational choice. In *Preferences, utility, and demand*, eds. J. Chipman, L. Hurwicz, M. Richter, and H. Sonnenshein, 29–58. San Diego, CA: Harcourt Brace Javanovich.

Rott, Hans. 1993. Belief contraction in the context of the general theory of rational choice. *Journal of Symbolic Logic* 58(4):1426–1450.

Rott, Hans. 2001. *Change, choice and inference: A study of belief revision and nonmonotonic reasoning*. Oxford: Oxford University Press.

Rott, Hans. 2004. A counterexample to six fundamental principles of belief formation. *Synthese* 139(2):225–240.

Rott, Hans., and Maurice Pagnucco. 2000. Severe withdrawal (and recovery). *Journal of Philosophical Logic* 28(5):501–547.

Samuelson, Paul A. 1938. A note on the pure theory of consumers. *Economica* 5:61–71.

Samuelson, Paul A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Sen, Amartya. 1970. *Collective choice and social welfare*. San Francisco, CA: Holden-Day.

Sen, Amartya. 1971. Choice functions and revealed preference. *Review of Economic Studies* 38:307–317.

Sen, Amartya. 1977. Social choice theory: A re-examination. *Econometrica* 45(1):53–89.

Sen, Amartya. 1993. Internal consistency of choice. *Econometrica* 61(3):495–521.

Sen, Amartya. 1996. Is the idea of purely internal consistency of choice Bizarre? In *Language, World and Reality*, eds. J.E.J Altham and T.R. Harrison, 19–31. A Festschrift for Bernard Williams, Cambridge: Cambridge University Press.

Sen, Amartya. 1997. Maximization and the act of choice, *Econometrica* 65(4):745–779.

Spohn, Wolfgang. 2009. A survey in ranking theory, In *Degrees of belief: An anthology*, eds. Franz Huber, Cristoph Schmidt-Petri, 185–229. Oxford: Oxford University Press.

Suzumura, Kotaro. 1976. Rational choice and revealed preference, *Review of Economic Studies* 43(1):149–158.

Suzumura, Kotaro. 1983. *Rational choice, collective decisions, and social welfare*. Cambridge: Cambridge University Press.

Chapter 9
Rational Belief Changes for Collective Agents

David Westlund

9.1 Introduction

Belief revision is a model for how an ideal agent should change his or her beliefs. The main intuition behind the belief revision model is a kind of epistemic conservatism. When you stop believing in something you should keep as many other beliefs as possible, and when you start believing in something you should not start to believe more than necessary. The formal model captures an intuition that Gärdenfors (1988, p. 8) calls *minimal change.*

There are connections between the intuition behind belief revision and philosophy of science. During the periods that Kuhn (1962) calls *normal science* changes within a discipline are conservative. An extra valuable part of the discipline's collection of theories – the paradigm – is kept intact, while small adjustments are made to the rest of the theories to keep the discipline in line with experimental findings.

Belief revision is also connected to changes to scientific as well as other kinds of theories. How a theory is changed is bounded by logic. Both to extravagant extensions and to expensive losses of information should be avoided during theory changes. The intuition of conservatism that is at the bottom of belief revision suits theory revision just as well.

Consequently, philosophers within the belief revision field have used their models to reason about problems within philosophy of science. Gärdenfors (1988, p. 88) argues that what Kuhn calls paradigm shifts can be captured by changes in how scientists value information. The idea is that when a paradigm shift occurs, the scientist will change which of her beliefs she considers the most important.

Levi (1991, p. 65) instead argues against incommensurability by showing that all possible belief changes can be done as a series of adjustments, all complying to the rules set up by the belief revision model. If all changes are of this sort they all fit into the same conceptual framework and thus they are commensurable.

D. Westlund (✉)
Fjärdingsvägen 29, 241 36 Eslöv, Sweden
e-mail: david@westlund.fm

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_9,

Even though there are connections between belief revision and philosophy of science, philosophers in the latter field have not taken much notice of former. Why is that? One possible reason is that science is a social conduct while belief revision is focused on single individuals. Scientists often work in teams, they argue about other scientists' theories, and they learn to use theories developed by earlier generations of scientists. Further, a theory as in the theory of evolution or the theory of gravity is a social entity that evolves over time. Theories do not depend on some specific person believing them and neither do they follow the development of one specific scientist's private theory. Kuhn's paradigms are social entities and the use of the revolution metaphor, the violent rising of a group against the power, shows the importance given to the collective in his work.

This social aspect of science is seldom discussed in depth by philosophers in the belief revision field. Gärdenfors (1988, p. 10) mentions that his model might be applicable to groups of individuals such as organizations, but does not make any attempt to defend or investigate this claim. Levi instead assumes that his model can be used for collectives as well as for individuals:

> I am inclined to think that social groups are sometimes agents. Sometimes they are not. If they are agents seeking to promote cognitive ends in fixing their beliefs, they are to be treated no differently than any other such agents (. . .). (Levi 1995, p. 621)

The connection between belief revision and normal science as well as the development of theories makes it interesting to study whether belief revision can be used as a model for collectives. To make such studies possible, it must be explicated what it means for a collective to believe something, and how the collective gets some kind of input that makes it change its beliefs. In this text it will be assumed that the beliefs and inputs of the collective are functions of the beliefs and inputs of its members. Given the assumption, one can study if individual belief revision agents give rise to a collective adhering to the same model of belief change. For belief revision to be applicable to collectives, this would have to be the case.

After a short introduction to the belief revision model an explication of collective beliefs and belief changes will be given. After that, the possibility of a collective following the rules of belief revision will be studied. The results will be mostly negative. The text will end with a conclusion and discussion of the results.

9.2 The Belief Revision Model

We will use the formal model presented in Alchourron et al. (1985) and Gärdenfors (1988).

Let **L** be language that includes classical propositional logic. An epistemic state **E** is represented by a pair $< \mathbf{K}, \leq >$. **K** is a set of sentences in the language **L** representing what the agent believes. **K** is closed under logical consequence – if α is implied by **K**, then $\alpha \in \mathbf{K}$. The second part of the epistemic state $\leq$ is an ordering over the sentences in the language **L**. This ordering determines the different sentences fate when there is a change of beliefs.

It is assumed that an agent's beliefs are in a stable state that will not change without any new information from the outside world. Since **K** is closed under logical consequence, deriving consequences from one's current beliefs will not result in any new beliefs. For there to be a change in **K**, the agent needs to get *epistemic input*. Informally, this could be anything that makes the agent change its mind, for example sensory input. Formally, it can be defined as three operators; sentences consistent with the current sentences in **K** are added with expansion (+), sentences within **K** are removed with contraction (–) and sentences inconsistent with the current belief set are incorporated with revision ($*$). Revision with α can be defined as a contraction of the negation of α followed by an expansion of α according to the Levi identity:

Levi identity: $\mathbf{K} * \alpha = (\mathbf{K} - \neg\, \alpha) + \alpha$

Because of this only contraction and expansion will be discussed in this text.

The expansion operator is assumed to fulfill the following rationality postulates from Gärdenfors (1988, pp. 49–51):

+1 $\mathbf{K} + \alpha$ is a belief set
+2 $\alpha \in \mathbf{K} + \alpha$
+3 $\mathbf{K} \subseteq \mathbf{K} + \alpha$
+4 If $\alpha \in \mathbf{K}$, then $\mathbf{K} + \alpha = \mathbf{K}$
+5 If $\mathbf{K_1} \subseteq \mathbf{K_2}$, then $\mathbf{K_1} + \alpha \subseteq \mathbf{K_2} + \alpha$
+6 **K** is the minimal belief set fulfilling $+1 - +5$

It has been shown (Gärdenfors 1988, p. 51) that $+1 - +6$ is equivalent to $\mathbf{K}+\alpha = \mathrm{Cn}(\mathbf{K} \cup \{\alpha\})$, where $\mathrm{Cn}(\mathbf{K}) =_{\mathrm{def}} \{\alpha\, |\mathbf{K} \vdash \alpha\}$.

Contraction is assumed to fulfill the following postulates:

-1 $\mathbf{K} - \alpha$ is a belief set
-2 $\mathbf{K} - \alpha \subseteq \mathbf{K}$
-3 If $a \notin \mathbf{K}$, then $\mathbf{K} - \alpha = \mathbf{K}$
-4 If α is not a tautology, then $\alpha \notin \mathbf{K} - \alpha$
-5 If $\alpha \in \mathbf{K}$, then $\mathbf{K} \subseteq (\mathbf{K} - \alpha) + \alpha$
-6 If $+\alpha \equiv \beta$, then $\mathbf{K} - \alpha = \mathbf{K} - \beta$ (Gärdenfors 1988, pp. 61–63)

The rationality postulates for contraction allows for many different outcomes of the operator. As an example, take the case of $\mathrm{Cn}(\{\alpha,\ \beta\}) - \alpha$. The obvious possibility is $\mathrm{Cn}(\{\beta\})$, but there are other possible outcomes also allowed by the postulates like $\mathrm{Cn}(\{\alpha \equiv \beta\})$ or $\mathrm{Cn}(\{\alpha \rightarrow \beta\})$. To choose between these possibilities, the epistemic entrenchment ordering is used. We will not go into the details of how this is done since it is a detail not needed for the results in the chapter.

9.3 Collective Agents

If an epistemic state is modeled after a single human, we will call that human an *individual agent*. If the state is modeled after groups of individuals, then it is a *collective agent*. Some examples of what could be considered a collective agent would the scientific community or parts of it such as the currently active particle physicists or the philosophy department. Informally, what we need for this text is any group of people such that we ascribe the group epistemic attitudes in the form of beliefs or theories. It will be assumed that the epistemic state of a collective agent is a function from the individual agents. The collective agent's epistemic state should depend only on its members' epistemic states, and nothing else.

9.4 Collection Functions

A collective agent's epistemic state will be represented in the same way as an individual agent's, as a pair $< \mathbf{K}, \ \leq >$ of a belief set and an epistemic entrenchment ordering. It will be assumed that the epistemic state of the collective depends on its members epistemic states. Let us use *F*() to denote a function from individual epistemic states to a collective epistemic state. We can now put up some restrictions on *F*(), where the first one is obvious.

> **Condition of epistemic dependency:** *F*() is a function from epistemic states $\mathbf{E_1}, \ldots \mathbf{E}_{\boldsymbol{n}}$ to an epistemic state $\mathbf{E_c}$.

Further, it will be assumed that what the collective believes only depends on what its members believes. Since beliefs are represented by belief sets, the belief set of a collective agent only depends on the belief set of its members. To express this formally, we need a function from many belief sets to one belief set. We will borrow terminology from computer science (Liberatore and Schaerf 1988), where such a function is called a merging function.

> **Definition, merging function:** A *merging function M*() is a function from n belief sets $\mathbf{K}_1, \ldots \mathbf{K}_{\boldsymbol{n}}$ to a belief set $\mathbf{K_c}$.

We can now formally capture the idea that the beliefs of the collective depends on its members beliefs.

> **Condition of belief dependency:** A function *F*() from epistemic states $< \mathbf{K_1}, \leq_1 >, \ldots, \ < \mathbf{K}_{\boldsymbol{n}}, \leq_n >$ to an epistemic state $< \mathbf{K_c}, \leq_c >$ fulfills the *belief dependency restriction* if and only if there is a merging function *M*() such that $\mathbf{K_c} = \mathbf{M}(\mathbf{K_1}, \ldots \mathbf{K}_{\boldsymbol{n}})$.

In the AGM model agents are supposed to be in reflective equilibrium, only changing their beliefs as a result of some input. The most obvious example of such input for an individual is sensory inputs such as hearing something or reading something, but there could be other possible sources for belief change than the senses, such as memories.

On the collective level it is harder to find a plausible source of change since collectives do not have senses or memories in the same way as individuals. Instead, it will be assumed that the epistemic input of the collective agent depends on the input of the individual agents. There is support for this assumption in ordinary language. For example, suppose that someone argues that the scientific community changed its beliefs after some experimental finding. What one then means is that all, or at least most, members of the scientific community changed their beliefs after the finding.

In this text, the only case that will be studied is when all individuals change their beliefs in the same way. In the AGM model, this would mean that if all individual agents expand or contract with a sentence, then the collective agent should expand or contract with the same sentence. We can express this more exact:

> **Condition of collective change:** Suppose that $F()$ is a function from epistemic states with belief sets $\mathbf{K_1} \ldots \mathbf{K}_n$ to an epistemic state with belief set $M(\mathbf{K_1}, \ldots, \mathbf{K}_n)$, and that o is one of the operators expansion or contraction. Then $F()$ fulfills the *condition of collective change* if and only if $M(\mathbf{K_1}\text{o}\alpha, \ldots,\ \mathbf{K}_n\text{o}\alpha) = M(\mathbf{K_1}, \ldots, \mathbf{K}_n)\ \text{o}\alpha$.

Note that the principle of collective change is a principle on the level of epistemic states, since if the operator is contraction $M(\mathbf{K_1} - \alpha, \ldots, \mathbf{K}_n - \alpha) = M(\mathbf{K_1}, \ldots, \mathbf{K}_n) - \alpha$ says something about the relation between the collective epistemic entrenchment ordering and the entrenchment ordering of its members. This is the only assumption that will be done about the epistemic entrenchment ordering of the collective agent in this text.[1]

These three conditions together give us that the epistemic state as well as the belief set of the collective only depends on the epistemic states and belief sets of its members, and that if all members of a collective change their beliefs in a certain way, then the collective changes its beliefs in the same way. A function fulfilling these three conditions will be called a collection function:

> **Definition, collection function:** A function $CF()$ is a collection function if and only if it fulfills the conditions of epistemic dependency, belief dependency and collective change.

[1] How to get a collective choice function from the choice functions of individuals is studied in the field of social choice (see Sen 1970). The problem studied there is in some ways similar to the subject of this article but there are some noticeable differences. First, in social choice many orderings are combined into one ordering. Here we combine unordered sets into one unordered set. Secondly, the problem discussed in this text have a dynamic aspect, it is the changes that are studied. There is no dynamic element in social choice. That being said, it could be worth studying if there is a connection between some of the results presented here and results from the studies of social choice.

We will assume that the epistemic state and epistemic input of a collective can be given by a collection function from its members' epistemic states and input. This assumption is central for the results in this text.

9.5 Limits on the Collective Agent's Belief Set

The belief dependency condition creates a connection between the beliefs of the collective and the beliefs of its members. Without the other conditions needed for a collection function this connection could be very weak. For example, consider the merging function $M()$ such that $M(\mathbf{K_1}, \ldots, \mathbf{K}_n) = \mathrm{Cn}(\{\ \})$ for any belief sets $\mathbf{K_1, \ldots, K}_n$. While it fulfills the belief dependency condition (it is a function from many belief sets to one belief set), the resulting belief set will always have only the logical tautologies as members. Obviously the connection between what the members of a collective believes and what the collective itself believes should be stronger. For example, it seems reasonable that if all members of a collective believes a sentence α then the collective agent also believes α.

The condition of collective change warrants such a limit. For example, the function $F()$ with the merging function $M()$ always giving $\mathrm{Cn}(\{\})$ as a result does not fulfill the condition, since it always gives the same belief set even after an expansion or contraction. Suppose that all members of a collective expands with the sentence β. From the condition of collective change it follows that the collective agent also should expand with the sentence β, so that $M(\mathbf{K_1} + \beta, \ldots, \mathbf{K}_n + \beta) = M(\mathbf{K_1}, \ldots, \mathbf{K}_n) + \beta$. In this case however, $M(\mathbf{K_1} + \beta, \ldots, \mathbf{K}_n + \beta)$ will be $\mathrm{Cn}(\{\})$ which is not allowed by the condition of collective change together with AGM postulate +2.

From the conditions on a collection function a limit on how much and how little the collective is allowed to believe can be inferred. We will start with the lower limit, how little a collective agent can believe.

Observation 1: Let CF() be a collection function with merging function $M()$. From the condition of collective change together with AGM postulates -3 and +4, $\mathbf{K_1} \cap \ldots \cap \mathbf{K}_n \subseteq M(\mathbf{K_1}, \ldots, \mathbf{K}_n)$.

***Proof 1*:** Suppose that α is any sentence such that $\alpha \in \mathbf{K_1} \cap \ldots \cap \mathbf{K_n}$. It can be concluded that $\alpha \in \mathbf{K}_m$ for any m such that $1 \leq m \leq n$. From postulate +4 we get that $\mathbf{K}_m + \alpha = \mathbf{K}_m$ for any m.

The condition of collective change gives us that $M(\mathbf{K_1} + \alpha, \ldots, \mathbf{K_n} + \alpha) = M(\mathbf{K_1}, \ldots, \mathbf{K_n}) + \alpha$. Since $\mathbf{K}_m + \alpha = \mathbf{K}_m$ for any m, $M(\mathbf{K_1}, \ldots, \mathbf{K}_n) = M(\mathbf{K_1}, \ldots, \mathbf{K}_n) + \alpha$. From postulate +2 it follows that $\alpha \in M(\mathbf{K_1}, \ldots, \mathbf{K_n}) + \alpha$, so $\alpha \in M(\mathbf{K_1}, \ldots, \mathbf{K_n})$.

This lower limit says that the collective must believe as least everything that is believed by all its members. A collective can not believe less than the consensus of its members. Intuitively, this seems reasonable.

Observation 2: Let *CF*() be a collection function with a merging function *M*(). Then it follows from the condition of collective change together with AGM postulates −3 and −4 that $M(\mathbf{K_1}, \ldots, \mathbf{K}_n) \subseteq \mathbf{K_1} \cup \ldots \cup \mathbf{K}_n$.

***Proof 2*:** Take any sentence α such that $\alpha \notin \mathbf{K_1} \cup \ldots \cup \mathbf{K_n}$. Since $\alpha \notin \mathbf{K_1} \cup \ldots \cup \mathbf{K}_n$ it can be concluded that $\alpha \notin \mathbf{K}_m$ for all m such that $1 \leq m \leq n$. From postulate −3 and $\alpha \notin \mathbf{K}_m$ it follows that $\mathbf{K}_m - \alpha = \mathbf{K}_m$ for any m.

The condition of collective change gives us that $M(\mathbf{K_1} - \alpha, \ldots, \mathbf{K}_n - \alpha) = M(\mathbf{K_1}, \ldots, \mathbf{K}_n) - \alpha$. Since $\mathbf{K_m} - \alpha = \mathbf{K_m}$ for any m such that $1 \leq m \leq n$, $M(\mathbf{K_1}, \ldots, \mathbf{K}_n) = M(\mathbf{K_1}, \ldots \mathbf{K}_n) - \alpha$. From −4 it follows that $\alpha \notin M(\mathbf{K_1}, \ldots, \mathbf{K}_n) - \alpha$ and thus $\alpha \notin M(\mathbf{K_1}, \ldots, \mathbf{K}_n)$.

This upper limit guarantees that the collective does not believe too much. All the collective's beliefs must be believed by at least one of its members.

By using observation 2, we can get some further results. The first one concerns the consistency of the collective agent.

Observation 3: Let *CF*() be a collection function with a merging function *M*(). From observation 2 it follows that if all belief sets $\mathbf{K_1}, \ldots, \mathbf{K}_n$ are consistent, then $M(\mathbf{K_1}, \ldots, \mathbf{K}_n)$ is consistent.

***Proof 3*:** If all belief sets $\mathbf{K_1}, \ldots, \mathbf{K}_n$ are consistent, the sentence $\alpha \wedge \neg\alpha$ is not an element of any of the belief sets. It can consequently not be an element of $\mathbf{K_1} \cup \ldots \cup \mathbf{K}_n$.

Everything follows from an inconsistency and belief sets are closed under logical consequence, so $\alpha \wedge \neg\alpha$ is an element of all inconsistent belief sets. From observation 2 we get that $M(\mathbf{K_1}, \ldots, \mathbf{K}_n) \subseteq \mathbf{K_1} \cup \ldots \cup \mathbf{K_n}$. $M(\mathbf{K_1}, \ldots, \mathbf{K}_n)$ can then not have $\alpha \wedge \neg\alpha$ as an element, and thus not be inconsistent.

This result seems very reasonable. Consistency is often considered to be a minimal requirement for rationality. For the collective to be an agent worth studying, it should be consistent in most cases. Unfortunately, observation 2 has another less intuitive consequence.

Observation 4: Let *CF*() be a collection function with a merging function *M*(). From observation 2 it follows that there is at least one belief set $\mathbf{K}_m$ where 1 ≤ m ≤ n such that $M(\mathbf{K_1}, \ldots, \mathbf{K}_n) \subseteq \mathbf{K}_m$.

***Proof 4*:** Let us assume that there is no $\mathbf{K}_m$ such that $M(\mathbf{K_1}, \ldots, \mathbf{K}_n) \subseteq \mathbf{K}_m$. From the definition of a belief set we know that $M(\mathbf{K_1}, \ldots, \mathbf{K}_n)$ is logically closed. From observation 2 we know that any element of $M(\mathbf{K_1}, \ldots, \mathbf{K}_n)$ must be an element of at least one of $\mathbf{K_1}, \ldots, \mathbf{K}_n$. Further, from what we want to prove we can ignore all $\mathbf{K}_i$ that is a subset of any other $\mathbf{K}_j$ from $\mathbf{K_1} \cup \ldots \cup \mathbf{K}_n$ since if $M(\mathbf{K_1}, \ldots, \mathbf{K}_n) \subseteq \mathbf{K}_i$ then $M(\mathbf{K_1}, \ldots, \mathbf{K}_n) \subseteq \mathbf{K}_j$ as well.

Suppose that $M(\mathbf{K_1}, \ldots \mathbf{K}_n)$ is not a subset of any one of the belief sets $\mathbf{K_1}, \ldots, \mathbf{K}_n$. Then from our assumption there must be at least one $\alpha \in M(\mathbf{K_1}, \ldots, \mathbf{K}_n)$ such that $\alpha \in \mathbf{K}_i$ but $\alpha \notin \mathbf{K}_j$ for any j, and one $\beta \in M(\mathbf{K_1}, \ldots, \mathbf{K}_n)$

such that $\beta \notin \mathbf{K}_i$ but $\beta \in \mathbf{K}_j$ for any j. From closure, $\alpha \wedge \beta$ is an element of $M(\mathbf{K}_1, \ldots, \mathbf{K}_n)$. However, since both α and β are not members of any one $\mathbf{K}_k$ for any index k the sentence $\alpha \wedge \beta$ is not an element of $\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_n$. Together with observation 2 this gives us a contradiction. We can conclude that there is a $\mathbf{K}_m$ where $1 \leq m \leq n$ such that $M(\mathbf{K}_1, \ldots, \mathbf{K}_\mathbf{n}) \subseteq \mathbf{K}_m$.

9.6 Specific Merging Functions and Families of Merging Functions

We now have some general results concerning the merging function for any collection function. However, not all collection functions are interesting. Let us start with an example. From observation 4 we know that the belief set of the collective has a maximum where it is the same as one of its members' belief set. What if there is an epistemic dictatorship, so the collective belief set is exactly the same as one of its members' belief set in all situations?

Definition, epistemic dictator function: A merging function $M(\mathbf{K}_1, \ldots, \mathbf{K}_n)$ is an epistemic dictator function if and only if there is an index i such that $M(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathbf{K}_i$ irrespectively of the elements of $\mathbf{K}_1, \ldots, \mathbf{K}_n$.

The following result is obvious.

Observation 5: An epistemic dictator function can be a merging function for a collection function.

***Proof 5*:** We know that $M(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathbf{K}_i$ so $M(\mathbf{K}_1, \ldots, \mathbf{K}_n) \circ \alpha = \mathbf{K}_i \circ \alpha$ irrespectively of whether o is the expansion or contraction operator. Further, since this is irrespectively of what $\mathbf{K}_i$ is, it is also true that $M(\mathrm{K}_1 \circ \alpha, \ \ldots, \ \mathbf{K}_n \circ \alpha) = \mathbf{K}_i \circ \alpha$. Obviously, $M(\mathbf{K}_1 \circ \alpha, \ldots, \mathbf{K}_n \circ \alpha) = M(\mathbf{K}_1, \ldots, \mathbf{K}_n) \circ \alpha$ which is what the condition of collective change demands.

Epistemic dictatorship is however not an interesting case. It is even questionable if a collection function with such a merging function would create an epistemic state for a collective agent at all, since only one agent's beliefs contributes to what the collective believes. So let us move on to other possible merging functions.

Suppose that there is good faith in all the members of a collective, we know that they are all accountable. The members might be experts in different areas or it might be a group with some kind of epistemic division of labor such as a group of students working on a paper together. In such cases, conflicts between the members' beliefs might be rare or even nonexistent. It is then reasonable that the collective believes everything believed by any of its members, provided that there is no inconsistency.

Union when possible: A merging function $M()$ is a *union when possible* merging function if and only if $M(\mathbf{K}_1, \ldots, \mathbf{K}_n) = \mathrm{Cn}(\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_n)$ when $\mathbf{K}_1 \cup \ldots \cup \mathbf{K}_n$ is consistent.

From earlier observations we get the following result:

Observation 6: No merging function in the union when possible family can be a merging function for a collection function.

***Proof 6*:** Let us use $\mathbf{K_1} = \text{Cn}(\{\alpha\})$ and $\mathbf{K_2} = \text{Cn}(\{\beta\})$ where $\alpha \neq \beta$ as an example. We can conclude that $\alpha \wedge \beta \in \text{Cn}(\mathbf{K_1} \cup \mathbf{K_2})$. We can also conclude that $\alpha \wedge \beta \notin \mathbf{K_1}$ since $\beta \notin \mathbf{K_1}$ and that $\alpha \wedge \beta \notin \mathbf{K_2}$ since $\alpha \notin \mathbf{K_2}$. Obviously, $\alpha \wedge \beta \notin \mathbf{K_1} \cup \mathbf{K_2}$. From observation 2 we know that $M(\mathbf{K_1}, \mathbf{K_2}) \subseteq \mathbf{K_1} \cup \mathbf{K_2}$ and that $\alpha \wedge \beta \notin M(\mathbf{K_1}, \mathbf{K_2})$. We can conclude that all merging functions in the union when possible family are inconsistent with the definition of a collection function.

Let us study another case. Sometimes we might want to represent the beliefs of the collective with what most of its members believe. This could for example be the case for different organizations where voting is used to settle questions or for big groups of people where there is no complete consensus but where most people agree on most issues. This idea can be captured by a merging function that works as majority voting.

It is well known that voting can cause inconsistencies. Suppose that there are three persons A, B and C voting on the three alternatives α, β and $\alpha \wedge \beta$. Further, A votes for α and for β, B votes for α but against β and C votes against α but for β. When A, B and C vote the result will be for α, for β but against $\alpha \wedge \beta$. From α and β you can derive $\alpha \wedge \beta$, so we would have an inconsistency.

Two different possible ways to solve this problem will be investigated here. Let us first solve it by always have the vote on atomic sentences. The collective will believe all the atomic sentences getting more than half of the votes, and the consequences of those sentences. If we adjust the example just used to epistemology, let the belief set of A be $\text{Cn}(\alpha,\ \beta)$, the belief set of B be $\text{Cn}(\alpha,\ \neg\beta)$ and the belief set of C be $\text{Cn}(\neg\alpha,\ \beta)$. There are two atomic sentences that will get a majority vote, α and β. In this case then, $\text{M}(\mathbf{K_1},\ \mathbf{K_2},\ \mathbf{K_3}) = \text{Cn}(\alpha,\ \beta)$.

Definition, simple epistemic majority voting function: A merging function $M()$ is a simple epistemic voting function if and only if $M(\mathbf{K_1}, \ldots, \mathbf{K_n})$ is the minimum belief set containing all atomic sentences α such that it is an element of at least $n/2 + 1$ of the belief sets, rounded down.

This captures the idea that the collective believe what most of its members believe, like in the sentence “Most scientists believe in the green house effect”.

Observation 7: The simple epistemic majority voting function can not be a merging function for a collection function.

***Proof 7*:** Suppose that there is a simple epistemic majority voting function $M(\mathbf{K_1},\ \mathbf{K_2},\ \mathbf{K_3})$ and that $\mathbf{K_1} = \text{Cn}(\{\alpha,\ \neg\beta,\ \chi\}), \mathbf{K_2} = \text{Cn}(\{\alpha,\ \beta,\ \neg\chi\})$ and $\mathbf{K_3} = \text{Cn}(\{\neg\alpha,\ \beta,\ \chi\})$. It follows that $M(\mathbf{K_1},\ \mathbf{K_2},\ \mathbf{K_3}) = \text{Cn}(\{\alpha,\ \beta,\ \chi\})$. From Observation 4 we know that $M(\mathbf{K_1},\ \mathbf{K_2},\ \mathbf{K_3})$ must be a subset of one of $\mathbf{K_1}$, $\mathbf{K_2}$ and $\mathbf{K_3}$, but in this case it is not. Therefore, the simple epistemic majority voting function can not be a merging function for a collection function.

Let us try to use another voting function to solve this problem. Instead of voting on individual sentences, we let the vote be on belief sets. The belief set of the collective will be the strongest belief set that gets a majority vote. If more than one such belief set gets a majority vote, then the intersection of the belief sets wins the vote. Since this voting scheme is a bit more advanced than the simple epistemic majority voting function, simple is not in its name.

Definition, epistemic majority voting function: A merging function $M()$ is an epistemic voting function for n belief sets if and only if $M(\mathbf{K_1}, \ldots, \mathbf{K_n})$ is the intersection of all belief sets that are intersections of $n/2{+}1$ of $\mathbf{K_1}, \ldots, \mathbf{K_n}$, rounded down.

This voting function does not have the same problem as the simple voting function.

Observation 8: If $M()$ is an epistemic majority voting function, then there is a $\mathbf{K_i}$ such that $M(\mathbf{K_1}, \ldots, \mathbf{K_n}) \subseteq \mathbf{K_i}$.

***Proof 8*:** Since $M(\mathbf{K_1}, \ldots \mathbf{K_n})$ is the intersection of at least $n/2 + 1$ of the belief sets there is a $\mathbf{K_i}$ such that $M(\mathbf{K_1}, \ldots, \mathbf{K_n}) \subseteq \mathbf{K_i}$. From this we can conclude that the limit from observation 4 does not exclude the epistemic majority voting function as a merging function for a collection function.

Unfortunately, a counter example for the epistemic majority function is easily created.

Observation 9: An epistemic majority voting function can not be a merging function for a collection function.

***Proof 9*:** Suppose that there is an epistemic majority voting function $M(\mathbf{K_1},\ \mathbf{K_2})$ and that $\mathbf{K_1} = \mathrm{Cn}(\{\alpha,\ \beta\})$ and $\mathbf{K_2} = \mathrm{Cn}(\{\alpha,\ \neg\beta\})$. It follows that $M(\mathbf{K_1},\ \mathbf{K_2}) = \mathrm{Cn}(\{\alpha\})$. Now suppose that $\mathbf{K_1}$ and $\mathbf{K_2}$ are contracted with β. From the condition of collective change, we get that $M(\mathbf{K_1} - \beta,\ \mathbf{K_2} - \beta) = M(\mathbf{K_1},\ \mathbf{K_2}) - \beta$. From -3 it follows that $M(\mathbf{K_1},\ \mathbf{K_2}) - \beta = M(\mathbf{K_1},\ \mathbf{K_2})$. A legitimate result of $\mathbf{K_1} - \beta$ according to the AGM postulates is $\mathrm{Cn}(\{\beta \rightarrow \alpha\})$, while we from -3 get that $\mathbf{K_2} - \beta = \mathbf{K_2}$ which gives that $M(\mathbf{K_1} - \beta,\ \mathbf{K_2} - \beta) = \mathrm{Cn}(\{\beta \rightarrow \alpha\})$. From this we can conclude that the condition of collective change does not hold for a collection function which uses an epistemic majority voting function for merging.

Note that this counter example relies on the individual agents having some specific epistemic entrenchment ordering. Since a collection function is from an epistemic state to an epistemic state, this is enough to show that the function is not defined for all cases.

The exact same counter example can be used in other cases as well. An interesting merging function is one that gives the consensus among all its members. The consensus will be a subset of all the individual belief sets so it is within the allowed limits set up by observation 1–5.

> **Consensus merging function:** A merging function $M()$ is a consensus merging function if and only if $M(\mathrm{K}_1, \ldots, \mathbf{K_n}) = \mathbf{K_1} \cap \ldots \cap \mathbf{K_n}$.
>
> **Observation 10:** A consensus merging function can not be a merging function for a collection function.
>
> ***Proof 10*****:** Exactly the same as proof 9.

9.7 Discussion

We wanted to study the possibility of collective rationality. More specifically, we investigated if belief changes on a collective level could fulfill the AGM postulates. To make a detailed study possible, assumptions were made about the connection between the collective and its members. First, it was assumed that the epistemic state of the collective is a function of its members' epistemic states. Secondly, it was also assumed that the belief set of the collective is a function of its members' belief sets. Thirdly it was assumed that if all members of the collective change their beliefs according to some epistemic input, then the collective changes its beliefs according to the same input.

The results that were derived from these assumptions are mostly negative. Observations 1–5 concern the limits of the collective belief sets. Observation 1 shows that the smallest possible belief set of the collective is the consensus of its members.

Observations 2–4 are about the upper limit on the collective belief set. The results exclude the possibility of a collective believing more than any of its member. This closes the possibility to view collectives as rational entities believing more than any of its members.

Observations 5–10 instead focuse on five different merging functions. Of these, only the epistemic dictatorship function is consistent with a collection function.

How should we handle these results? One possibility is to conclude that we can not speak about collectives as rational belief agents. They do not have beliefs, or if they have them, they don't change them. There is no similarity between how an individual changes her beliefs and the development of theories or science.

Another possibility is that there is rationality on the collective level, but that the rationality follows other rules. Consider a case where the members of a collective have beliefs about something's place on a scale (for example something's length, weight or temperature). In such a case it might seem reasonable for the collective to believe less than consensus. It could seem reasonable to assume that the collective believes that the thing could be anywhere on the scale between the highest point believed by any member, and the lowest point believed by any member. Observation 1 states that the collective believes at least the intersection of its members' beliefs. Since all the agents believes that the thing is at some certain point on the scale, they also believe that it is at the point the first agent believes or at the point the second agent believes or ... or at the point where the *n*th agent believes. This is thus also believed by the collective. That however excludes all other points on the scale except those believed by at least one of the members of the collective.

An example can illustrate this. Suppose that two persons are taking a walk. A car drives by them faster than the speed limit. The two persons disagree on the exact speed of the passing car. One of them believes that the car traveled in 60 km/h, while the other believes that the speed of the car was 75 km/h. Now we merge these two agents' belief sets into a collective belief set. The collective agent can only have one of the following beliefs about the speed of the car:

1. The collective agent can believe that the speed of the car was 60 km/h.
2. The collective agent can believe that the speed of the car was 75 km/h.
3. The speed of the car is either 60 km/h or 75 km/h.

In this case, one alternative that might seem reasonable is that the collective agent believes that the speed of the car can be anything between 60 and 75 km/h, that is either 60 or 61 or or 74 or 75 km/h. This is however weaker than the lower limit where the collective agent must believe that the speed is either 60 or 76 km/h, nothing in between. One could try to find other intuitive examples like this and try to construct rules for how the collective changes its beliefs from that.

A last possibility would be to find an alternative to the AGM postulates for individual as well as collective agents. Observation 2–5 relies on the belief set being closed under logical consequence. If one instead uses the belief base model developed by Sven-Ove Hansson (1991) where beliefs are represented by sets of sentences not closed under logical consequence, one might get other results. For observation 9 and 10 the solution could instead be to add extra postulates that block the possibility of creating the counter example.

References

Alchourrón, C.E., Peter Gärdenfors, and David Makinson. 1985. On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic* 50:510–530.

Gärdenfors, Peter. 1988. *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge, MA: MIT.

Hansson, Sven-Ove. 1991. *Belief base dynamics*. Uppsala: Uppsala University.

Kuhn, Thomas S. 1962. *The Structure of scientific revolutions*. Chicago: University of Chicago Press.

Levi, Isaac. 1991. *Fixation of belief and its undoing*. Cambridge: Cambridge University Press.

Levi, Isaac. 1995. Cognitive value and the advancement of science. *Philosophy and Phenomenological Research* 50(3):619–627.

Liberatore, Paolo, and Marco Schaerf. 1998. Arbitration (or how to merge knowledge bases). *IEEE Transactions on Knowledge and Data Engineering* 10:76–90.

Sen, Amartya. 1970. *Collective choice and social welfare*. San Fransisco, CA: Holden-Day.

Chapter 10
The Best of All Possible Worlds: Where Interrogative Games Meet Research Agendas

Emmanuel Genot

10.1 Introduction

Erik J. Olsson and David Westlund have recently argued that the standard belief revision representation of an epistemic state is defective.[1] In order to adequately model an epistemic state one needs, in addition to a belief set (or corpus, or theory, i.e. a set closed under deduction) $\underline{K}$ and (say) an entrenchment relation E, a *research agenda* $\underline{A}$, i.e. a set of questions satisfying certain corpus-relative preconditions (hence called $\underline{K}$-questions) the agent would like to have answers to. Informally, the preconditions guarantee that the set of potential answers represent a partition of possible expansions of $\underline{K}$, hence are equivalent to well-behaved sets of *alternative hypotheses*.

This addition, according to the authors, could shed a new light on some old problems, and extend the range of application of belief revision theory. On the one hand, among the problems, is the role of pure contraction, that is contraction not followed by revision. Genuine examples of pure contraction are hard to come by, but such a contraction can be motivated by the intention to give some hypothesis a try, when the reason to do so is not to respond to some new information, but to better conform to some standard of epistemic economy. Contraction is needed to make room for the hypothesis. Yet belief revision theory as it stands cannot represent the commitment to investigate the new hypothesis as an alternative to the old one, which has been removed by contraction. The addition of an agenda shows that ‘pure’ contraction is indeed a response to some other change. On the other hand, agendas are a welcome addition if one is interested in applying the theory to philosophy of science. For example, Olsson and Westlund argue that adopting an *ad hoc* hypothesis (or other *ceteris paribus* restrictions) can be justified if a commitment

E. Genot (✉)
Department of Philosophy, University of Lille 3, Lille, France
e-mail: emmanuel.genot@etu.univ-lille3.fr

[1] The original theory is presented in Olsson and Westlund (2006).

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_10,
© Springer Science+Business Media B.V. 2011

is taken to investigate 'exceptionless' solutions (hence inscribing new questions on the agenda). Subsequently, the number (or some other qualitative evaluation) of such commitments can be used to appreciate the degree to which a research program has degenerated.

Let $\underline{\mathrm{K}}$ be a corpus of beliefs of some agent, and $Ag(\underline{\mathrm{K}})$ the associated agenda. Which questions should appear on $Ag(\underline{\mathrm{K}})$, among those possible questions satisfying $\underline{\mathrm{K}}$-relative preconditions? It seems that the content of $Ag(\underline{\mathrm{K}})$ should depend partly on the agent's current interests, and partly on the questions she had asked earlier, at previous states. With the exception of *why*— and *how*— questions (which will not be addressed here) expansion solves questions (partially or completely), by reducing the range of alternative hypotheses. Contraction never solves questions, and open new ones. But contraction may weaken a state to the extend that for some questions corpus-relative preconditions no longer hold. As a result, some questions may be deleted form the agenda without being answered.

While Olsson and Westlund give an account of how an agenda should react to expansion, the effect of contraction, though addressed, is not studied in full details. They define an operation to be applied to questions after expansion, and propose a postulate for state expansion, but no dual operation is proposed for contraction, and if some tentative postulates are stated, the characteristics of the contracted state remain formally underspecified, even if conceptually clear. Completing the theory with a functional account of question update following belief set contraction is thus not only important for the general framework—since it allows for a completed description of the effect of revision on agendas—but also because of its intended applications in philosophy of science.

The aim of this essay is to show how questions in $Ag(\underline{\mathrm{K}} \div a)$ may be made functionally dependent on questions in $Ag(\underline{\mathrm{K}})$. Using results and concepts borrowed from Jaakko Hintikka's logical analysis of questions, we propose to examine the possible effects of contraction on a corpus, characterize the kind of continuity that exists between a state $\underline{\mathrm{K}}$ and its possible contractions, so to speak 'question-wise'. The nature of this continuity naturally leads to two proposals for updating agendas, which in turn highlight two possible interpretations of Olsson and Westlund's initial theory, building an analogy with game-theoretical concepts of games in 'strategic' (or 'normal') form, and games in 'extensive' form. But this analogy may have deep consequences on belief-revision theory, and the possibility to apply it to study belief dynamics and (scientific) theory change.

We begin by contrasting Olsson and Westlund's perspective on questions (as sets of hypotheses) with another view, also epistemically motivated (Section 10.2). We first adopt a 'broad' epistemic perspective, i.e. without commitment to any specific formalism or modelling (Section 10.2.1), which serves to introduce key ideas of Hintikka's *interrogative model of inquiry*. This model is briefly introduced (Section 10.2.2), in order to contrast with the view of questions as sets of (rival) hypotheses.[2]

[2] There are many examples of logical analysis of questions, which will not be reviewed here. Our exposition is grounded in Hintikka's account (especially Hintikka et al., 1999, and Hintikka, 2003),

The two views are then brought together (Section 10.2.3) in connection with the Peircean notion of abduction, and the problem of hypothetical reasoning.

Section 10.3 reviews parts of Olsson and Westlund's initial take at the theory of agendas. After a brief exposition Olsson and Westlund's formal proposal for representing questions, and operate on them (Section 10.3.1), we expose the problem of continuity through contraction (Section 10.3.2).[3] It is suggested that *yes-or-no*–questions may play a special role in representing this continuity. The section concludes (Section 10.3.3) relating the problem to game-theoretic concepts, in order to consider the problem of update as a strategic problem.

Section 10.4 returns to interrogative and inquiry games, beginning with a formal analogy (Section 10.4.1) between, in the one hand, corpus relative questions and their associated sets of *yes-or-no*-– questions, and on the other hand *strategic* and *extensive* (forms of) games. A solution for updating questions *qua* strategic games is proposed first (Section 10.4.2) due to its close relation to corpus-relative questions as they were first introduced. A second solution is then offered (Section 10.4.3) closer to the spirit of the interrogative model of inquiry.

In conclusion, we extend our analogy with games in strategic and extensive forms to other topics in belief revision theory, an suggest that inquiry games may throw some light on some conceptual issues in the interpretation of belief revision theory.

Before proceeding, one last remark is to be made about the limitation of the present essay. The reader may wonder why we do not discuss in this essay the topic of ampliative inference, and in particular the role of statistical and probabilistic inference, given their importance in scientific practice and philosophy of science: research questions about universal generalization cannot be answered unless some means is available to go past the finite data. One inessential reason is that the focus, in belief revision theory, on propositional representation masks the quantificational complexity of answers. There is however a more substantial reason: the problem is less related to principles pertaining to the framing of question, than to possible rules for accepting answers. The following considerations will explain why.

Nature's answers to experimental questions are only partial. As an example, a functional dependency can be conjectured, but the data will only give a partial graph (e.g. values for an observed variable for a finite set of values of a control variable). Alternatively, statisticians and learning theorists have studied how such conjectures may arise from the data. One could expect that when a potential answer to a research question is a universal statement, a research agenda also includes some advice as to how the the partial answer obtained from observation or experimentation is to be

since his analysis pays systematic attention to both the epistemic and strategic aspects of questioning.

[3] It should be noted that Olsson and Westlund's theory is taken as a first step toward a more detailed account of *interrogative contraction*, a topic which will not be addressed here. This kind of contraction is intended to model 'pure' use of contraction, while our proposal addresses the topic or 'regular' contraction as an analytic step of *revision*. The relevance of our proposal to interrogative contraction is yet to evaluate, and may be a subject for further research.

inductively expanded. If this is so, an agenda should suggest recommendation about testing procedures.[4]

Though recommendations of this sort will be part of an inquiry setting, they depend on the a available sources of answers, as well as the available answers, which are contextual parameters. Moreover, as suggested by Hintikka, testing procedures, statistical ‘inverse inference’ methods, sampling procedures, etc., should be viewed, in the light of the interrogative model, as rules to ‘shelve Nature’s answers’, the choice of which may also depend on the standard of acceptance of an inquirer.[5]

As a consequence, since this essay discusses primarily the dynamics of questions, rather than the procedures to collect, classify, accept or reject answers, the problem of answer-gathering methodology will remain unexamined. When discussing the effect of acceptance of answers, we will assume that answers have been accepted in compliance with some contextually appropriate standard.

Notation. We will use (mostly from Section 10.3 on) the following conventions. Let $\mathcal{L}$ be a propositional language with the usual connectives and syntax, lowercases $a, b, \ldots$ being propositional variables of arbitrary boolean degree. We assume $\mathcal{L}$ contains the propositional constants $\top$ (*verum*) and $\bot$ (*falsum*), with their usual interpretation.

Uppercases $A, B, \ldots$ are used as variables for sets, and special names are used whenever needed. We also presuppose the usual set-theoretic operations and notation. $\mathrm{Cn}(\cdot)$ denotes a consequence operator, that is a function from $2^{\mathcal{L}}$ to $2^{\mathcal{L}}$ which sends a set of sentences to the set of its classical (truth-functional) consequences.

$\underline{\mathrm{K}}$ denotes a set closed under classical consequence (that is: $\mathrm{Cn}(\underline{\mathrm{K}}) \subseteq \underline{\mathrm{K}}$). Sub- or superscripts will be added whenever needed. A ‘hybrid’ notation may occasionally be used, mixing object- and metalanguage levels such as $\vee E$ ($\wedge E$), where $E \subset \mathcal{L}$ (it is assumed that E is *finite*) to denote the (complex) proposition formed by the conjunction (disjunction) of elements of E. We will use $(a \veebar b)$ to denote exclusive disjunction of a and b, i.e. to abbreviate: $(a \vee b) \wedge \neg(a \wedge b)$, and $\veebar\{a_1, \ldots, a_n\}$ to denote the exclusive disjunction $(a_1 \veebar \ldots \veebar a_n)$.[6]

Finally, let $\underline{\mathrm{K}} + a$, $\underline{\mathrm{K}} \div a$ and $K * a$ denote, respectively, the *expansion*, *contraction* and *revision* of corpus $\underline{\mathrm{K}}$ by information (proposition) a. Set-theoretic construction

[4] Consider an research agenda to find linear models to explain some phenomenon. Should the agenda include recommendation to proceed through randomized experiments, or using some sampling method? If the former, should one randomize simultaneously or one variable at a time? If the later, how should one sample? (Thanks to an anonymous referee for raising these questions.)

[5] See Hintikka (1987b) for an analysis of probabilistic reasoning and statistical tests in the interrogative model framework. Standards of acceptance are discussed in Hintikka (2007a), where their context- and subject-dependency is related to decision-theoretic aspects of inquiry, i.e. the role of conclusions in decision-making.

[6] Since exclusive disjunction is not associative, any n-ary $\veebar$ should be treated as a distinct operator. We will nevertheless use it as is, relying on the readers’ logical acumen (or interpretative charity) to restore proper use and notation if they find our choices inappropriate.

corresponding to expansion $\mathrm{Cn}(\underline{\mathrm{K}} \cup \{a\})$ will be used as an alternate notation for expansion. Expansion of $\underline{\mathrm{K}}$ by a is said to be *consistent* iff $\perp \notin \underline{\mathrm{K}} + a$ (equivalently, if $\underline{\mathrm{K}} + a \neq \mathcal{L}$).

10.2 Two Views on Questions

10.2.1 An Epistemic Perspective on Questions

Informally speaking, a question Q can be raised with respect to a corpus $\underline{\mathrm{K}}$, if according to $\underline{\mathrm{K}}$ at least one of the answers to Q must hold. It is common to identify this condition as the *presupposition* of Q, and to say that $\underline{\mathrm{K}}$ entails, or satisfies the *presupposition* of Q. We will only consider *propositional* (or *whether*-questions), furthermore with only finitely many potential answers.[7] Let's have a closer look at the notion of presupposition relative to those questions.

Epistemically speaking, one can say that $\underline{\mathrm{K}}$ specifies (for a given agent) a space of possible situations (worlds, scenarios, etc.), and that each answer to Q would further restrict the agent's attention to a subset of these situations—provided the answer is both obtained and accepted. An answer provides information, insofar as information is understood qualitatively as elimination of possibilities—here, restriction of the range of admissible epistemic/doxastic alternatives. The set of potential answers to Q divides the space of $\underline{\mathrm{K}}$-compatible alternatives (but it can fall short from partitioning it in a strict sense, see below).

What precedes can be expressed in logical terms saying that, in the case of *whether*- (or 'propositional') questions, $\underline{\mathrm{K}}$ entails the disjunction of the potential answers (this disjunction being exclusive whenever the division is a partition). Hence, the presupposition of a *whether*- (or propositional) question is nothing but the disjunction of its potential answers. The potential answers themselves can in turn be (for all practical purposes) identified with a *set of hypotheses* (rival if it is a partition, partially compatible if not).

A question is a step in a goal-directed *knowledge-seeking by questioning* activity—at least in the cases we are interested in. Hence the intended effect of a question Q being answered is as important as the presupposition of Q. This effect is the restriction of attention to a subset of the admissible scenarios, those in which the (content of) the answer hold. Following Hintikka again, let's call this effect of a question the *desideratum* of Q. Hence, any potential answer being received and accepted brings about the desideratum (but turning a reply into an answer

[7] Finiteness is to be understood as holding up to logical equivalence between potential answers, since there can be infinitely many 'equivalent' ways to ask the 'same' question. Question equivalence is detailed p. 20, n. 42.

may require several steps, see *infra*). A question can thus be identified as a pair presupposition-desideratum.[8]

In a broad sense, a relevant reply to a question Q is whatever information which restricts attention in some way which would affect the way to ask Q, that is, which would lead the questioner to re-formulate Q or to consider a different set of hypotheses. For example, if at least one of the potential answers to Q is known *not to hold*, then Q is rhetorical. This may not seem at first natural, since a rhetorical question is generally viewed as a question of which the answer (or part of the answer) is known. But the following considerations will make the connection clearer. First, it can be useful to distinguish between partial answers and complete ones. Let us first say that a *partial* answer to a question Q is any information which restricts the range of admissible alternatives to a subset of the hypotheses (answers) with respect to which Q is formulated. Then, a *complete* answer will be an information that restricts attention so a *singleton* subset of those hypotheses (notice that complete answers are thus special cases of partial answers). It is easily seen that, under this definition, a question is rhetorical if it has been partially answered, hence if some potential answer is known not to hold.

Relevance of a reply depends on background information. Bas van Fraassen illustrates this with a famous example:

> Almost anything can be an appropriate response to a question [. . .] as '*Peccavi*' was the reply telegraphed by a British commander in India to the question how the battle was going (he had been sent to attack the province of Sind).[9]

Following Hintikka, let the additional information (needed to derive an answer from a reply) be referred to as *conclusiveness condition*, since it is what 'turns' the reply into a conclusive answer. In the above example, '*Peccavi*' is the *reply*, and the *conclusiveness condition* is given by the knowledge of Latin, at least enough to translate 'I have sinned', and subsequently understand the pun.[10]

Hence, just noas a question can be identified through a pair (its presupposition and desideratum), so can an answer. A reply will be part of a conclusive answer if some conclusiveness condition is known. Obtaining a conclusiveness condition is sometimes possible through mere deduction (as in the above example, assuming knowledge of Latin), but it can also involve more convoluted patterns of reasoning. This phenomenon is well known in philosophy of language, especially in connection with Gricean pragmatics and implicatures of seemingly irrelevant answers. A worn example, first used by Francois Recanati, is the question: 'Do you know how to

[8] Since a question can be raised only if its presupposition is known (or believed) to hold, interrogative strategies involve sometimes elaborate plans to establish the presupposition of a given question (see also n. 14).

[9] See van Fraassen (1980, p. 138); as van Fraassen adds in note, the reply is attributed to Sir Charles Napier (1782–1853), Commander-in-chief of India, after whom the city of Napier, New Zealand, is named.

[10] Strictly speaking, obtaining an answer is a mere deductive move if the question was: 'Do you have Sind?'.

cook?', responded with: 'I'm French!'. The questioner will probably have to 'guess' what the addressee would answer to a question about the link between being French and being a decent cook, and this guessed answer is precisely the conclusiveness condition needed for the received response to bring about the desideratum of the original question.[11]

With a bit of simplification, one can call a *direct* answer to a *whether*-question any potential answer which is mentioned, or referred to, in the question itself, thus having a 'trivial' conclusiveness condition. Unfortunately, this approximate definition is not really satisfying as shown by the following example. To the question:

Do you plan to serve filet or fish for diner? (1)

one can have a variety of responses, or replies (and therefore, of answers):

I have planned to serve fish (2a)

I have planned to serve a soup, and then fish. (2b)

I have not planned not to serve a soup, or not to serve fish. (2c)

I have planned to serve salmon in sour cream sauce. (2d)

It seems that (2a) is a 'natural' candidate for being labeled a *direct* answer, while (2b), (2c) and (2d) are *indirect*. Yet (2b) and (2c) require only a *deduction*, while (2d) has to be backed by suitable background knowledge. On the other hand, some inferences may seem less natural than some uses of widespread background knowledge, as showed (on our opinion) by the difference between (2c) and (2d).

Maybe the direct-indirect distinction should be dispensed with altogether since all can be re-cast in terms of conclusiveness conditions.[12] A finer-grained (yet, in view of the above example, insufficient) distinction could be made between 'deductive' conclusiveness conditions, and 'non-deductive' ones.[13]

However, this problem is secondary in BRT-related applications, since the distinction is blurred by the fact that corpora, relative to which any question is asked, and to which any reply is added, are closed under deduction. We should not forget however that it has to be addressed in any non-idealized (epistemic) theory of questions, and the next section presents a brief overview of the topics this discussion should cover.

[11] Recanati considers that the reply 'clearly [...] provides an affirmative answer' (see Recanati, 2001).

[12] This is the solution adopted by Hintikka, in contrast to others, e.g. Van Fraassen (1980).

[13] The latter category is meant to include conclusiveness conditions which are deductive *modulo* suitable background knowledge, as well as those which require some additional assumption (maybe some form of abductive reasoning) as illustrated in Recanati's example.

10.2.2 Interrogative Games

Consider a situation in which an inquirer, given some background knowledge (BK), tries to establish the truth of some hypothesis. If her BK is not sufficient to establish deductively the conclusion (or to disprove it), then she will have to obtain some more information by putting questions to *sources*. Assume further that all the questions she asks can be answered (or, more realistically, that she selects only questions answerable by available sources), and that all answers can be treated as true (that is, are together consistent, and consistent with BK).

This situation can be considered as a 'game' where an Inquirer can, in order to prove her hypothesis, either use *deductive moves* or *interrogative moves*. This kind of game can be considered as a modern descendant of the questioning games practiced in Plato's Academy and Aristotle's Lyceum. This connection has lead to valuables insights in the so-called theory of fallacies.[14]

Because of our restrictive assumptions, every interrogative move will be answered—though, in order to be allowed, the presupposition of the corresponding question has to be already established, either by appeal to BK, or using already answered questions. Moreover, *yes-or-no*–questions may always be asked: their presuppositions are trivial (being instances of the excluded middle), at least for well-defined terms.[15]

Borrowing from game theory, such a game can be considered as a two-player zero-sum game with perfect information. One player, Inquirer, plays against Nature and tries to establish her hypothesis using available information: she has a winning strategy iff she can prove her hypothesis using (deductively) only answers from sources and information in BK. The game is strictly competitive, hence is zero-sum (Nature wins when Inquirer looses, and reciprocally), and with perfect information since: (i) all questions are answerable, and (ii) information is available at any time (that is, at any state the game can reach) for any player once it has been obtained.[16]

Though these assumptions may seem unrealistic, they nevertheless coincide with the typical Sherlock Holmes case (though Holmesian 'deductions' incorporate guesses, see below). The 'logic' which describes this situation can be called

[14] See in particular Hintikka (1987a), which includes a simple introduction to interrogative games from the perspective of Socratic questioning. As an example, the well-known *Fallacy of Many Questions* is a case where a question *Q* is asked though its presupposition has not been established: I cannot (rightfully) ask whether you have stopped beating your dog if I cannot take for granted a 'positive' answer to the question whether you (ever) beat it. This is not much a fallacy than it is an illegitimate 'move' in an 'interrogative game'. This re-interpretation of fallacies dates back to R. Robinson (1971), who addresses (in a very spirited manner) the so-called *Fallacy of Begging the Question*. Its fruitfulness has been defended by Hintikka (1997) in contrast with other approaches.

[15] Asking a *yes-or-no*–question about some vague term, for example, may need some 'contextual standard' to be fixed. This problem is studied in David Lewis' paper (1979).

[16] If Inquirer were allowed to take a guess (that is, if she asks questions which are not answerable), the game would proceed with *imperfect* information, since Inquirer would not know in which state she is. But she would still have *perfect recall*, since in a given play of the game, she could re-use any information obtained during that play.

the logic of *pure discovery*, 'a type of inquiry in which all answers [...] can be treated as being true [and where] we do not have to worry about justifying what we find' (Hintikka, 2007d, p. 98). A moment's reflection will convince the reader that this logic is simply ordinary (fisrt-order) logic, supplemented with interrogative steps, hence, as Hintikka writes, 'a logic that is little more than the good old deductive logic viewed strategically' (Hintikka, 2007d, p. 2). This intuition can easily be represented formally, through some variant of Beth tableaux.[17]

The simplest formal representation is obtained with a tableau system, where the left column represents the 'Oracle' side, and the right column the 'Inquirer' side (possibly using rewriting rules to avoid traffic between columns). The aim of Inquirer is to build a countermodel to the hypothesis she's investigating (placed on the top of the right column) while information in the left is (i) BK and (possibly) other active hypotheses; and (ii) answers to interrogative moves; and (iii) whatever conclusion that can be reached applying deductive (analysing) moves to the information in that column. Player Nature is not represented, and is indeed reduced to its 'answering' role of Oracle. This can be though of as a formal counterpart of the well-known metaphor of the researcher putting Nature to question.[18]

Since our discussion is restricted to the *propositional* case, the only interrogative moves we will consider are those prompted by occurrences of *disjunctions* in the left column, while 'vacuous presuppositions' (for *yes-or-no*–questions) can be added on the left side any time. The difference between the outcome of deductive moves (illustrating 'reasoning by cases') and interrogative moves is illustrated by Fig. 10.1: instead of producing two subtableaux (logical move, on the left) an interrogative move (if the source answers) allows for Inquirer to continue the game with the answer, now in her information set, without considering the other possible answer(s) (as illustrated by the right-hand tableau).[19]

Adding a rule permitting Inquirer to ask *yes-or-no*–questions whenever she wants amounts to allow for the introduction of a disjunction ($a \vee \neg a$) whenever Inquirer wants, which is consistent with the usual interpretation of tableaux (everything on

[17] These modified Beth tableaux are presented in Hintikka and Halonen (1999), and are used in more details in Hintikka (2007b). They bear a strong resemblance with the 'logical dialogues' of the Lorentz–Lorenzen–Rahman tradition (see e.g. Rahman and Keiff (2005).

[18] For an alternate interpretation, giving a more 'active' role to the left column player, see n. 19.

[19] This interpretation of Beth Tableaux is often put forward by Hintikka, and is closely related to ∃loise vs. ∀belard games of Game-Theoretic Semantics (GTS) (Harris, 1994, is an early attempt at a GTS for interrogative tableaux). Another interpretation, in the tradition of Lorenzen, Lorenz and Rahman, is to see it as a game where the right-column player, Proponent, tries to defend a thesis against criticism of the left-hand player, Opponent. Both player can attack and defend statements according to certain particle rules (governing the use of logical constants and quantifiers) and structural rules (governing the overall conduct of the game). Proponent has a winning strategy if she can have the last word, whatever Opponent may do, using only information he has previously conceded (and logical moves). For the latter interpretation, see Rahman and Keiff (2005), and for its relation to the former, see Rahman and Tulenheimo (2007), where some kind of interrogative moves are introduced in dialogues.

Fig. 10.1 Deductive vs. interrogative moves

['Oracle']	['Inquirer']
$b \vee c$	
$b \mid c$	$\mid$

['Oracle']	['Inquirer']
$b \vee c$	
b	

the left column being true).[20] Obviously, if the 'game' interpretation is to be carried further, some rules against 'delaying tactics' have to be introduced in order to prevent Inquirer from keeping asking questions to avoid loosing.[21]

Suitably developed, such a model can account for the various features of 'deductive' reasoning *à la* Sherlock Holmes, i.e. deductive reasoning interwoven with interrogative steps.[22] Obviously, this view contrasts with a widely held view, according to which Sherlock Holmes' reasoning is improperly called 'deductive', since it should rather be described as some kind of non-monotonic reasoning. This view is authoritatively stated by David Makinson:

> In the stories of Sherlock Holmes, his companion Watson often speaks of the master's amazing powers of deduction. Yet those who have some acquaintance with deductive logic [. . .] realize that the game is not the same. None of the conclusions drawn by Sherlock Holmes follow deductively, in the strict sense of the term, from the evidence. They involve presumption and conjecture, and the ever-present possibility of going wrong. According to Watson, Sherlock Holmes usually got it right. (Mackinson, 2005, p.1)

The case against deductive logic is not that clear-cut if one realizes that 'presumption and conjecture' represent possible 'guesses' of Holmes' part, were an interrogative move cannot be responded by any other source than himself. This goes beyond 'pure discovery' (since in this case the answer cannot be simply treated as true), hence is beyond deductive logic. One can concede to Makinson that Holmesian deductions involve 'educated guesses' in the form of some interrogative steps responded by Holmes himself, without sufficient evidence (that is, without independent source). When Holmes is forced to give up some previous conclusion, non-monotonic reasoning is clearly needed to make sense of his change of mind. But in other cases, deduction *cum* interrogation suffices.[23]

[20] Addition of this rule defines, which is equivalent to the *cut rule*, defines what Hintikka, Halonen and Mutanen name *extended interrogative logic* (see Hintikka et al., 1999, p. 53), while the unextended interrogative logic is the cut-free (analytical) version obtained if Inquirer is *not* allowed to ask *yes-or-no*–questions.

[21] See Rahman and Keiff (2005) for such a rule in the context of proof games (called *formal dialogues*).

[22] For a more complete presentation, see Hintikka et al. (1999), which also includes various correspondence results between 'interrogative' reasoning and deductive reasoning, as well as some metatheorems we will use in this paper. In particular, it shows that the problem of finding the 'best' interrogative strategy reduces (in the case of 'pure discovery') to the problem of finding the best *deductive* strategy (at least for unextended interrogative logic) which, in view of the semi-decidability of first-order logic, is not solvable computationally.

[23] In Hintikka et al. (1999, sec. 8) the problem of reasoning with uncertain answers addressed briefly, but no explicit connection is proposed with non-monotonic logics. A connection is made

10.2.3 Abduction, Hypotheses and Belief Revision

The ‘abductive step’ is commonly taken to include the determination of a set of potential answers to some question, a view which is often attributed to Peirce, and to which Olsson and Westlund explicitely refer.[24] Peirce seems to have included in abduction the preference an inquirer may have for one hypothesis over another prior to any test of the hypothesis (induction, in Peirce’s terminology). The importance of this often neglected aspect of Peirce’s conception of abduction has been repeatedly stressed by Hintikka in connection with the Interrogative Model of Inquiry (IMI hereafter).[25] Moreover, abduction is presented by Peirce as involving some kind of acceptance, which falls obviously short of qualifying as belief:

> It is to be remarked that, in pure abduction, it can never be justifiable to accept the hypothesis otherwise than as an interrogation. But as long as that condition is observed, no positive falsity is to be feared; *and therefore the whole question of what out a number of rival hypotheses ought to be entertained becomes purely a question of economy.* (Peirce, 1940, p. 154) (Emphasis added.)

The primary concern of the IMI framework is to give a model of what happens once the abductive step has been *fully* carried, i.e. once a *principal* question has been identified and one of its potential answers chosen, in order to begin investigation.[26] On the other hand, Olsson and Westlund’s framework considers a question as a *set of hypotheses*, each being a possible candidate for the critical scrutiny represented by inquiry. It is then quite natural to look at both frameworks as complementary, rather than rival or opposed.

One problem that can be solve by this complementarity is the problem of modeling investigations into a set of (mutually exclusive and together exhaustive) hypotheses. Considering this kind of investigation as an expansion, followed by a contraction in order to give another hypothesis a try, is at odds with the idea of minimal mutilation, or epistemic economy, put forward by belief revision theorists, if one intends to impose that the initial ‘belief’ state, *including the questions opened at the outset of inquiry*, should be recovered as it was before investigation. And this

with probabilistic reasoning, and especially the problem of *cognitive fallacies*, later developed in Hintikka (2004).

[24] For example, Isaac Levi (1991, p. 71) writes that: ‘The task of constructing potential answers to a question is the task of abduction in the sense of Peirce’. Hintikka (1999, p. 104) agrees with Levi to the extend that ‘from a strategic viewpoint, in that the choice of the set of alternative answers amounts to the choice of questions to be asked’, but insists that: ‘in abduction one may prefer one possible conjecture over others’ that is, favor one answer over others. O&W explicitly refer to Peirce (in Isaac Levi’s interpretation), and consider that a set of potential answers being considered as ‘given’ (by abduction) is a ‘methodological decision that has to [their] knowledge never been questioned’ (Olsson and Westlund, 2006, p. 179, n. 3).

[25] See especially Hintikka (1988), and, more recently Hintikka (2007d, Essay 2).

[26] This may be a way to understand what is meant by acceptance ‘as an interrogation’, or at least a sensible interpretation. Let’s for the moment simply say that what we’re aiming at is to formally reconstruct one possible understanding of Peirce’s idea.

constraint seems sensible since reasoning ‘for the sake of the argument’ should not alter the set of unsettled questions.[27]

Therefore, a model which allow for ‘parallel’ rather than ‘serial’ investigation can be well-motivated, and provide some incentive to combine BRT *cum* agendas and IMI. A question of the kind considered by the theory of agendas may provide a set of ‘principal questions’, and each potential answer to each question can be the object of a separate investigation. The details of such a model are not at all trivial (especially when considering how some ‘local’ effects of inquiry may have ramifications on the global level). This question will be addressed in conclusion. But at least it shows that some multi-level model may be of interest.[28]

Other ways are certainly open, more ‘conservative’ with respect to the BRT tradition. Nevertheless the import of the IMI may help to solve some difficulties. Indeed, the IMI stresses the importance of interrogative *strategies*. And this notion of strategy (or a shadow of it, yet an impressive one) will prove particularly useful to address a problem left open by Olsson and Westlund’s pioneering work on agendas, the problem of how an agenda should be updated given that the corpus to which it is associated undergoes a contraction.

10.3 Research Agendas, Expansion and Contraction

10.3.1 Formal Representation of Questions

Olsson and Westlund (O&W hereafter) identify a (propositional) question to be included on an agenda associated with a corpus $\underline{K}$ ($\underline{K}$-questions, following their terminology) as a (finite) set of *potential answers* which partitions the possible (consistent) expansions of $\underline{K}$ by elements of Q: i.e. $\underline{K}$ cannot be consistently expanded by the conjunction of two or more elements of Q (in what follows, partitions of consistent expansions will always be relative to the set of potential answers). Hence $Q = \{a_1, \ldots, a_n\}$ is a $\underline{K}$-question iff: (i) $\veebar Q \in \underline{K}$; and (ii) there is no $Q' \subset Q$ such that $\veebar Q' \in \underline{K}$. Let $\mathrm{Q}_{\underline{K}}$ denote the set of $\underline{K}$-questions, and $Ag(\underline{K})$ denote the agenda associated with corpus $\underline{K}$. There is no constraint on which question should be included in a $\underline{K}$-agenda (for some $\underline{K}$), save for their being $\underline{K}$-questions, that is, the only constraint is: $Ag(\underline{K}) \subseteq \mathrm{Q}_{\underline{K}}$.

[27] Assume that one wants to investigate the consequences of adding hypothesis a_i to the belief set $\underline{K}$, and forms the expansion $\underline{K} + a_i$. Let furthermore $(a_i \leftrightarrow (b_1 \wedge b_2)) \in \underline{K}$. Assume now that one wants to investigate the consequences of hypothesis a_j, where a_j is a ‘rival’ to a_i, and then wants to restore the initial state before expansion by a_i. Since epistemic economy recommends that contraction by a_i only removes b_1 or b_2, but not both, it follows that either the question whether b_1 or the question whether b_2 will remain settled: then ‘purely hypothetical’ reasoning is not likely to be adequately modelled. Another option is to define a *question-relative* contraction, and such a solution was outilned in Olsson and Westlund (2006), yet not fully articulated.

[28] There are other important motivations, but they require the discussion of interrogative strategies, and we will wait to have introduced and discussed this notion before returning to this question.

Unlike O&W, but following a common practice, we do not refer to conditions for inclusion of Q in $\mathrm{Q}_{\underline{\mathrm{K}}}$ as 'presuppositions' of Q, but as *preconditions*, and we use 'presupposition' to refer to the disjunction of potential answers of a given question. That is, the presupposition of Q is $\vee Q$, and $\underline{\mathrm{K}}$ *satisfies* the presupposition of $\underline{\mathrm{K}}$ whenever $\vee Q \in \underline{\mathrm{K}}$. If $\vee Q' \in \underline{\mathrm{K}}$ for some $Q' \subset Q$, we say that Q is (a) *rhetorical* (question) with respect to $\underline{\mathrm{K}}$. A non-rhetorical question w.r.t. $\underline{\mathrm{K}}$ of which the presupposition is satisfied by $\underline{\mathrm{K}}$ is said to be a *genuine question* (w.r.t. $\underline{\mathrm{K}}$). $\underline{\mathrm{K}}$-questions are, of course, a special case of genuine questions w.r.t. $\underline{\mathrm{K}}$.

Let's say that an expansion of $\underline{\mathrm{K}}$ by b *partially answers* $Q = \{a_1, \ldots, a_n\}$ if (and only if) there is a $Q' \subset Q$ such that $Q' \in \mathrm{Q}_{\underline{\mathrm{K}}+b}$, and *completely answers* Q if (and only if) $Q' = \{a_i\}$, in which case (following O&W) we say that Q is *settled*. Notice again that, under this definition, complete answers are a special case of partial answers. Hence, a question Q is rhetorical with respect to $\underline{\mathrm{K}}$ whenever $\underline{\mathrm{K}}$ entails some partial answer to Q.

Following O&W (2006, p. 172), let $Q/_{\underline{\mathrm{K}}}a$, the $\underline{\mathrm{K}}$-truncation of Q by a, denote the set: $Q/^{a}_{\underline{\mathrm{K}}} = \{b \in Q : \neg b \notin \mathrm{Cn}(\underline{\mathrm{K}} \cup \{a\})\}$. It is immediate that $Q/_{\underline{\mathrm{K}}}a \neq Q$ iff $\underline{\mathrm{K}} + a$ partially answers Q. Truncation is instrumental to the definition of agenda updating upon expansion. One obtains $Ag(\underline{\mathrm{K}} + a)$ substituting to each question in $Ag(\underline{\mathrm{K}})$ its $\underline{\mathrm{K}}$-truncation by a (see next section).

In the original paper, no dual operation is defined, to be applied upon *contraction* of a corpus. We will address the topic of contraction, yet without defining any 'dual' operation. We will instead propose to consider procedures to update questions in $Ag(\underline{\mathrm{K}})$, and hence $Ag(\underline{\mathrm{K}})$ itself, upon contraction of $\underline{\mathrm{K}}$, as possible choices to update a *questioning strategy*. Before so doing, one needs some basic results about questions, and a clear assessment of the possible effects of contraction on agendas.

10.3.2 Agenda Continuity Through Change

The informal idea behind our proposal is easy to understand combining the game-theoretic framework of the interrogative model with the BRT framework used by O&W. Consider a game the goal of which is to form some new belief (or to choose between possible expansions represented by the set of answers to some question). A $\underline{\mathrm{K}}$-question is so to speak calculated to bring about the situation in which this (one of these) new belief(s) is formed. If, during the game, some information is received which changes the epistemic situation, it may affect the initial question. This information may be a response to some interrogative move, but it can be also added for other reasons as well. Let's first discuss these kind of changes.

First, the problem of questions updating is obviously related to the case of seemingly irrelevant answers such as Charles Napier's, or the one of Recanati's example. In both cases, the information is different from what was expected, and calls for yet another move in the game: deductive (given suitable background knowledge) in the former case, and interrogative in the latter (most likely to be responded by Inquirer herself through a guess, the justification of which may be found, once again, in the background knowledge). It can also happen that an answer to some question lead to revise another, previously unrelated, belief: for assuming that Sir Napier had no knowledge of Latin, then learning what his answer was may lead to change one's belief.

The case of *expansion* is easily solved by truncation. The postulate proposed by O&W is the following:

$$Ag(\underline{\mathrm{K}} + a) = \{Q' : Q' = Q/_{\underline{\mathrm{K}}}a \text{ for some } Q \in Ag(\underline{\mathrm{K}})\}^{29} \tag{3}$$

Truncation, and postulate (3), guarantee some continuity through expansion, since if Q is a $\underline{\mathrm{K}}$-question, then $Q/_{\underline{\mathrm{K}}}a$ is a $\underline{\mathrm{K}} + a$-question.[30]

On the other hand, since contraction never solves questions, it seems a natural constraint (or a natural tentative postulate for agenda continuity) to impose that a $\underline{\mathrm{K}}$-question inscribed on the $\underline{\mathrm{K}}$-agenda should be inscribed on any agenda associated with a contraction of $\underline{\mathrm{K}}$, i.e.:

$$Ag(\underline{\mathrm{K}}) \subseteq Ag(\underline{\mathrm{K}} \div a) \tag{4}$$

Unfortunately, this cannot be done, as shown by O&W's own example, which illustrates the possible failure of 'exhaustiveness' precondition:

> [...] the question whether god is evil or benevolent presupposes that there either exist a god that is evil or one that is benevolent. Take away the belief that god exists and the presupposition of the question will be removed, leaving the question itself hanging in the air. (Olsson and Westlund, 2006, p. 174)

Counterexamples where 'exclusiveness' precondition fails are easily constructed too. However, *yes-or-no*–questions cannot have their preconditions weakened by contraction. They have no non-trivial presuppositions, since their presuppositions are always satisfied (as long as the underlying logic remains classical, which we assume). Indeed, for any $\underline{\mathrm{K}}$ and a, $\{a, \neg a\}$ is either a $\underline{\mathrm{K}}$-question, or is settled with respect to $\underline{\mathrm{K}}$.[31]

This simple fact turns out to be of special importance in view of a result proved by Hintikka and his associates, the so-called *Yes-No Theorem*, which states that any conclusion which follows from a set E of premises together with answers to arbitrary questions, follows from E together with answers to *yes-or-no*–questions only.[32]

[29] Cf. Olsson and Westlund (2006, p. 172).

[30] See Olsson and Westlund (2006 p. 172), proof given n. 10.

[31] By classical logic, $(a \veebar \neg a) \in \underline{\mathrm{K}}$ whenever $\underline{\mathrm{K}}$ is consistent. (If is inconsistent, then $\underline{\mathrm{K}} = Ł$, hence $\{a, \neg a\} \cap \underline{\mathrm{K}} ne \varnothing$ and $\{a, \neg a\}$ is trivially settled.) If $a \in \underline{\mathrm{K}}$, then $\{a, \neg a\}$ is not a $\underline{\mathrm{K}}$-question since $\{a\} \subset \{a, \neg a\}$, but then it is settled. The same holds with $\neg a$.

[32] The theorem is proved in Hintikka et al. (1999, p. 55), and its philosophical consequences are developed in Hintikka (2007c). A formal version for corpora and agendas is given in Genot (2009). It requires the use of the equivalent of the *cut rule* of proof theory (see n. 20). Hence, for any procedure using *yes-or-no*–question–such as those we will introduce for question update—the question whether the procedure admits of cut elimination or not will be of considerable interest (and in our case, a topic of further research).

Hence, to each $\underline{K}$-question Q can be associated a set of *yes-or-no*–questions which (given $\underline{K}$) will ‘do the same job’ (epistemically speaking) as Q does. Let’s refer to this set as $|^{Y}_{N}\text{-}Q|$. Now, the following general fact about *yes-or-no*–questions holds:

$$\text{For all } \underline{K},\ a \text{ and } b,\ \text{if } \{a, \neg a\} \in \mathrm{Q}_{\underline{K}},\ \text{then } \{a,\ \neg a\} \in \mathrm{Q}_{\underline{K \div b}}\text{[33]} \tag{5}$$

As expressed by (5), any *yes-or-no*–question open with respect to a given corpus $\underline{K}$, remains open with respect to any contraction of $\underline{K}$. And this will also hold for *yes-or-no*–reductions, and as a special case we have:

$$\text{For any } a,\ Q \text{ and } \underline{K},\ \text{if } Q \in Ag(\underline{K}),\ \text{then } \left|^{Y}_{N} - Q\right| \subseteq \mathrm{Q}_{\underline{K}\div a}\text{[34]} \tag{6}$$

In the light of (6), it could be proposed, for any $\underline{K}$ and a, to include (at least) in $Ag(\underline{K} \div a)$ any question in $Ag(\underline{K})$ which is still a $\underline{K} \div a$-question, and for those Q in $Ag(\underline{K})$ which are not, to add any *yes-or-no*–question in $|^{Y}_{N}\text{-}Q|$ to $Ag(\underline{K} \div a)$. This would yield the following postulate:

$$\begin{aligned} Ag(\underline{K} \div a) \subseteq {} & (Ag(\underline{K}) \cap \mathrm{Q}_{\underline{K}\div a}) \\ & \cup\{Q' : Q' \in \|^{Y}_{N}\text{-}Q\| \text{ for some } Q \in Ag(\underline{K}) \text{ and } Q \notin \mathrm{Q}_{\underline{K}\div a}\} \end{aligned} \tag{7}$$

However, it is possible to do even better, updating a question having lost one or both of its preconditions by transforming it.

10.3.3 Updating Questions: Preliminaries

For some $\underline{K}$-question Q, a contraction of $\underline{K}$ by some a affecting Q may have one of the following outcomes: (i) alternatives in Q are no longer exhaustive; (ii) alternatives in Q are no longer exclusive; or (iii) both (i) and (ii). Let us consider only the two former cases, and let $\underline{K}' = \underline{K} \div a$. In case (*i*), Q cannot be asked before some $Q' \in \mathrm{Q}_{\underline{K}'}$, $Q \subset Q'$ has been partially answered by some b in such a way that $Q'/_{\underline{K}'}b = Q$. Case (ii) is slightly more complex: with respect to $\underline{K}$, if Q receives an answer, this answer *narrows* the range of alternatives, *by at least one* if it is partial, and *to at most one* if it is complete. But with respect to $\underline{K}$’, a complete answer is not enough to narrow the range of alternatives to at most one, since another may be compatible with the one known (from the answer) to hold.

[33] It is easily proved by contraposition: assume that $\{a, \neg a\}$ is not a $\underline{K} \div b$-question. Then either $\underline{K} \div b$ entails a or $\neg a$. Assume the former: since, by Inclusion postulate for contraction, $\underline{K} \div b \subseteq \underline{K}$, we have (by Closure) $a \in \underline{K}$. Hence Q is not a $\underline{K}$-question. A symmetric conclusion follows from the assumption that $\neg a \in \underline{K} \div b$.

[34] By definition of a $\underline{K}$-question and of a $\underline{K}$-agenda, and construction of $|^{Y}_{N}\text{-}Q|$, if $Q \in Ag(\underline{K})$, then $\|^{Y}_{N}\text{-}Q\| \subseteq \mathrm{Q}_{\underline{K}}$, which, combined with (5), yields that $\|^{Y}_{N}\text{-}Q\| \subseteq \mathrm{Q}_{\underline{K}\div a}$ as desired.

Using *yes-or-no*–question, the problem raised is made simpler to analyze. Let's first say that for some $\{b, \neg b\} \in \|^{Y}_{N}\text{-}Q\|$, b is a *positive* answer if $b \in Q$. Hence, b is positive iff any reply which settles $\{b, \neg b\}$ in $\{b\}$ also settles Q (and negative if it only helps truncate Q by one). In case (*i*), at least one more *yes-or-no*–question is needed to cover an exhaustive set of exclusive alternatives, since every $\{b, \neg b\} \in \|^{Y}_{N}\text{-}Q\|$ may receive a negative answer, according to $\underline{\text{K}}$'. In case (*ii*) any positive answer to some $\{b_i, \neg b_i\} \in \|^{Y}_{N}\text{-}Q\|$ hardly settles Q, since some $\{b_j, \neg b_j\} \in \|^{Y}_{N}\text{-}Q\|$ may receive, according to $\underline{\text{K}}$', a positive answer too.

Consider a situation in which our inquirer, call her Alice, is lost in a foreign city, and wants to know where she is. She's got a small map grabbed at her hotel, not very detailed, but showing small pictures of remakable landmarks, monument and buildings. Alice can then try to find some vantage point and attempt to locate herself with respect to the recognizable landmarks. Depending on those she may see, and their relative position, she will be able to locate her position with respect to the city map. The 'principal' question is for her *Where am I?*, and it could be represented as a set of alternative positions on the map.[35]

In the above example, each landmark may be treated as a parameter. Alice could estimate where she could be, and frame some 'question' considering (a subset of) the set of all (observable) relative positions of the landmarks. Each configuration thus considered would correspond to one of those possible locations, currently indiscernible from her point of view. And the elements of this set would be jointly exhaustive (relative to Alice's estimate) and pairwise exclusive, thus corresponding to a research question the kind O&W consider. Alice could also try to find (identify) one of the landmarks, and once it is done, try for another, and 'triangulate' to obtain her position. The forme method gives her a 'search pattern' to answer her principal question, hence the kind of question represented by the O&W-like research question, while the second uses instrumental questions.[36]

[35] Obviously, not *every* point of the map may be taken an alternative, on pain of having not only an infinite set thereof, but an uncountable one as well. Moreover, in a situations like Alice's, the set of alternative locations usually will not even be fully specified (at least, not given a full attention) before some step of the search has been reached. The analogy with *yes-or-no*–questions will be conspicuous to everybody who as shared Alice's predicament, since a good way to find where one is, is first to divide the map in two, with some imaginary line meeting the point one comes from, or using grid of the map (if there is one) and try to find in which half one is, using some landmark. Once this is done, one can proceed to further divisions, and will eventually identify some *area*. The set of areas, though specified at the outset (by the grid of the map, or some imaginary equivalent one could 'picture'), is indeed used in what is equivalent to *yes-or-no*–questions.

[36] Each square delineated by the map grid may be associated with the different visual patterns of landmarks one could see standing somewhere in the corresponding area. Usually some orientation (aligning the map, and oneself, to the North) is used to narrow down the overwhelming number of alternate possibilities that may be associated to each square, depending on the precise point one stands at, the direction one faces, etc. It is then clear that, though one may have some way to define a set of alternatives, in many cases, one will not proceed to *actually define* it before taking some steps to narrow down the range of alternatives. One will rather proceed, as described above (see n. 35), with a sequence of questions, each excluding at least one more (set of) alternatives (with respect to some method to list or determine them).

This example shows that a $\underline{\text{K}}$-question (w.r.t. some $\underline{\text{K}}$) Q may be viewed either as a one-shot question, or as a blueprint for series of questions. For now, let's consider that the elements of the series are simply $|{}^{Y}_{N}\text{-}Q|$—though obviously some $Q' \in \|{}^{Y}_{N}\text{-}Q\|$ can itself give rise to a series of instrumental questions, and so on. Subsequently, 'updating' Q can be done in two ways: updating the one-shot question, or adding instrumental questions to the series. Knowing which potential answer to the one-shot question describes the actual situation is the *desideratum of the principal question Q*: if one describes inquiry as a 'game', then it is the ultimate goal of that game, and any epistemic utility is to be calculated with respect to this goal.

Game theory tells us that preferences of the players, determining utilities, are defined neither over the set of single moves—more properly, moves in isolation—nor over the set of action profiles of a single player (i.e. sequences of actions available to *one* player) through their expected utility, as in decision theory. Preferences are defined over the *set* of action profiles of *all players*. This means here that epistemic utility will be ascribed to some series of questions and answers (moves from Inquirer followed by moves from Oracle) depending on its bringing Inquirer closer to the desideratum. This last 'calculation' in turn depends on several factors—whether the question is answerable or not, how reliable is the source of answer, how specific the answer will be if obtained, etc.—some of which count as 'actions' of the other player, Oracle, while others can be modeled as 'random variable' affecting the action profiles of the players. In the latter case some probability space will have to be defined, as well as a function from the product of this space and the set of action profile to some set of outcome.[37]

Hence, $\underline{\text{K}}$-question can be considered as descriptions of *one-shot games*, or as one-shot descriptions of games involving several (interrogative as well as inferential) steps; *yes-or-no*-reductions, on the other hand, display (some of) the possible (interrogative) steps.

These last remarks cast a somewhat different light on the problem of agenda update. It may be viewed as a *strategic* problem of adjusting one's strategy to a new set of information in an evolving game. The next section examines how some game-theoretic concepts can be imported in the discussion, throw some light on the general problem of agenda updating, and help suggest some technical solution.

[37] For the notions of action profile, strategic games, etc., see Osborne and Rubinstein (1994, Chap. 2). In Alice's case, the possibility to recognize the landmarks on the basis of the pictures on the small map is one of those factors (since two different buildings my have a similar pictorial representation), and ultimately depends on previous choices of the map designers. Relying on the map as a source of answers (rather than relying on some native) requires the map to be treated, abstractly, as a player of the game. Since distances and how they affect recognition (the well-known Cartesian example of a squared tower appearing round from afar is a case in point) has to be taken into account, but does not depend on some foreseeable 'action', reliability of the visual system may be modeled through a probability measure.

Some of these factors may or may not affect the 'phrasing' of the principal question. In particular, one can ask a question while knowing that one lacks the resources to obtain a complete answer, or even before knowing which other 'players' one will face. They may also influence the choice of the first hypothesis to test, and this is one way to understand Peirce's notion of (epistemic) economy.

10.4 Updating Agendas as a Strategic Problem

10.4.1 Normal vs. Extensive Forms

As already mentioned, in game theory utilities are defined from preferences defined over the set of possible combinations of actions of all the players. A game can be represented through a game-matrix, displaying the end-states (combinations of) actions of players may lead to. These matrices are especially useful to study whether the game under scrutiny admits of some solution (i.e. whether there are equilibria, and which, if any, is optimal, etc.). In a two-player game, one player's choices will be assigned rows, the other's columns, and each cell will display a pair of utilities. This representation is specially suited for *one-shot games*, where players make their choices independently. 'One-shot' is here to be understood in a broad sense: it may be that a player's action (choice) involve several sub-choices, but as long as the consequence depends of the overall results—as long as the outcome is resolved once all these actions are done—the strategic form is appropriate. [38]

On the other hand, one might want to stress that players face different possible *choices* at each phase of the game, or emphasize the role of sequential choices of players—especially since they may depend on the knowledge one player has of the other's actions, preferences, knowledge, etc. Then, the game will be represented by some tree-like structure. Such representation is especially useful when one wants to represent players' knowledge of the state she's reached: several states may be indiscernible from the player's point of view (e.g. in some games with imperfect information) so that a player cannot tell at which node of the tree she is. Hence the dynamic representation offered by trees may display features which are hidden in the matrix (normal form) representation. This representation is equally important when several choices have a different outcome depending on their order—i.e. whenever there is some kind of dependency between moves.

We can now make the analogy with research questions and research agendas more precise. In one sense, asking $\underline{\mathrm{K}}$-questions is very much like playing a strategic game (or a game in normal form), while choosing to ask corresponding *yes-or-no*–questions results in some extensive game.[39] But this analogy holds only at certain conditions. It can be shown, for example, that the equivalence in epistemic effect (utility, payoff) for Inquirer with respect to some $\underline{\mathrm{K}}$-question Q, between asking

[38] The following matrix (borrowed from Osborne and Rubinstein 1994, p. 13) represent abstractly a two-player game where each player, Row and Column, has two possible actions: the set $\{T, B\}$ for Row, and the set $\{L, R\}$ for Column. Each pair of value is a pair of utilities, the first for player Row, the second for player Column. Clearly, utilities are associated to combinations of actions of both players. For example, x_1 is the value for Row of the outcome of (T, R) while its value for Column is x_2.

[39] This analogy can be pursued formally, and various correspondence result proved (see Genot 2009). However, we will limit our exposition to a mostly informal, and at most semi-formal, overview.

Q and asking all questions (save at least one) in $|^{Y}_{N}\text{-}Q|$ depends on the insensitivity of Oracle to the order in which the questions in $|^{Y}_{N}\text{-}Q|$ are asked.[40] As long as certain assumptions about sources (and about the game in general) can be made, $\underline{\text{K}}$-questions, through their *yes-or-no*–reductions, can be considered as 'normal-form blueprints' for *yes-or-no*–strategies in extensive interrogative games.[41]

Updating a question can then be thought of either as a process which has, as an input, a 'normal form' question, and produces a normal form question as an output (which is O&W's choice with truncation), or as a process having as input and output some (structured) sets of questions, or 'extensive form' questions. It clearly appears now that (7) was a kind of compromise between the two possibilities. Let us now define the two related forms.

10.4.2 Strategic Update of Questions

We have thus far stressed that O&W's $\underline{\text{K}}$-questions are (already) in some kind of normal form: potential answers are pairwise exclusive and jointly exhaustive. Obviously, such a form needs not be unique, as the following example shows:

Are you planning to serve fish, or chicken? (8a)

Are you planning to serve fish and no chicken, or chicken and no fish? (8b)

Assuming suitable BK, (8a) and (8b) are equivalent $\underline{\text{K}}$-questions.[42] Yet (8b) leaves far less implicit, and bears some resemblance with disjunctive normal forms common in propositional logic. In fact, the *Normal Form Theorem* can be put into service to show that for every *genuine* question Q with respect to some corpus $\underline{\text{K}}$,

[40] Let $Q = \{a_1, \ldots, a_n\}$, and let's assume that oracle will answer in the most direct and informative manner, that is either by some (strictly) partial answer: $(\neg a_i \wedge \cdots \wedge \neg a_j)$ (with $1 \leq i, j \leq n$), or some complete answer $a_i \in Q$. Then the result of asking Q and receiving an answer will be either $\underline{\text{K}} + a_i$ or $\underline{\text{K}} + (\neg a_i \wedge \cdots \wedge \neg a_j)$. The consequence is clearly that the result of asking every $\{a_i, \neg a_i\} \in \||^{Y}_{N}\text{-}Q\|$ will be the same if Oracle answers to each the same way as it answers to Q, that is iff the *union* of answers is equivalent either to $(\neg a_i \wedge \cdots \wedge \neg a_j)$ or to a_i, hence iff Oracle is insensitive to the order of questions (since union does not preserve ordering).

[41] Other assumptions, besides insensitivity to order and to formulation, may include the absence of a time-limit, the availability of all sources at any time (if Oracle is a 'team' player) at various stages of the game, etc. The reader is encouraged to imagine his or her own counterexamples, throwing in various constraints. Real-life inquiries (or crime novels and TV shows) abound of cases in which those assumptions *cannot* be made.

[42] Question equivalence is easily defined as *two-way inclusion*, with *question inclusion* defined as follows: a question Q *includes* a question Q' iff any expansion which answers the first also answers the second, that is if for all $a \in Q$, if $a \in \text{Cn}(\underline{\text{K}} \cup \{d\})$ (for some d), there is a $b \in Q'$ such that $b \in \text{Cn}(\underline{\text{K}} \cup \{d\})$. In particular, since the condition for inclusion holds when $d = a$ for some $a \in Q$, it is equivalent (by the Deduction property of Cn) to the condition that any potential answer to Q entails some potential answer to Q'. As a consequence, two equivalent questions have (pairwise) equivalent potential answers (proof left to the reader)

there is a $\underline{\mathrm{K}}$-questions Q' such that any answer to Q' entails an answer to Q.[43] Indeed, assume that one learns that the dinner may be a two-courses dinner, the rest of the background remaining the same. Then, fish and chicken are no longer exclusive. Then, one can transform (8a) or (8b) into the following:

$$\text{Are you planning to serve fish and no chicken, or chicken and no fish, or both chicken and fish?} \tag{9}$$

The update (8)–(9) generalizes into a way to deal with questions having lost their *exclusiveness* preconditions. Since for any $\underline{\mathrm{K}}$-question, answers are exclusive (with respect to $\underline{\mathrm{K}}$), the two following forms are equivalent:[44]

$$Q = \{a_1, \ldots, a_n\} \tag{10a}$$

$$Q' = \{(a_1 \wedge \neg a_2 \wedge \cdots \wedge \neg a_n), \ldots, (\neg a_1 \wedge \cdots \wedge \neg a_{n-1} \wedge a_n)\} \tag{10b}$$

If (only) exclusiveness is lost, say between a_i and a_j following a contraction of $\underline{\mathrm{K}}$ by some b, then the following update will give a $\underline{\mathrm{K}} \div b$-question:

$$Q' \cup \{(\neg a_1 \wedge \cdots \wedge a_i \wedge \cdots \wedge a_j \wedge \cdots \wedge \neg a_n)\} \tag{11}$$

Since exhaustiveness is dealt with adding answers to a (10a)-like form, and since other answers being made compatible by contraction only call for repetition of the procedure sketched above, these remarks can easily serve as a basis for a 'normalization' procedure to update $\underline{\mathrm{K}}$-questions in case of contraction.[45]

Together with truncation, this procedure is all that one needs to offer postulates for updating agendas in both cases of expansion and contraction. All there is to do is to add a postulate to the effect that *normalized* questions—i.e. adding potential answers opened through contraction and dealing with compatibilities the way suggested above—are to be substituted to the original question in the agenda corresponding to the new (contracted) corpus. Since this normalization is strategic (in the

[43] See Genot (2009) and Enquist and Olsson (2008). For the *Normal Form Theorem* see e.g. Smullyan (1968, p. 13). The result holds when using the procedure to obtain normal form up to disjunctions having as disjunct conjunctions of potential answers or their negations, without necessarily going up to disjunction of *literals*. Yet nothing prevents the definition (or the usefulness) of an *atomic* interrogative normal form. Enqvist and Olsson (2008) proposes a *Topic Strategy* for updating questions which allows to 'break down' initial potential answers and rearranging answer to obtain an new $\underline{\mathrm{K}}$-question which falls somewhere in between the State Description Strategy (mentioned earlier n. 45) and the full atomic interrogative normal form.

[44] Equivalence in the sense of n. 42 clearly holds, since for any $a_i \in Q$:

$$a_i \leftrightarrow (\neg a_1 \wedge \cdots \wedge a_i \wedge \cdots \wedge \neg a_n) \in \mathrm{Cn}(\underline{\mathrm{K}})$$

Hence each element of Q is equivalent to exactly one element of Q'.

[45] The procedure is detailed in Genot (2009), where it is also proved equivalent to the *State Description Strategy* presented in Enqvist and Olsson (2008).

sense of the strategic form of games), we propose to call it the *strategic normalization* of a question. If $Q \in \mathrm{Q}_{\underline{\mathrm{K}}}$, we denote $\underline{\mathrm{K}}normQ_{\underline{\mathrm{K}}\div a}$ the strategic normalization of Q with respect to $\underline{\mathrm{K}} \div a$, then a tentative postulate will be:

$$Ag(\underline{\mathrm{K}} \div a) \subseteq \left\{Q' : Q' = \|Q\|^{S}_{\underline{\mathrm{K}}\div a} \text{ for some } Q \in Ag(\underline{\mathrm{K}})\right\} \tag{12}$$

In particular, the procedure and the postulate guarantee that, if one contracts *non-vacuously* $\underline{\mathrm{K}}$ by some a, and then expands the result by a, then the agenda is recovered iff the belief set is, which translates formally:

$$\text{If } a \notin \underline{\mathrm{K}} \text{ then } Ag((\underline{\mathrm{K}} \div a) + a) = Ag(\underline{\mathrm{K}}) \text{ iff } (\underline{\mathrm{K}} \div a) + a = \underline{\mathrm{K}} \tag{13}$$

Since we have not presented the normalization procedure in full, we will not prove (13), but we will soon present another procedure, which will make the result more intuitive. In order to do so, we have to move to 'extensive' questions.

10.4.3 Extensive Update of Questions

Transforming a $\underline{\mathrm{K}}$-question into an 'extensive' question, and subsequently updating it, may even be intuitively clearer than using strategic forms. Let us return to our diner example. In order to know what will be served, the questioner could very well ask (at most) the two following questions:

Are you planning to serve chicken? (14a)

Are you planning to serve fish? (14b)

Assuming again suitable BK, asking one of (14a) or (14b) will suffice: a positive answer to one will exclude the possibility of the other receiving one, while a negative answer will (by Disjunctive Syllogism) entail a positive answer to the other. It is obviously possible to ask both, even if the second may appear to be rhetoric (given BK, plus previous answers): but one may 'delay', so to speak, expansion until after all answer are received (especially if one asks the same question to several sources). However, such a possibility relates to more 'specific' cases of inquiry games, while we are by now only interested in the more abstract and general features of such games. Learning that there may be a third choice (say, filet) would amount to add a new *yes-or-no*–question.

Now, what would be the consequence of learning that there is a third possibility, i.e. a two-courses dinner, while no change in $\underline{\mathrm{K}}$ occurs to the effect that only fish and chicken are the only possible choices? One has the option of asking the following question:

Are you planning a two-courses dinner? (15)

and then ask (14a) or (14b) if the answer is negative, or both if the answer is positive — and if it is not excluded that one of chicken or fish can be served twice, for if it is, a positive answer to (15) entails a positive answer to both (14a) and (14b). Another option is to ask (14a) *and* (14b), without asking (15). If one expects a positive answer to the latter, and assumes that chicken (fish) will not be served twice, the first option is more economical, in terms of number of questions asked. The second option, on the contrary, may be more cautious. Other inquiry-specific considerations may guide the choice between those options.[46]

In order to generalize these remarks, let us first use trees to illustrate the way some question $Q = \{a_1, \ldots, a_n\}$, through the set $|^Y_N\text{-}Q|$, can give rise to a questioning strategy, or rather a family of strategies.[47] The left-hand tree in Fig. 10.2 illustrates one of these strategies. Each node is a state the game can reach, with Inquirer's moves labeled 'I', and Oracle moves labeled 'O'. A question mark indicates an *interrogative* move from Inquirer. The presupposition has to follow from $\underline{\text{K}}$, together with information available at this node: whatever follows from $\underline{\text{K}}$ together with information added at preceding nodes of the same branch. The absence of question mark indicates that Inquirer draws an inference from $\underline{\text{K}}$ together with the preceding answers. The sign ø indicates that Inquirer has obtained a complete answer to $Q = \{a_1, \ldots, a_n\}$ (no further question is needed). The right-hand tree is simplified, omitting labels of players and interrogative moves.

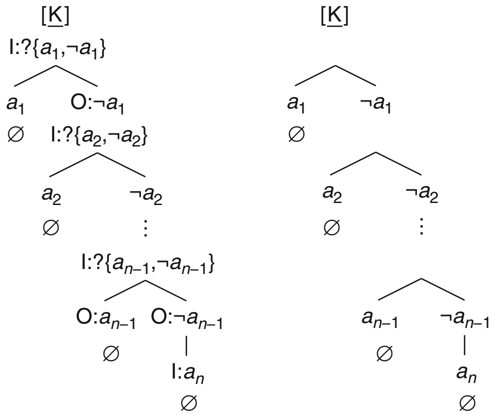


Fig. 10.2 Fig. 10.2 Tree form of $Q = \{a_1, \ldots, a_n\}$ (using $|^Y_N\text{-}Q|$) and its simplified form

[46] Consider a situation where a restaurant has been booked for the evening for some social event, with only one menu, and where a columnist of some celebrity magazine tries to know what the menu is, but cannot access directly the information. Assume that, for some reason, only the *chef* does know the whole menu, but the journalist can only ask the kitchen *commis*. Then asking (15) would be pointless, and he would have better results asking (14a) and (14b) to different *commis*—if the information he has already obtained allows him to restrict his range of questions to those two.

[47] If we assume that the order in which the questions are put to sources is irrelevant, then the family is a equivalence class: only the set of answers matters, not their sequence.

The overall tree corresponds to a strategy which can be expressed in plain words as (roughly): 'Keep asking questions in $|^{Y}_{N}\text{-}Q|$ as long as you receive negative answers, and stop as soon as you've received a positive one or asked the penultimate'. Since, given $\underline{\mathrm{K}}$ and the answers obtained at the preceding nodes, $\{a_n, \neg a_n\}$ would indeed be rhetoric, the last question is never asked.[48]

To handle loss of exhaustiveness following contraction of $\underline{\mathrm{K}}$ by some b, one can proceed by adding at the 'root' of the tree the *yes-or-no*–questions corresponding to the i newly opened alternatives. Figure 10.3 displays the general form of such an updated tree.

Handling loss of exclusiveness, without adding any new question to the list $|^{Y}_{N}\text{-}Q|$ (contrast this with to the solution suggested adding (15) to the list) can be done as follows. Assume that a_i and a_j are made compatible by some contraction of $\underline{\mathrm{K}}$ by b'. Then, Inquirer can arrange the list so that $\{a_i, \neg a_i\}$ is followed by $\{a_j, \neg a_j\}$, but if she obtains a_i as an answer to the former, she will nevertheless ask the latter in addition. Other compatibilities are easily dealt with repeating the procedure, which is illustrated in Fig. 10.4 for a_i and a_j.

Generalizing from these examples, one can obtain the basis for a second normalization procedure to update *yes-or-no*–strategies based on *yes-or-no*–reduced $\underline{\mathrm{K}}$-questions in case of contraction.[49] It is possible to define 'extensive questions' (ordered n-uple of questions corresponding to questioning strategies), and to consider agendas as sets of those questions. Let $Ag^e(\underline{\mathrm{K}})$ denote agendas of this kind, and Q^e the 'extensive' form of Q. If $Q \in \mathrm{Q}_{\underline{\mathrm{K}}}$, we denote $\|Q^e\|^E_{\underline{\mathrm{K}} \div a}$ the extensive normalization of Q^e with respect to $\underline{\mathrm{K}} \div a$, then a tentative postulate will be:

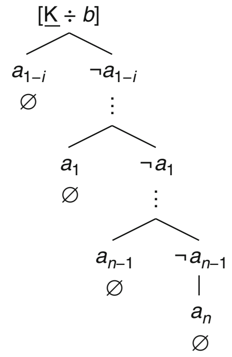
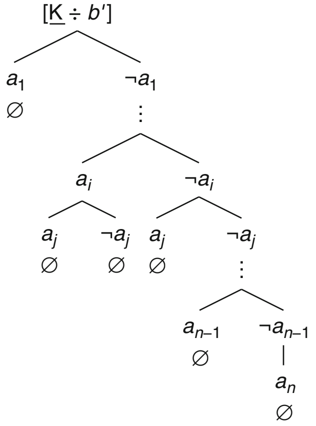


Fig. 10.3 Update of Q after loss of *exhaustiveness*

[48] It is easily checked that this strategy (or any other with another ordering) will indeed deliver an answer to $Q = \{a_1, \ldots, a_n\}$, if all questions in $|^{Y}_{N}\text{-}Q|$ are answerable—hence if Q is, since we assumed our source not to be sensitive to the way questions are put. It is obviously still assumed that the set Q is finite, so the strategy will not guarantee an answer to, say an 'existential' question about an infinite domain.

[49] The procedure is detailed in Genot (2009).

Fig. 10.4 Update of Q after loss of *exclusiveness* (between a_i and a_j)



$$Ag^e(\underline{\mathrm{K}} \div a) \subseteq \{Q' : Q' = \|Q^e\|^E_{\underline{\mathrm{K}} \div a} \text{ for some } Q^e \in Ag^e(\underline{\mathrm{K}})\} \tag{16}$$

It can be shown that there is a one-to-one correspondence between the *strategic* and the *extensive* normalization of a given question Q.[50] Hence, every result obtained with extensive questions can be translated to strategic questions. In particular, one will easily verify that (13) holds thanks to (16), since expansion will 'cut' the tree below the root of the original question, and prune the redundant branches.

10.5 Conclusion: From a Extensive Point of View

We would like to conclude with some programmatic remarks, extending the 'extensive' point of view that we have been defending to other problems related to belief change. We would stress that the distinction between 'one-shot' (strategic) games and 'repeated' (extensive) games permeates all the theory of belief and theory change. More accurately, just as we have argued that questions *à la* O&W should be viewed as 'blueprints' for extensive games (and indeed partial, underdetermined blueprints), we will argue that BRT, treating as functional dependency the relation between states, should be indeed viewed as a one-shot reconstruction of a multi-step process.

Recall the 'logic of pure discovery' briefly described in Section 10.2.2. A moment's reflection shows the following correspondence: if one begins a interrogative game with $\underline{\mathrm{K}}$ as a set of premises, and can indeed treat every answer as true, then, if we let A be the set of answers (assumed here to be finite, for simplicity sake) then the set of conclusions Inquirer could obtain or reach (by logical moves)

[50] See Genot (2009).

after the inquiry has been carried is Cn($\underline{\text{K}} \cup A$). The connection here is not very deep, but as it is it already raises an interesting question: what kind of process at the level of inquiry does give rise to *revision*? And if we can say that facts about expansion 'supervene' in some sense on facts about inquiry, can we say the same about revision?

Let us venture here a conjecture: the standard BRT treatment of this question is similar to the original treatment of research questions in O&W's theory: it treats as a 'one-shot' operation what is indeed an extensive process. We won't offer a full argument here, but simply illustrate our point with lost Alice.

Assume that Alice believes to have located her whereabouts, on the right half of the map. She thinks to have recognized (through the pictures) two remarkable buildings on the map, and have 'triangulated' her position. Unfortunately for her, the pictures are not very precise, and the scale inaccurate, and she has mistaken the two buildings for two others. But since she is cautious, she tries for another landmark which, according to the map, should be in sight if she is correct. Failing to see it, Alice realizes that she has fallen victim to a 'visual' *qui pro quo*.[51] What should Alice do?

According to one of the way to read the BRT story, Alice belief state (before revision) can be represented as a sphere system, in the style of the well-known sphere-based semantics for counterfactuals, introduced by David Lewis. The innermost sphere includes the possible worlds compatible with what she takes for granted (her current belief set), and the successively larger spheres surrounding it, the belief-contravening possible worlds. Spheres are ordered by some 'epistemic preference' ordering relation: the farther you go from the innermost sphere, the more 'favored' beliefs you have to relinquish.[52]

Receiving some belief-contravening information, and assuming that it has some propositional content a, Alice's reasoning can be represented as a two-step process of: (i) first 'falling back' to the closest a-permitting (or a-compatible) sphere, a step corresponding to contraction (by $\neg a$); and (ii) then expand with a from there, i.e. form the intersection of the contracted sphere with a, and take this set as the new belief sphere, the innermost of a new system. A re-ordering of worlds will also result from this change, too.

Now, according to good old-fashioned common sense, Alice should probably try to identify properly the two buildings she has misidentified, possibly using the pictures on some other part of the map. She could try to locate on the map the landmark she has just noticed, checking whether some picture may correspond to

[51] The phrase is taken here in the sense of blundered substitution, rather than in its legal sense (the former is common in Latin country, and acknowledged in English too, though according to the *Oxford Concise Dictionary, 4th ed. 1950* rarely so used).

[52] Use of this 'topological' semantics is common since Grove's seminal paper (1988), which has started a semantic undercurrent working with possible-worlds modelling within the AGM-tradition. We choose it because of its relation with ordinary (relational) possible worlds semantics, but as is well-known, Grove proved its equivalence with other constructions, using entrenchments or selection functions.

it, too. Anyone who has been in such a predicament will have a story of their own, paralleling Alice's. With some technical ingenuity, this process could be represented as an interrogative game. And the output of such a process will be a 'fall back': a set of places Alice could be, consistent with the data she has taken into account (which may or may not be the whole set of 'answers' to her questions), to be substituted to the one she previously (and mistakenly) believed to have been identified.

The conclusion is almost immediate: BRT relates the input (the belief state together with the new information) to the output,[53] through a one-step process, while it could—and if we are correct, should—be reconstructed as a multi-stepped process. A more complete and realistic model of belief change should try to explain how agents build their 'fall back theories', so to speak, on the fly.

Yet describing this process itself is not without challenges. If, for example, one adopts the view that actual change occurs only when inquiry has been carried to a term—according to the inquirer's opinion, or some preset standard—then how should one describe the changes occurring *within* inquiry? How they relate to changes in the overall belief structure is an challenging question, to which the answer may lead to reconsider several traditional questions related to BRT such as the very existence of 'routine expansion', of expansion into inconsistency, or the status of 'input assimilation', etc.

Accepting information and using it—even possibly inconsistent information, since information, unlike knowledge, need not be true—is not the same as believing it. Hence, one can delay credence, or belief, until inquiry has been carried far enough. As Hintikka reminds us, the well-known fictional detective Maigret, if asked about his beliefs about the case at hand, generally answers that he believes nothing, for 'the moment for believing or not believing has not yet come'.[54]

As a simple example, the priority to input in standard BRT becomes a triviality in a multi-layer model including some 'information-based' inquiry level: expansion, or revision, will *always* satisfy the success postulates, since the assimilation of input indicates that inquiry has done the job of sorting out what is believable and what is not. As Hintikka puts it:

> It is not that Maigret has not carried his investigation far enough to be in a position to know something. He does not know enough to form a belief. In serious inquiry *belief, too, is a matter whether an inquiry has reached far enough*. (Hintikka 2007a p. 32), emphasis added.)

From the BRT side, the interrogative model has been almost completely ignored, while from the IMI point of view, belief revision has been criticized, sometimes even harshly.[55] If the way we propose to explore is as promising as we think it is, then

[53] Or the set of possible outputs, in the case of relational approaches to BRT (originated in Lindstrom and Rabinowicz 1991).

[54] G. Simenon, *Maigret and the Pickpocket*, quoted in Hintikka (2007a, p. 32).

[55] Belief revision theory has not been as much a target of Hintikka and his associates as non-monotonic logics have. However, on the one hand, they are "two faces of the same coin" (to borrow D. Makinson's well known image), and on the other, the criticism addressed by Hintikka to the

studying the interplay between BRT and the IMI is not only suitable, but necessary to build a more complete and realistic model of belief change.

Erik J. Olsson once expressed the belief that in the best of all possible worlds, Hinitkka's IMI could be merged with his theory of belief revision *cum* agendas.[56] We hope to have shown not only that this world is possible, but also why it may indeed be the best we formal epistemologists could leave in, and that it is in our power to make it actual.

References

Enqvist, Sebastian, and Erik J. Olsson. 2008. Contraction in interrogative belief revision. Unpublished manuscript.

Emmanuel J. Genot. 2009. Extensive questions. In *Logic and its applications, third indian conference, ICLA 2009, Chennai, India, January 7–11, 2009. Proceedings.* Berlin: Springer.

Grove, Adam. 1988. Two modellings for theory change. *Journal of Philosophical Logic* 17: 157–170.

Harris, Stephen. 1994. GTS and interrogative taleaux. *Synthese* 99:329–343.

Hintikka, Jaakko. 1987a. The fallacy of fallacies. *Argumentation* 1(3):211–238.

Hintikka, Jaakko. 1987b. The interrogative approach to inquiry and probabilistic inference. *Erkenntnis* 26(3):429–442.

Hintikka, Jaakko. 1988. What is abduction? The fundamental problem of contemporary epistemology. *Transactions of the Charles S. Peirce Society* 34:503–533.

Hintikka, Jaakko. 1997. What was aristotle doing in his early logic, anyway? *Synthese* 113: 241–249.

Hintikka, Jaakko. 1999. *Inquiry as inquiry: A logic of scientific discovery*. Dordrecht: Kluwer.

Hintikka, Jaakko. 2003. A second generation epistemic logic and its general signifiance. In *Knowledge contributors*, eds. V.F. Hendricks, K.F. Jørgensen, and S.A. Pedersen, chapter 3, 33–56. Dordrecht: Kluwer, Reprinted in Jaakko Hintikka (2007d), chap. 3.

Hintikka, Jaakko. 2004. A fallacious fallacy?. *Synthese* 140:25–35. Reprinted in Jaakko Hintikka (2007d), chap. 9.

Hintikka, Jaakko. 2007a. Epistemology without belief and without knowledge. In *Socratic epistemology*, chapter 1, 11–37. Cambridge: Cambridge University Press.

Hintikka, Jaakko. 2007b. Logical explanations. In *Socratic epistemology*, chapter 7, 161–188. Cambridge: Cambridge University Press.

Hinitkka, Jaakko. 2007c. Presuppositions and other limitations of inquiry. In *Socratic epistemology*, chapter 4, 83–107. Cambridge: Cambridge University Press.

Hintikka, Jaakko. 2007d. *Socratic epistemology*. Cambridge: Cambridge University Press.

Hintikka, Jaakko, and Ilpo Halonen. 1999. Interpolation as explanation. *Philosophy of Science* 66(3):414–423.

latter holds, *mutatis mutandis* against the former: "[...] in the ultimate epistemological perspective [ampliative logics] are but types of enthymemic reasoning, relying on tacit premises [...]. An epistemologist's primary task here is not to study the technicalities of such modes of reasoning, fascinating though they are in their own right. It is to uncover the tacit premises on which such euthymemic reasoning is in reality predicated on." (Hintikka, 2007a, p. 21). The equivalent of "tacit premises" in BRT are, if we are correct, the entrenchment, or ordering of 'fall back theories', of which one of the tasks of (formal) epistemology should be to study the *formation*.

[56] In a talk given in Lille, 23rd of January 2007, which provided the initial motivation of the present work.

Hintikka, Jaakko, Ilpo Halonen, and Arto Mutanen. 1999. Interrogative logic as a general theory of reasoning. In *Inquiry as inquiry: A logic of scientific discovery*, 47–90. Dordrecht: Kluwer.

Levi, Isaac. 1991. *The fixation of belief and its undoing*. Cambridge: Cambridge University Press.

Lewis, David K. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic* 8(3): 339–359.

Lindström, Sten, and Wlodek Rabinowicz. 1991. Epistemic entrenchment with incomparabilities and relational belief revision. In *The logic of theory change*, eds. A. Fuhrmann and M. Morreau, 93–126. Berlin: Springer.

Mackinson, David. 2005. *Bridges from classical to nonmonotonic Logic*, volume 5 of *texts in computing*. London: King's College Publications.

Olsson, Erik J., and David Westlund. 2006. On the role of research agenda in epistemic change. *Erkenntnis* 65:165–183.

Olsson, Erik J. Belief revision and agenda management. This volume, XX–XX.

Osborne, Martin J., and Ariel Rubinstein. 1994. *A course in game theory*. Cambridge, MA: MIT Press.

Peirce, Charles S. 1940. *The philosophy of Peirce: Selected writings*. Routledge.

Rahman, Shahid, and Laurent Keiff. 2005. On how to be a dialogician. In *Logic, thought and action*, ed. D. Vanderveken, LEUS. Dordrecht: Kluwer.

Rahman, Shahid, and Tero Tulenheimo. 2007. From games to dialogues and back. In *Logic, games and philosophy: Foundational perspectives*, eds. Ondrej Majer, Ahti Pietarinen, and Tero Tulenheimo. Berlin: Springer.

Recanati, François. 2001. What is said. *Synthese* 128:75–91.

Robinson, Richard. 1971. Begging the question 1971. *Analysis* 31:113–117.

Smullyan, Raymond. 1968. *First-order logic*. Berlin: Spinger.

van Fraassen, Bas C. 1980. *The scientific image*. Oxford: Oxford University Press.

Chapter 11
Functional vs. Relational Approaches to Belief Revision

Erik J. Olsson

11.1 Introduction

This chapter addresses a central issue in the study of theoretical rationality: how much of science is governed by such rationality and how much is rather a matter of taste or arbitrary choice?[1] The issue has been the concern of a long-standing debate between a functionalist and a relationalist approach in the part of philosophical logic devoted to the logic of belief revision.[2] Roughly, the functionalist holds that rationality uniquely determines the outcome of a given revision process. Relationalists disagree. According to them, considerations of theoretical rationality alone do not yield a unique recommendation for how to change one's view. Such considerations can only delineate a set of possible results among which the inquirer must ultimately choose without recourse to theoretical reason.

In my view, what has driven the debate is partly a lack of conceptual clarity, a claim to be substantiated as we proceed. In an effort to improve on the present state of affairs I shall differentiate between three different ways of drawing the line between functionalism and relationalism. How these distinctions bear on the functionalist-relationalist controversy will be the subject matter of most of the present essay.

By *strong functionalism* will be meant the view that the agent's old beliefs and the new datum are together sufficient rationally to determine the agent's new belief state after revision. The denial will be referred to as *weak relationalism*. Thus weak

E.J. Olsson (✉)
Lund University, Lund, Sweden
e-mail: Erik_J.Olsson@fil.lu.se

[1] For an overview of the history of this subject, see Stölzner (2000). Otto Neurath seems to have been the first to suggest, in his critiscism of Descartes dating back to 1913, that in practical as well as in theoretical matters decisions must be made even if thorough investigation terminates in a set of equally reasonable alternatives. Neurath suggests basing such decisions on an "auxiliary motive", that is, ultimately by tossing a coin. For his discussion of Descartes, see Neurath (1981).

[2] See for example Lindström and Rabinowicz (1989, 1991), Doyle (1991), Galliers (1992), Rabinowicz and Lindström (1994), Hansson (1998), Levi (2004) and Tennant (2006a, b).

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_11,
© Springer Science+Business Media B.V. 2011

relationalism is the view that what can be said about rationality exclusively in terms of the old beliefs and the new datum is insufficient to single out a new state of belief.

Weak relationalism is compatible with the position that the new belief state after belief change is rationally determined given the new datum, the prior belief state and other relevant features of the agent's cognitive makeup. I will refer to this position as *weak functionalism*. The denial of weak functionalism will be called *strong relationalism*. Thus a strong relationalist holds that there is no set of cognitive features that, in conjunction with the old beliefs and the new datum, suffices to determine, in all cases, a unique new belief state after revision.

Finally, a given theory of belief revision is *relation-based* if its central primitive concept is that of a belief revision relation that is not a function. In general, a belief revision relation relates an old belief set and an input to several rationally admissible revised belief sets. A belief revision theory that takes as its chief primitive notion a belief revision relation that is a function is said to be *function-based*.

Based on these distinctions, eight combinations are prima facie possible (see Table 11.1). For instance, the position wFwRR amounts to a *w*eakly *F*unctional and *w*eakly *R*elational *R*elation-based theory, i.e. a relation-based theory to the effect that while the new belief state resulting from revision is not rationally determined by the old beliefs and the new datum alone (weak relationalism), it is thus determined once other cognitive aspects are taken into account (weak functionalism).

Table 11.1 The prima facie possible overall positions in the debate between functionalists and relationalists

Abbreviation	Result of belief revision determined all things considered?	Result of belief revision determined by old beliefs and new input?	Primitive concept a function?
wFsFF	Yes (weak functionalism)	Yes (strong functionalism)	Yes (function-based)
wFsFR	Yes (weak functionalism)	Yes (strong functionalism)	No (relation-based)
wFwRF	Yes (weak functionalism)	No (weak relationalism)	Yes (function-based)
wFwRR	Yes (weak functionalism)	No (weak relationalism)	No (relation-based)
sRsFF	No (strong relationalism)	Yes (strong functionalism)	Yes (function-based)
sRsFR	No (strong relationalism)	Yes (strong functionalism)	No (relation-based)
sRwRF	No (strong relationalism)	No (weak relationalism)	Yes (function-based)
sRwRR	No (strong relationalism)	No (weak relationalism)	No (relation-based)

Two of these positions can be excluded on purely logical grounds. If the belief revision output is uniquely determined by the old beliefs and the input alone, then clearly it is also uniquely determined once more aspects of the agent's cognitive states are taken into account. In other words, strong functionalism entails weak functionalism and hence also the negation of the opposite position, viz. strong relationalism. Therefore, sRsFF and sRsFR are contradictory.

Indeed, strong functionalism, the view that the old belief state and the new datum are sufficient to determine the new belief state after revision, is a philosophically dubious position. It is part of belief revision folklore that "logical" properties alone are insufficient to characterize the revision process which is taken to involve various extra-logical features of a cognitive state, such as entrenchment, plausibility, rational choice or assessments of informational value. For that reason, we can exclude all positions that involve a commitment to strong functionalism, i.e., not only sRsFF and sRsFR but also wFsFF and wFsFR.

This leaves us with four serious contenders: wFwRF, wFwRR, sRwRF and sRwRR. They have in common a commitment to weak relationalism, i.e. to the view that the old beliefs and the new datum alone are insufficient to determine rationally the revised state of belief. Positions wFwRF and wFwRR involve weak functionalism as well. Positions sRwRF and sRwRR do not. Positions wFwRF and sRwRF are function-based. Positions wFwRR and sRwRR are not.

The next task will be to assess relationalist arguments to the effect that it is incoherent to advance weak or strong relationalism within a function-based framework. This, if true, would lead to the further exclusion of wFwRF and sRwRR, leaving only wFwRR and sRwRR still open. The latter represents a "fully relational" position which is not only strongly (and hence also weakly) relational, but also relation-based.

11.2 Are wFwRF and sRwRF Incoherent?

The most well-known advocates of a relation-based approach to belief revision are Sten Lindström and Wlodek Rabinowicz (L&R) and Neil Tennant.[3] They motivate their view essentially by referring to the basic relationalist intuition that the result of belief revision is not determined by rationality alone. According to L&R, the main idea behind their theory is "to allow for their being several equally reasonable revisions of a theory with a given proposition" (1994, p. 69). Tennant also stresses the importance to "countenance variety". As he puts it, "a theory of relational theory change should be able to furnish such variety, by treating contraction and revision more generally as relational, not functional, notions" (2006a, p. 490), the implication being that function-based theories fail to countenance variety.

[3] Lindström and Rabinowicz (1989, 1991), Rabinowicz and Lindström (1994), and Tennant (2006a, b).

Neither L&R nor Tennant makes a clear distinction between strong and weak relationalism. As a consequence, it is unclear whether what they find objectionable is already the combination of function-basedness and weak relationalism or merely the combination of function-basedness and strong relationalism, i.e. whether they wish to rule out wFwRF and hence also sRwRF, or only the latter.

My first observation will be that these two positions are both at least consistent. This will be shown by example. The AGM theory of Alchourrón, Gärdenfors and Makinson (1985) will be seen to instantiate position wFwRF, and – as Tennant et al. would agree – there is no reason to think to think that AGM should be considered inconsistent, ill-defined or otherwise internally defective. A theory instantiating position sRwRF can be obtained by slightly modifying the AGM theory.

In the AGM theory, a belief state is represented as a logically closed set of sentences, called a belief set. There are three principal types of belief change: expansion, revision and contraction. In expansion a new belief is added without any old beliefs being given up. In revision, the new information is added in a way that preserves consistency. Even if the new information is inconsistent with the original belief set, revision guarantees that the new belief set is consistent, provided that the information is itself non-contradictory. Finally, to contract a belief means to remove it from the belief set.

In the AGM theory, these types of changes are conceptualized as functions or "operations" from belief states to belief states. Thus AGM theory is function-based. The simplest of the three operations, expansion of a belief set K with a sentence α, denoted $K + \alpha$, is defined as the logical closure of the union of K and $\{\alpha\}$, i.e., $K + \alpha = Cn(K \cup \{\alpha\})$. Closing under logical consequence ensures that the result of expansion is indeed a belief set.

It is less obvious how to define the more interesting notion of genuine revision. As a preliminary, the AGM trio argued that a reasonable revision operation, denoted *, should satisfy eight so-called rationality postulates:

$(\mathrm{K}^*1) K^*\alpha = Cn(K^*\alpha)$

$(\mathrm{K}^*2)\ \alpha \in K^*\alpha$

$(\mathrm{K}^*3) K^*\alpha \subseteq Cn(K \cup \{\alpha\})$

(K^*4) If $\neg\alpha \notin K$, then $Cn(K \cup \{\alpha\} \subseteq K^*\alpha$

$(\mathrm{K}^*5) K^*\alpha - K_\perp$ if and only if $\vdash \neg\alpha$

(K^*6) If $\vdash \alpha \leftrightarrow \beta$, then $K^*\alpha = K^*\beta$

$(\mathrm{K}^*7) K^*\alpha \wedge \beta \subseteq (K^*\alpha) + \beta$

(K^*8) If $\neg\beta \notin K^*\alpha$, then $(K^*\alpha) + \beta \subseteq K^*\alpha \wedge \beta$

The revision postulates are intended to capture the intuition that revisions should be, in a sense, minimal changes so that information is neither lost nor gained without compelling reasons. Thus changes in belief should obey a principle of informational economy. As Gärdenfors puts it, "the main thrust of the criterion of informational economy is that the revision of a belief set not be greater than what is necessary in order to accept the epistemic input" (1988, p. 53). According to K*3, for instance, $K^*\alpha$ must not contain more information than what is included in $Cn(K \cup \{\alpha\})$.

According to AGM theory, the postulates for revision together express all that can rationally be said about the revision result exclusively in terms of the old beliefs and the new datum. Technically they delimit a class of total functions from **K**×**L** to **K**, where **K** is the set of all logically closed sets of L-sentences, representing the class of rationally admissible revision functions. As is well known, the AGM revision postulates are together indeed insufficient to single out a single belief revision function as uniquely rational: the class of revision functions that satisfy the revision postulates contains more than one element.[4]

The fact that the AGM postulates fail to determine a unique belief revision function means that AGM is weakly relationalist. For the old belief set and the new datum are together insufficient to determine rationally the new belief set, which will also depend on what revision function is being employed. For the record, this is a consequence that Gärdenfors welcomes: "[t]he postulates (K*1)–(K*8) do not uniquely characterize the revision $K^{*}\alpha$ in terms of only K and α. This is, however, as it should be. I believe it would be a mistake to expect that only logical properties are sufficient to characterize the revision process." (1992, p. 11, notation and spelling adapted)

It is also part of the AGM theory that a unique revision result is guaranteed provided that more features of the agent's cognitive apparatus are taken into account than just the old beliefs and the new input. To see why this is so, first note that revision can be reduced to contraction followed by expansion. In order to revise K by α, first contract by $\neg\alpha$ and then expand the result by α, where the contraction is performed for the purpose of making room for α. This proposal is codified in the so-called Levi identity: $K^{*}\alpha = (K \div \neg\alpha) + \alpha$ (where $\div$ denotes contraction).

Now since expansion is already functionally defined, the Levi identity reduces the problem of how to define a revision function to the problem of how to define a contraction function. The basic mechanism for this purpose in the AGM model is that of partial meet contraction, as defined by the following identity: $K \div \alpha = \cap\gamma(K\bot\alpha)$. Here $K\bot\alpha$ is the set of all inclusion-maximal subsets of K that do not imply α, and γ is a selection function such that $\gamma(K\bot\alpha)$ is a non-empty subset of $K\bot\alpha$, unless the latter is empty, in which case $\gamma(K\bot\alpha) = \{K\}$. The intuition behind the use of the selection function is that it should select the "best" elements of $K\bot\alpha$ according to the agent's theoretical preferences, which can be taken to be part of the agent's cognitive constitution. Partial meet contraction amounts, then, to taking as the new state of belief what is common to the best maximal subsets of K that do not imply α, i.e., to suspend judgment between these subsets. This recipe is bound to lead to a uniquely determined result.

In this way, all three types of belief change addressed by the AGM theory can be accounted for in such a way that the result of change is uniquely determined given a suitable portion of the agent's cognitive state, including not just the agent's old beliefs, but also her theoretical preferences. Thus AGM is not only a function-based weakly relational theory; it is also weakly functionalist.

[4] For a formal derivation, see Theorem 3 in Tennant (2006b).

The AGM theory allows for the possibility that the function selecting a set of “best” inclusion-maximal subsets of K that do not imply α returns a set with more than member. If this were the end of the story and the agent were free to choose as he pleases between those members, we would have a strongly (and hence also weakly) relational function-based account of belief change, i.e. a theory instantiating position sRwRF.

What makes AGM weakly functional is the invocation of judgment-suspension as a tie-breaking rule. This is a powerful and intuitive strategy that seems eminently rational: what we ought to believe when confronted with several alternative theories or states of belief equally worthy of choice, and more so than any other theory or state of belief, is precisely what they all have in common. Other weakly functionalist theories are also based on suspension of judgment as a tie-breaking strategy. Levi’s (1991) and Hansson’s (1991) approaches are cases in point. They, too, instantiate position wFwRF, i.e. they belong to the class of weakly functionalist and weakly relationalist approaches that are also function-based.

The upshot is that the internal consistency of wFwRF and sRwRF is hardly in doubt. Yet, our point of departure in this chapter was the observation that researchers with relationalist inclinations have been questioning the coherence of those positions, arguing that relationalism should, in the interest of avoiding internal tension, be combined with a relation-based rather than a function-based framework. It is time to take a closer look at their argumentation, if only to pinpoint where they err.

The relational-based approach was first introduced in Lindström and Rabinowicz (1989). In that paper, a belief revision relation is defined as a ternery relation $\mathbf{R} \subseteq \mathbf{K} \times \mathbf{Con} \times \mathbf{K}$ satisfying the following axioms for all belief sets A, B, C, and all consistent propositions α and β:

(R0) $(\exists D \in \mathbf{K})\, A\, \mathbf{R}_{\alpha} D$
(R1) If $A\, \mathbf{R}_{\alpha} B$, then $\alpha \in B$
(R2) If $A \cup \{\alpha\}$ is consistent and $A\, \mathbf{R}_{\alpha} B$, then $B = A + \alpha$
(R3) If $Cn(\{\alpha\}) = Cn(\{\beta\})$, then $A\, \mathbf{R}_{\alpha}\, B$ if and only if $A\, \mathbf{R}_{\beta} B$
(R4) If $A\, \mathbf{R}_{\alpha} B$, $B\, \mathbf{R}_{\beta} C$ and $B \cup \{\beta\}$ is consistent, then $A\, \mathbf{R}_{\alpha \wedge \beta}\, C$

As L&R read $A\, \mathbf{R}_{\alpha}\, B$, it means that B is a possible result for a given agent of revising A by the addition of α as a sole piece of new information. Axiom R0 expresses that belief revision should be defined for all belief sets A and consistent propositions α. Axioms R1-R4 mirror closely the AGM postulates for revision. Thus R3, corresponding to AGM postulate K*6 usually referred to as the postulate of *extensionality*, expresses that in revision only the logical content of the input sentence is important, not its syntactic formulation. A belief revision relation $\mathbf{R}$ is said to be functional if, in addition to $\mathrm{R0} - \mathrm{R4}$, it satisfies:

$$\text{(R5) If } A\, \mathbf{R}_{\alpha} B \text{ and } A\, \mathbf{R}_{\alpha} C, \text{ then } B = C$$

Tennant (2006a) presents a relation-based belief revision theory much along the same lines. He studies two primitive relational notions: $\downarrow (J,\ K,\ \alpha)$, meaning that

J is a contraction of *K* with respect to α, and ↑ (*J*, *K*, α) meaning that *J* is a revision of *K* with respect to α. Tennant's theory differs in some respects from L&R's treatment. Unlike L&R, Tennant argues for a principal case analysis of the relevant belief change relations. For revision, the principal case is where a consistent belief state *K* is revised with respect to a non-contradictory sentence α that is inconsistent with *K*. (L&R's relation is defined also for inconsistent belief sets and for new input consistent with the belief set.) A second difference is that Tennant formulates AGM style axioms in the form of a number of elimination rules and one introduction rule.

According to L&R, a relation-based theory "is natural if we think that the agent's policies for belief change may not always yield a unique belief set as the result of revising a given belief set A with a proposition x" (1989, p. 25). In their 1994 paper, they write in retrospect:

> Some years ago, we proposed a generalization of the well-known approach to belief revision due to Peter Gärdenfors (cf. Gärdenfors 1988). According to him, for each theory *G* (i.e. each set of propositions closed under logical consequence) and each proposition α, there is a unique theory, *G**α, which would be the result of revising *G* with α as a new piece of information. There is a unique theory which would constitute the revision of *G* with α. Thus, belief revision is seen as a function. Our proposal was to view belief revision as a relation rather than as a function on theories. The idea was to allow for there being several equally reasonable revisions of a theory with a given proposition (notation adapted).

The implication is that Gärdenfors's theory fails to "allow for there being several equally reasonable revision of a theory with a given proposition".

Similarly, Tennant motivates his choice of a relation-based rather than a function-based formal framework as follows:

> AGM-theory provides an account of expansion, contraction, and revision of theories with respect to sentences. But it does so by treating the "operations" of contraction and revision as thought they were functional, with uniquely defined values, for any given rational agent, on all possible inputs ⟨*K*, α⟩. An alternative and arguably more reasonable approach would be to treat contracting and revising as non-deterministic processes that can produce a variety of possible values on any given input ⟨*K*, α⟩. A mark of rationality, on the part of any agent, would be to countenance such variety rather than to insist on uniquely defined outcomes. Hence a theory of relational theory-change should be able to furnish such variety, by treating contraction and revision more generally as relational, not functional, notions (Tennant 2006a, p. 490, notation adapted).

The implication, again, is that AGM fails to "furnish such variety".

Tennant et al. apparently believe that AGM fails to be a theory that treats contracting and revising as "processes that can produce a variety of possible values on any given input ⟨*K*, α⟩", i.e. that AGM is, in our terminology, a strongly functionalist theory. This is incorrect. As we have seen, AGM allows for a variety of possible values on any given input ⟨*K*, α⟩. The values will differ depending on what contraction/revision function is being employed. AGM is therefore to be classified as a weakly relationalist theory and not as a strongly functionalist one. This observation undercuts the main relationalist motivation for introducing a relation-based theory of belief revision as an alternative to AGM style function-based theories.[5]

This is not to say that it may not be fruitful to study relation-based belief revision for methodological or systematic purposes. This is also an aspect that Tennant stresses, suggesting reformulating the AGM theory within a relational setting because this has the "methodological advantage" (Tennant 2006a, p. 493) of helping us to "identify certain inadequacies of AGM-theory that might more easily escape attention in the functional setting" (ibid.). Tennant is here referring to his "degeneration" theorems which show that the AGM postulates for revision are too liberal in the principal case of revision of a non-contradictory sentence which is inconsistent with the old beliefs (Tennant 2006b).

Tennant's reformulation has the additional systematic benefit of making it possible to derive, rather than stipulate, K*6, the AGM principle of extensionality, with which we have already made acquaintance. Tennant argues that stipulating rather than deriving the extensionality principle is essential to the functional framework. While there may be systematic reasons for studying belief revision in a relation-based framework, this does not bear on the internal coherence of developing some form of relationalism within a function-based framework.

11.3 A functionalist-Relationalist Dilemma

We have seen that there seems to be no good reasons to rule out either position wFwRF or sRwRF as inconsistent, incoherent or otherwise internally defective. Hence, wFwRF, wFwRR, sRwRF and sRwRR are all still in the race. My point of departure in this section will be strong relationalism, which is an ingredient in the last two positions. What can be said in favor of the view that the result of belief revision may not be rationally determined even if all cognitive aspects of the agent have been taken into account?

In a paper presented at a workshop in 1989 and later published in a proceedings volume in 1991, L&R give what I take to be an argument for strong relationalism.

[5] Tennant's reference to "non-deterministic processes" might suggest a *causal* difference between AGM and a relation-based theory. On this reading, Tennant is taking AGM to be a theory according to which the new belief state after revision is causally determined by the old belief state and the new input, and he is seeing the relationalist position as the opposite view that the most that can be said is perhaps that there is a certain probability (less than 1) that a given belief state will ensue. However, it is generally agreed that theories of belief revision in the tradition of AGM are normative theories of rationality and not descriptive theories. Such theories of belief revision do not purport to account for how people actually change their views but rather for how they ought to do it. For that reason, it is difficult to see how a *causal* distinction between determinism and indeterminism could be relevant in this field of research. Rather what is meant by claiming or implying that AGM is deterministic must be that, according to AGM, the result of belief revision is *rationally* determined, so that given an old belief state and a new datum, there is only one new belief state that is compatible with the AGM principles of rationality. This amounts to claiming or implying, falsely, that AGM is a strongly functionalist theory when in fact it is a weakly relationalist one.

Their point of departure is Adam Grove's paper from 1988 where two related models of functional belief revision are presented, one in terms of a family of spheres around the agent's theory G, viewed as a set of possible worlds, and the other in terms of an epistemic entrenchment ordering of propositions. Grove's spheres may be thought of possible fallback theories relative to the agent's original theory. By a fallback theory is meant a theory that may be reached by deleting propositions that are not sufficiently entrenched. In other words: fallback theories are theories that are closed upwards under entrenchment so that, if T is a fallback, A belongs to T and B is at least as entrenched as A, then B also belongs to T. Figure 11.1 illustrates Grove's family of spheres around a given theory G. We notice that the spheres around a theory are nested, i.e., simply ordered. For any two spheres, one is included in the other.

The next picture illustrates how revision is supposed to work in the Grove model. The area labeled H in Fig. 11.2 represents the revision of G with a proposition A. The result of revising G by A is taken to be the strongest A-permitting fallback theory of G expanded by A. This corresponds to the taking intersection of A with the smallest sphere around G that is compatible with A. This clearly gives a unique result. (If A is inconsistent, the revision by A is taken to be the inconsistent theory, i.e. the empty set of worlds.)

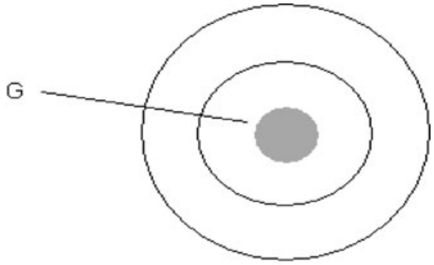


Fig. 11.1 A theory and its family of spheres

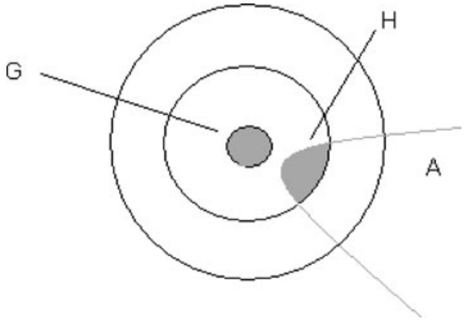


Fig. 11.2 Revision in the Grove model

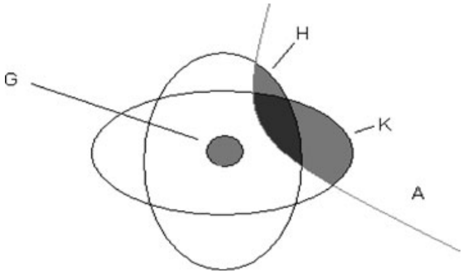


Fig. 11.3 Revision in Grove model with incomparability

But assume now that some propositions may be incomparable with respect to entrenchment. Two propositions are incomparable if neither is at least as entrenched as the other. Hence, allowing for incomparability means relaxing the assumption that the entrenchment ordering is connected. As a result, the family of fallbacks around a given theory no long has to be nested. It will no long be a family of spheres but, to use L&R's term, rather a family of "ellipses". Allowing for incomparability vis-à-vis entrenchment means opening up for the possibility that there may be several different ways to revise a theory with a given proposition. See Fig. 11.3 for an illustration.

In the picture, the two ellipses represent two different fallback theories for G. Each of them is a strongest A-permitting fallback. Hence, both H and K is the intersection of A with a strongest A-permitting fallback. It is natural, therefore, to say that both are possible revisions of G by A.

Still, so far this is merely a hypothetical defense of relational belief revision. What L&R have argued is that a case could be made for relational belief revision if propositions can plausibly be incomparable with respect to entrenchment. The question remains as to whether propositions can be thus incomparable.

L&R's view is that they can. We can, they say, be unable to compare propositions "perhaps because the propositions are so different from each other, or perhaps because they are totally unrelated" (1991, p. 106). The vagueness of this short account of the roots of incomparability makes it difficult to assess. In particular, it is unclear what L&R mean by propositions being "different". Their reference to "totally unrelated" propositions suggest that they have in mind unrelatedness with respect to topic. Still, there are many topic-wise unrelated propositions that are easily comparable. For instance, I consider my belief that the earth is round much more entrenched than my topically unrelated belief that we will have pork for lunch today. Hence, topic-difference cannot be a *source* of entrenchment incomparability. On this reading of their proposal, L&R still owe us an explanation of why some cases of topic-difference lead to incomparability and some do not.

Isaac Levi has provided a more compelling defense of entrenchment incomparability. Without going into any technical details, incomparability results, says Levi, not because we are comparing propositions that are different content-wise, but "due to conflict or indeterminacy in the agent's values and goals" (Levi 2004, p. 206).

This yields indeterminacy in the sense that the agent's assessment of informational value needs to be represented not as a single measure (a so-called M-measure) but as a (convex) set of such measures. Each such measure, as Levi shows via his concept of damped informational value, gives rise to a permissible entrenchment ordering of the agent's beliefs. Levi notes that each permissible entrenchment ordering yields a nested system of spheres in the sense of L&R, and so "[i]f we consider all unions of the sets of fallbacks associated with each permissible ordering, we have a system of fallbacks of the sort considered by Lindström and Rabinowicz with an associated entrenchment ordering that allows for incomparabilities" (ibid., p. 211).

In other words, agents with conflicting theoretical goals and values may end up in a situation where there is no unique way to order beliefs with respect to epistemic entrenchment but several equally admissible orderings. This would seem to open up for the possibility that a given change in belief may give rise to an indeterminate result; that there could be several equally rational ways to change beliefs. A strongly relationalist theory can accommodate this indeterminacy. A functionalist theory, it may appear, cannot. From this perspective, the strongly relationalist approach may seem to be the philosophically more meritorious position.

The question however is whether the strong relationalist, in her argumentation, has really exhausted all the resources of theoretical rationality. Maybe it is true that there is not always a unique way to order beliefs with respect to how entrenched they are. But many researchers, among them advocates of AGM or its variants, would be unhappy with letting this be the end of the story. In such cases, they would say, theoretical rationality dictates that we invoke a rule for ties as a further criterion. The result of belief revision all things considered should be a belief state containing all and only the beliefs that are common to all admissible belief states. This new belief state will be unique.

But the relationalist could counter as follows: The relational view may not be so plausible so long as we confine ourselves to the consideration of small changes of specific beliefs within a system of belief where the system itself does not undergo any dramatic changes. But consider a case of a *bone fide* scientific theory change. Suppose we have a theory which must be changed in the light of the outcome of one or more experiments and, in addition, criteria for rational theory choice, such as empirical adequacy, simplicity, fruitfulness and the like. Given all this, there is no guarantee that one single unique theory will satisfy our adequacy criteria optimally. Rather, we should not be surprised to find that several theories tie for optimality. And, crucially, what guarantee do we have for thinking that suspending judgment between these optimal theories will give rise to a theoretical position that is itself optimal or even qualifies as a scientific theory at all?[6] The result of suspending judgment between the different theories that surface concerning the extinction of the dinosaurs would hardly itself pass as a theory on the matter in the scientific sense.

[6] The objection raised in this paragraph is due to Sten Lindström (personal communication).

The problem of justifying judgment-suspension in the apparent absence of a guarantee that it leads to a good theory motivates much of Isaac Levi's later research on belief revision, from Levi (1991) to Levi (2004). His approach has been to insist that suspending judgment between optimal theories results in a theory that is not only rationally admissible but even optimal. Levi bases his initially somewhat dubious view on an intricate analysis of the theoretical values that are involved in theory choice and how they combine into his measure of informational value.

While Levi should be credited for stressing the importance of the problem of ties and for providing a detailed and sophisticated solution to it, it is difficult to avoid the impression that his measure of informational value was chosen mainly for the reason that it gives the desired result vis-à-vis judgment-suspension. For related criticism, see Rott (2006). I believe that Levi's persistent efforts in this direction have been less than convincing, and that the prospects of making any further progress are dim. But, as always, the only good objection is another theory. I will therefore proceed to propose an alternative approach to the dilemma which has been my main concern in this section. The alternative resolution is not only simpler and, I believe, initially more appealing than Levi's proposal; it also has an independent standing that the latter to some degree lacks.

11.4 Weak Functionalism and Dynamic Caution: A Preliminary Defense

The dispute between strong relationalism and weak functionalism concerns, again, what to accept when there are several theoretical options (hypotheses) that are "best", i.e., equally good, and better than any other option (hypothesis). The bold strategy recommended by the strong relationalist is to pick one of the best theoretical options arbitrarily, or in some other way that does not appeal to the principles of theoretical rationality, as the new view on the matter. This approach, though undoubtedly representing a minimal change approach, is intuitively repugnant. It doesn't seem rational to choose to fully to believe something when one might just as well have fully believed something else instead. The skeptical strategy to which the weak functionalist pledges allegiance involves taking as the new position what all the best theoretical options have in common, i.e., to suspend judgment between them. The drawback of this strategy is that there is, *pace* Levi, no compelling reason to think that suspending judgment between optimal positions leads to a position that is itself optimal or even satisfactory. Thus we are presented with what seems to be a forced choice between two unattractive positions.

Otherwise put: we have to face the predicament that none of positions in the functionalist-relationalist controversy that have so far survived critical scrutiny – wFwRF, wFwRR, sRwRF or sRwRR – seems acceptable. For they all involve a commitment to either weak functionalism or strong relationalism, and we have just seen that both seem unacceptable from the standpoint of pre-systematic intuition.

I believe, however, that the dilemma is based on a misconception of what it means rationally to suspend judgment. It seems true that the belief state corresponding to suspension need not itself be one of the optimal belief states; typically, it won't. It is also true that one should never settle for a suboptimal option. But this only shows that if one accepts (only) the belief state corresponding to suspension, one must do so *without settling for it*. In other words, that belief state should not be one that we rest content with. Rather, accepting it should generate or preserve, as the case may be, a commitment to settle the original issue as to which of the best options should be accepted. The goal to settle that issue should still be on the agent's *research agenda*.

What I am suggesting is that the belief state corresponding to suspension of judgment should not be chosen as the end-point of inquiry but as an intermediate result in an on-going investigation. Usually, an agent who has come to the conclusion that what the best theories have in common is what can be assumed to be the case at the present stage of inquiry would nevertheless continue asking which one of those best theories should eventually be accepted. Still on the agenda, this question serves to motivate further inquiry and deliberation aiming at the ultimate acceptance of one of the theories among which judgment was suspended.

What I am proposing, more precisely, is that epistemic states be viewed as more complex objects consisting not only of a belief set, an epistemic entrenchment ordering (or some other suitable ordering or choice mechanism) but also of a research agenda. The research agenda can be represented as a set of questions (Olsson and Westlund 2006). A question, in turn, can be represented as a set of potential answers. In the following, the entrenchment ordering (choice mechanism etc) will be disregarded.[7]

Suppose, for instance, that as the effect of receiving new information three alternative states of belief B_1, B_2 and B_3 present themselves as being as good as any other. Subsequent inquiry and deliberation reveals that B_1 and B_2 are equally good and better that B_3. On the present view, the inquirer is now justified in believing all and only what B_1 and B_2 have in common, provided that she retains on her agenda the task of deciding which one of B_1 or B_2 is ultimately to be chosen. On my reconstruction, the new epistemic state should be something like $E = \langle B_1 \cap B_2,\ \{\{B_1, B_2\}\}\rangle$, where $B_1 \cap B_2$ is the new state of belief and $\{\{B_1,\ B_2\}\}$ the new research agenda. Intuitively, this epistemic state is better than either of the two "bold" alternatives of choosing as the new epistemic state either $E_1 = \langle B_1,\ \emptyset\rangle$ or $E_2 = \langle B_2,\ \emptyset\rangle$ or, for that matter, $E_3 = \langle B_1 \cap B_2,\ \emptyset\rangle$. The latter represents suspending judgment while maintaining an empty agenda.

However, it is one thing to assert, as I have done, that dynamic caution is admissible and optimal; it is quite another to supply a decision theoretic argument to the same effect. Supplying such an argument would involve, among other things, supplying a theory for how more complex epistemic states – states that involve research agendas – are to be evaluated and compared with respect to informational

[7] For more recent accounts of the agenda in belief revision, see Enqvist (2009) and Genot (2009).

value. This issue still needs to be addressed in its entirety. What can be said already at this point is that $\langle B_1 \cap B_2, \ \{\{B_1, B_2\}\}\rangle$, should be decision-theoretically more advantageous than $\langle B_1 \cap B_{2,} \ \emptyset\rangle$ from the point of view informational value. In the former case, but not in the latter, there are things on the research agenda, meaning that, everything else being the same, the prospects of a future increase in informational value is relatively high. It is more difficult to rationalize in decision theoretical terms why $\langle B_1 \cap B_2, \ \{\{B_1, B_2\}\}\rangle$ should be considered preferable to, say, $\langle B_1, \ \emptyset\rangle$. Intuitively, this has to do with the increased risk of error involved in accepting the latter as opposed to the former, but – as Levi has taught us – risk of error is a problematic notion in the context of belief contraction.

The new agenda-based proposal raises another open question of some urgency. As Rott (2006), p. 191, points out, Levi's theory can be seen as an attempt to satisfy three principles at once. Suppose the agent has decided to contract α from her corpus. First, there is what Rott calls the Decision-theoretic Rule which says that "The corpus after a contraction must be optimal, that is, it must minimize the loss of informational value among all corpora expelling the hypothesis α" (ibid., notation adapted). Second, the Rule for Ties in Contraction should be obeyed: "given a set of optimal contraction strategies, one should always choose the weakest of them if it exists" (Levi 2004, p. 119). Finally, there is the Intersection Equality: "If members of a set **S** of contractions from *K* are equal in informational value, their intersection is equal in informational value to the informational value of any element of **S**" (Levi 2004, p. 125, notation adapted). In the special case of **S** being the set of contractions that are optimal in the sense of minimizing loss of informational value the Intersection Equality ensures that the intersection of **S**, representing judgment-suspension in a sentence-based framework, is also optimal. Levi, as we saw, achieves the satisfaction of these three principles *ad hoc* by introducing a measure of damped informational value that significantly lacks independent support.

Now, it would be interesting to find out whether our new, more complex framework, involving research agendas, could come to Levi's rescue here. In other words, does the shift to the agenda-based framework allow the simultaneous satisfaction of Levi's three principles without there being any need for substantial extra assumptions? This, too, is a question I must leave uninvestigated.

Summing up, the present way out of the dilemma between the "bold" and "skeptical" approaches is based on the observation that the skeptical approach has not been advocated in its most plausible form. AGM style theorizing on the matter conveys the impression that the skeptical position resulting from judgment-suspension is fine as it is, i.e., that there is no need to inquire any further. However, there is a more plausible form of the skeptical approach that combines taking what the theories or belief states have in common as the new belief content with updating the research agenda in the manner suggested. What I am getting at can be described as the difference between a "static" and a "dynamic" caution; between being satisfied with having suffered informational loss and accepting such loss only as part of a temporary retreat. The kind of caution that I favor is dynamic.

11.5 Conclusion

In an attempt to contribute conceptual clarity I have distinguished three ways of drawing the line between functionalism and relationalism: weak functionalism vs. strong relationalism, strong functionalism vs. weak relationalism and finally function-basedness vs. relation-basedness. These distinctions were then employed to shed light on the functionalist-relationalist debate in philosophical logic. In the final section, I argued for a weak functionalism according to which the belief state resulting from revision is rationally determined given the new datum, the prior belief state and other relevant features of the agent's cognitive makeup. My brand of such functionalism was seen to be one with a "relationalist touch": it concedes to the relationalist that the skeptical position resulting from suspension of judgment is unreasonable if interpreted as a static end-point of inquiry. Rather, such suspension must be interpreted dynamically as an intermediate position in an on-going investigation aiming at the eventual acceptance of one of the optimal theories among which judgment was originally suspended. Making this precise is a task left for future research. It requires, among other things, a formal framework that includes a representation of the agent's research agenda.

References

Alchourrón, C., P. Gärdenfors, and D. Makinson. 1985. On the logic of theory change: partial meet contractions and revision functions. *Journal of Symbolic Logic* 40:510–530.

Doyle, J. 1991. Rational belief revision: Preliminary report. In *Principles of knowledge representation and reasoning*, ed. J.A. Allen, 163–174. Los Altos, CA: Morgan Kaufmann.

Enqvist, S. 2009. Interrogative belief revision in modal logic. *Journal of Philosophical Logic* 38(5):527–548.

Galliers, J.R. 1992. Autonomous belief revision and communication. In *Belief revision*, ed. P. Gärdenfors, 220–246. Cambridge tracts in theoretical computer science 29: New York, NY: Cambridge University Press.

Gärdenfors, P. 1988. *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge, MA: MIT.

Gärdenfors, P. 1992. Belief revision: An introduction. In *Belief Revision*, ed. P. Gärdenfors, Cambridge tracts in theoretical computer science 29: Cambridge University Press.

Genot, E.J. 2009. The game of inquiry: The interrogative approach to inquiry and belief revision theory. *Synthese*, published online.

Hansson, S.O. 1991. *Belief base dynamics*. Uppsala: Acta Universitatis, Upsaliensis.

Hansson, S.O. 1998. Revision of belief sets and belief bases. In *Belief change*, eds. D.M. Gabbay, and P. Smets, 17–75. Handbook of defeasible reasoning and uncertainty management systems, vol. 3. Dordrecht: Kluwer.

Levi, I. 1991. *The fixation of belief and its undoing*. Cambridge: Cambridge University Press.

Levi, I. 2004. *Mild contraction: Evaluating loss of information due to loss of belief*. Oxford: Clarendon.

Lindström, S., and W. Rabinowicz. 1989. On probabilistic representation of non-probabilistic belief revision. *Journal of Philosophical Logic* 89(18):69–101.

Lindström, S., and W. Rabinowicz. (1991), Epistemic entrenchment with incomparabilities and relational belief revision. In *The logic of theory change*, eds. A. Fuhrmann, and M. Morreau, 93–126. Berlin: Springer.

Neurath, O. 1981. Die Verirrten des Cartesius und das Auxiliarmotiv (Zur Psychologie des Entschlusses). In *Gesammelte philosophische und methodologische Schriften*, vol. 1, eds. R. Haller, and H. Rutte, 57–67. Wien: Hölder-Psichler-Tempsky. The paper is dated 1913.

Olsson, E.J., and D. Westlund. 2006. On the role of the research agenda in epistemic change. *Erkenntnis* 65(2):165–183.

Rabinowicz, W., and S. Lindström. 1994. How to model relational belief revision. In *Logic and philosophy of science in Uppsala*, eds. D. Prawitz, and D. Westerståhl, 69–84. Dordrecht: Kluwer.

Rott, H. 2006. The value of truth and the value of information: On Isaac Levi's epistemology. In *Knowledge and inquiry: Essays on the pragmatism of Isaac Levi*, ed. E.J. Olsson, 179–200. New York, NY: Cambridge University Press.

Stölzner, M. 2000. An auxiliary motive for Buridan's ass: Otto Neurath on choice without preference in science and society. *Conceptus* 33(82):23–44.

Tennant, N. 2006a. New foundations for a relational theory of theory-revision. *Journal of Philosophical Logic* 35(5):489–528.

Tennant, N. 2006b. On the degeneracy of the full AGM-theory of belief revision. *Journal of Symbolic Logic* 71(2):661–676.

Chapter 12
Knowledge as True Belief

Isaac Levi

12.1 Sources of Knowledge and Knowing That

"Knowledge" is an honorific intended to distinguish sources of information that are approved from those that are not and also to distinguish full beliefs that are prized from full beliefs that are despised. These are two distinct functions. The function of sources of information is quite different from the function of states of full belief. We should not expect that the characterization of knowledge in the two cases should be the same or that conflict between the two can be settled by appeal to "our ordinary" concept of knowledge. (There is no such thing.)

X's state of full belief K is used to distinguish between serious possibilities (logical possibilities consistent with K) that are open to "real and living" doubts and serious impossibilities inconsistent with K whose falsehood is from the point of view of the inquirer absolutely certain.[1] X fully believes that *h* if and only if the potential state of full belief that *h* is a consequence of K. The set of consequences of K is the set of full beliefs to which X at the given time is committed. The state of full belief K so conceived is the evidential basis on which X conducts inquiry, makes judgments of uncertainty and takes decisions. The full beliefs that are consequences of K provide information that gratify curiosity to various degrees and either avoid error or fail to do so.[2]

I. Levi (✉)
John Dewey Professor Emeritus, Columbia University, New York, NY, USA
e-mail: levi@columbia.edu

[1] My first effort to elaborate the conception of knowledge and its relation to full belief discussed here appeared in Levi (1983) versions of which were read at several places in the early 1970s. The final version was submitted for publication in 1975 but did not appear until 1983. In 1976, I published another version of the same ideas. The first three chapters of Levi (1980) are a more leisurely and carefully elaborated statement of the same general conception. More recent statements of this view are found in Levi (1991, 2004, ch. 1).

[2] The full belief that *h* is to be distinguished from the state of full belief K of which it is a consequence. Both the full belief that *h* and the state of full belief K are potential states of full belief. The set of potential states of full belief *K* is a Boolean Algebra. I shall suppose that *K* is an atomic algebra. *W* is the set of atoms. I assume that the Algebra is closed under meets of subsets of *K* of

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_12,

The full beliefs that are to be prized are those that carry valuable information and avoid error. But we prize all extralogical information to some degree as long as error is avoided. And even if logical truths carry no information, we are committed as rational agents to believe them (insofar as we are able). For this reason, I argue that X knows that *h* if and only if X fully believes that *h* and that *h* is true.

This view of knowledge emphasizes the interests of the inquirer who uses full beliefs in deliberation and inquiry. The basic problem of epistemology from this point of view is not the definition of knowledge but identifying conditions for justifying *changes* in states of full belief. (Levi 1980, ch. 1–3, 1991, ch. 1) X should remain in the state of full belief in which X is currently situated unless X has good reason to add or to give up beliefs that are consequences of that state. I understand this to be the attitude implicit in Peirce's rejection of efforts to create doubts where there is no serious problem. Reasons are not to be demanded for current beliefs but for changing beliefs either by adding to the current stock or depleting it.

Edward Craig (1990) and I agree that knowledge is best understood in terms of the purposes for which it is designed. But instead of focusing on knowledge as meritorious belief used as evidence and as a standard for serious possibility, Craig holds that "the concept of knowledge is used to flag approved sources of information" (Craig 1990, p. 11). Craig is right to point out that knowledge may be characterized as certified sources of information. But where I think of knowledge so conceived as serving a quite different function from knowledge as a species of belief, Craig explores attempts to derive a characterization of the latter type of knowledge from conditions on what constitutes an approved source of information.

To do this Craig restricts attention to sources of information that are *informants*. Informants are agents with points of view whom we may consult as to their *beliefs* on issues concerning which we are in doubt. Craig is fully aware that there are other sources of information that lack beliefs or other attitudes. Encyclopedias and other such records, symptoms and other manifestations that are sources of information that do not testify to the truth of propositions by expressing beliefs; for they may

cardinality of *W*. The members of *K* are thus the power set of *W*. *K* is the set of states of full belief that an inquirer may coherently adopt at some given time. I do not mean to impose any conceptual bound on *W*. If X wishes to consider potential states that are more specific than the elements of *W*, I assume that X may move to a space of potential states of full belief more fine grained than *K*. But I shall retain the restriction that the new set of potential states like the old one is an atomic algebra. The atoms in *K* are, thus, not to be confused with conceptual counterparts of possible worlds. There are no possible worlds immune to splitting. Thus, subsets of *W* are not to be confused with propositions understood as sets of possible worlds. If one wishes, one can think of subsets of *W* as doxastic propositions in *K*. Or one can think of the joins of such subsets (which are potential states of full belief) as doxastic propositions in *K*. X's full belief that *h* is full belief in the truth of a doxastic proposition understood as a potential state of full belief that *h*. X's full belief that *h* could be the potential state of full belief K that is X's current state of full belief. But as a general rule it is a potential state of full belief that is weaker than K. Hence, I do not think of X's full belief that *h* as a state. *A fortiori* it is not a mental state. Consider, however, the set of full beliefs to which X is committed at a given time. This is the set of all consequences of X's state of full belief K and may be used to characterize that state.

lack beliefs. Craig restricts attention to testifiers who have beliefs expressed by the testimony they offer. Why does he do that?

Craig contends that what an inquirer requires of an informant approved as a source of information is that the testimony of the informant express true belief plus "any detectable property which has been found to correlate closely with holding a true belief as to whether *p*" (Craig 1990, p. 25). Various current characterizations of knowledge such as true belief that tracks the facts, true belief causally connected to the facts, true belief attained by a reliable method and true belief with a reason are not "far off the mark" as properties that correlate well with properties indicative of true belief that are more or less user friendly. This is not because any one of these types of analyses succeeds in explicating the meaning of "knowledge". Craig and I agree in turning our backs on such analyses. But Craig thinks that there are certain general beliefs that "we all hold" "about the extent to which the world is a system of causally inter-related states, more specifically beliefs about the extent to which belief-states are themselves the end-product of a causal process; the belief that for nearly all human beings there is such a thing as the method by which their beliefs are acquired; and the fact that human beings are usually conscious of certain stages of the processes by which they arrive at beliefs" (Craig 1990, p. 34).

I myself do not share in these beliefs that according to Craig we all hold. I do not understand the thesis that the world is a system of causally interrelated states and doubt that for nearly all human beings there is such a thing as the method by which their beliefs are acquired or that human beings are usually conscious of some of the stages by which they arrive at beliefs.

Craig has, nonetheless, contributed an important insight into why certain types of accounts of conditions for knowledge have the attraction they do. These accounts are of the sort I had in mind when I made reference in Levi (1980) to pedigree theories of knowledge. If what we want from agents that have beliefs is that they be informants who are approved sources of information and think that their status as approved sources of information depends on their beliefs being not only true but truth indicative, Craig's approach explains the allure of pedigree theories of knowledge as due to the idea that knowers are certified informants – sources of information that provide information by manifesting their beliefs – in virtue of the pedigrees of their beliefs.

Craig is right to emphasize the importance of the difference between approved sources of information and sources of information that are not so certified. The former are often taken to be sources or repositories of knowledge. However, encyclopedias are storehouses of knowledge, so are the rings on tree trunks. Like informants, these are often approved sources of information. Obviously different types of sources of information provide the information in different ways and these differences can impact on the circumstances under which the sources are approved. But an approved source of information does not have to be an informant as Craig is well aware. Craig distinguishes between informants and other sources of information that have "evidential value" from which information can be gleaned. (Craig 1990, p. 35)

Craig insists, however, "the concept of knowledge, as we operate it in everyday practice, is tied to informants rather than sources of information in the sense just (approximately) characterized." (*loc. cit.*)

Clearly, encyclopedias are not informants. And yet they are considered sources of knowledge. And the Garden of Eden had its Tree of Knowledge. In everyday practice, I submit, sources of information that are approved are called "sources of knowledge" *whether these sources are informants or not*. Informants can have beliefs it is true. And being an agent having beliefs is a necessary condition for having knowledge. But it is not a necessary condition for being a source of knowledge. The properties of a source of information that render it worthy of approval for use by inquirers – the properties Craig claims to be seeking – characterize the conditions for being a source of knowledge. They do not in any obvious way characterize the conditions for having knowledge that *h*.

Consider then the special case of a source of knowledge where an informant is an approved source of information. It is at least an open question whether the informant's qualifications as a source of knowledge must depend on whether the informant's testimony expresses beliefs having properties correlated with the beliefs possessing an appropriate pedigree. What does matter is whether the inquirer can rely on the informant's testimony when appropriately interpreted. From the inquirer's point of view, the informant should be considered a source of knowledge and not merely a source of information. But the informant need not know the information about which he or she is testifying. The informant may testify to the opposite of what the informant believes and yet the informant's testimony may be a reliable source of information. And, as is more frequently the case, informants will "bullshit" in Harry Frankfurt's technical sense and may testify reliably when they have no view.

But even if we concede that in most cases, informants that are sources of knowledge also know what they are talking about, the informants' knowing that the information they furnish is true can mean that the informants have true beliefs whether or not these beliefs have a proper pedigree. If informant X's full beliefs serve as X's evidence and standard for serious possibility, what relevance does pedigree have to serving this function as long as the beliefs are true? Craig resists the obvious answer.

> It is not just that we are looking for an informant who will tell us the truth about *p*, we also have to be able to pick him out, distinguish him from others to whom we would be less well advised to listen. (p. 18)

But even if Craig were right (which he may or may not be) that the pedigree of an informant's beliefs matters to the inquirer seeking to use the informant as a source of knowledge – i.e., as a certified reliable source of information, it does not follow at all that this pedigree is necessary to qualify the informant's beliefs as knowledge.

Keep in mind that the informant is also an agent who may be a scholar or researcher or witness to events or anyone who has convictions. Whether the agent's beliefs constitute knowledge depend upon whether these beliefs are to be prized or despised in their capacities as beliefs and not as items in a store of knowledge. The purposes are quite different. An agent uses the beliefs he or she possesses as the evidence and standard for serious possibility in deliberations and inquiries he or she conducts in seeking to answer the questions and take the decisions he or she faces. To be good in this functioning, the beliefs (full beliefs) should be true and carry information (i.e., contribute to the standard for serious possibility). The agent who has these full beliefs takes for granted that the beliefs are true full beliefs. From the inquirer's current point of view, his or her current state of full belief and his or her current state of knowledge coincide.

Of course, when considering the full beliefs of others or of him or herself in the past or the future, the same inquirer distinguishes between full belief and knowledge. And when considering him or herself as a source of knowledge that others may depend upon, the same agent must take into account whether they have the information available to certify him a source of knowledge. So the inquirer can readily distinguish between what he or she now knows and whether he or she is a certified expert informant.

Consider then some hypothesis *h* concerning whose truth the inquirer is in doubt. The inquirer may seek data from reliable sources of information or may reason inductively from the information already available to him or her to some conclusion on this matter. The aim of the endeavor in both cases is to obtain valuable information that is error free. Justifying the acquisition of the new information is not part of this aim. Justification purports to show that the new beliefs adopted best promote the aim. In this context too, what the inquirer prizes is true full belief (or true information).

So my contention is that whether the inquirer already has beliefs that he prizes or seeks such belief, what is prized is true belief and nothing else. The pedigree of belief is of no account.

When an agent evaluates a source of information, he continues to prize true belief. But the agent is now evaluating a source of information and not beliefs. In this context, the reliability of the source increases in importance and when the source is an informant, it may (but may not) be the case that the pedigree of the *informant's* beliefs (as opposed to the inquirer's beliefs) becomes important to ascertaining whether the informant is a source of knowledge. It continues to play no role in determining whether the agent's full beliefs are knowledge.

Craig is right to insist that we characterize knowledge in terms of the functions it serves. But those functions are different when we consider sources of knowledge and knowledge as a species of belief. To insist, as Craig does, that beliefs are to be evaluated exclusively in terms of their relevance to sources of information and not in terms of their use as standards for serious possibility and evidence is to foster the alienation of agents from their own convictions.

Most commentators who understand knowledge to be a species of belief (Craig included) insist with near unanimity that knowledge is belief that is true and satisfies some other condition. "A central part of epistemology, as traditionally conceived, consists of the study of the factors in virtue of which someone's true belief is an instance of knowledge." (Jason Stanley 2003, p. 1). These factors are often supposed to characterize some conception of the provenance of true belief.

Craig's discussion is a skillful argument showing how insistence on warrant or pedigree sometimes derives from an equation of "X knows that *h*" with "X is a certified informant concerning the truth of *h*. He and Philip Kitcher (2006) are quite clearly committed to this equation. There are others.

Michael Williams writes as follows:

> Still, we might ask, why does justification *matter*? One answer to this question invokes the normative character of epistemic classification. A knower possesses a special kind of *entitlement – epistemic entitlement* – to hold to the proposition in question. Epistemic entitlement confers further entitlements: to use the claim as a basis for inferences or to authorize other people so to use it. So while, in a way, one may be entitled to make guesses and act on them – it isn't illegal – one is not normally entitled to advise others on the basis of mere guesswork. (Williams 2001, p. 20–21)

According to Williams, the epistemic entitlement possessed by a knower confers upon the knower other entitlements: (a) to use the claim as a basis for inferences and also (b) to authorize others so to use it.

To my way of thinking, using beliefs as a basis for inference is to use beliefs as evidence in inquiry aimed at improving one's state of full belief. It is not an entitlement conferred on beliefs that the inquirer knows. The inquirer who has certain beliefs is *committed* to believing also their logical consequences and to using those beliefs as evidence in seeking new information. The agent's state of full belief is the agent's standard for distinguishing serious possibilities from propositions that are not serious possibilities. Serious possibilities may be (but need not be) assigned positive probability. Propositions that are not serious possibilities must be assigned 0 probability. The probability of h conditional on e is well defined provided that e is a serious possibility whether or not its conditional probability is positive. If e is not a serious possibility, the conditional probability is undefined. More generally, the agent's beliefs are characterized by the undertakings to use them as premises in the agent's inquiry and deliberation. The sort of socially licensed authorization that Williams's epistemic entitlements confer has no bearing on the agent's conduct of his or her own inquiries and deliberations.

By way of contrast, offering information to others and claiming that the information is authoritative calls for being in a position to establish one's authority. The agent is then a licensed informant in Craig's sense.

Williams concedes that one may be entitled to make guesses and act on them as long as one does not advise others on the basis of mere guesswork. But an agent's full beliefs, which are the evidence on the basis of which the agent deliberates and acts are, according to that agent, more than mere guesswork. Indeed, they are the constituents of the agent's knowledge. Of course, it matters to that agent whether the testimony of others is to be trusted. That is because the agent seeks help from others in improving his or her body of beliefs. So the agent seeks certification for their testimony as the product of approved sources of information – i.e., sources of knowledge.

The equation of knowledge that and source of knowledge is also found in the account of the relation between knowledge and assertion defended by Timothy Williamson (1996, pp. 489–523). Williamson's favorite rule of assertion is the

Knowledge Rule that he himself says gives a condition "on which a speaker has the *authority* to make an assertion." (*op. cit.*, 509). Since the knowledge rule runs: "One must assert that P only if one knows that P" (*op. cit.*, 494), knowing that P seems to imply a license to assert in the sense of testifying as witness or authority.

Williamson also holds that the inquirer's knowledge constitutes the inquirer's evidence. Like Williams, Williamson thinks of knowledge as a source of knowledge and knowledge as evidence or standard of serious possibility as being one and the same. Knowledge as a source of knowledge is not a species of full belief (as Williamson also maintains for quite different reasons). Knowledge as evidence (which Williamson endorses) and a standard for serious possibility is a species of full belief. To my way of thinking, the question of how to distinguish between sources of information that are sources of knowledge (in a sense in which knowledge is not a species of belief) and sources of information that are not sources of knowledge is a separate issue from distinguishing between belief used as evidence that is knowledge (in a sense that is a species of belief) and belief used as evidence that is not.

12.2 Knowledge and Belief

Consider the following thesis:

> *Knowledge as true full belief*:
>
> All rational agents ought to fully believe and, hence, agree that X knows that *h* if and only if X fully believes that *h* and it is true that *h*.

Is it a necessary truth that knowledge is true full belief? I do not understand the question. For one thing "necessary" is ambiguous and for another many of the interpretations of it (conceptual necessity, metaphysical necessity for example) are beyond my comprehension. I mention two interpretations that are intelligible and according to which the answer to the question is affirmative.

The Knowledge as True Full Belief thesis is put forward as a prescription concerning what all rational agents ought to fully believe. It is not a thesis about the "ordinary" concept of knowledge (as if there were such a thing) or an extraordinary concept. The verb "to know" and the noun "knowledge" may be used by rational agents in all sorts of ways in expressing their beliefs and other attitudes. I do not mean to suggest that rationality mandates a specific usage for either the verb or the noun. And I do not think that there is enough consensus in ordinary linguistic usage to warrant insisting on one way of using it.

But I do think that "knowledge" is an honorific term when applied to belief. It is used to identify those beliefs that are prized and to distinguish them from those that are despised or are merely tolerated. A specification of necessary and sufficient conditions for knowledge that all rational agents ought to believe is an expression of an ideal to which all rational agents ought to be committed. Full belief of the best kind is true belief. Knowledge as an honorific should therefore be applied to full belief of the best kind. In this sense, the Knowledge as True Belief Thesis is a necessary truth.

In further response to the question as to whether Knowledge as True Full Belief is a necessary truth, consider the following feature of full belief or absolute certainty. X's state of full belief at time *t* is X's standard for distinguishing between logical possibilities that are serious possibilities and logical possibilities that are not. That is to say, X's standard for serious possibility determined by X's state of full belief rules out the logical possibility that X fully believes that *h* while *h* is false as a serious possibility and, hence, guarantees that X fully believes that everyone of X's current full beliefs are true. It is easy to see that according to X at *t*, there is no serious possibility that the following claim is false: X knows that *h* at *t* is true if and only if X fully and truly believes that *h* at *t* is true. Thus, *according to X at t*, X's state of full belief at *t* is the same as X's state of knowledge at *t*.[3]

From this it does not follow that according to X at a different time t' or according to Y distinct from X at any time, X's state of full belief at *t* and X's state of knowledge at t' coincide. But even so, Y ought to fully believe that all full beliefs Y and X share in common are true and, hence, known by both X and Y because these full beliefs belong, according to Y at t'. Thus, that X knows that *h* if and only if X fully believes that *h* and it is true that *h* is endorsed from the point of view of every rational inquirer as the Knowledge as True Full Belief Thesis requires.

According to the Knowledge as True Full Belief Thesis, X and Y may disagree as to whether X knows that *h* or not. Should that happen, both X and Y should agree that one of them is wrong. And they should rationally agree that this is so. If X or Y attributes knowledge to X, the biographical remark that X (Y) attributed knowledge to X is clearly relational. But the *attribution* of knowledge to X is not true (false) in a relativist sense. And it is not dependent on the context of utterance in the senses currently fashionable in contextualist epistemology. It is true if and only if X fully believes that *h* and does so truly.

The Knowledge as True Full Belief Thesis clashes with the near unanimity found among the tribe of epistemologists that knowledge is true belief with an appropriate pedigree. Pedigree theories of knowledge include not only views that insist that knowledge is true justified belief but also views that do not require the beliefs to be justified but do insist that the beliefs are not acquired by accident, have the proper

[3] In Levi (1980), I took X's state of knowledge at t to be X's standard for serious possibility rather than beginning by taking X's state of full belief at t to be X's standard for serious possibility. I pointed out then as I have done here that from X's point of view at *t*, the two characterizations coincide. I also pointed out that X distinguishes between Y's knowledge and Y's full beliefs at any time or X's knowledge and X's full beliefs at times other than *t*. This raises the question as to whether from X's point of view at *t*, Y's standard for serious possibility at *t*' is Y's state of full belief or Y's state of knowledge at *t*'. This question was not explicitly addressed but was discussed in Levi (1979) reprinted in Levi (1984, p. 153). The view taken is substantially the one adopted here.

origin or are formed in a legitimate way. In spite of the clash, advocates of pedigree theories agree with the view taken here that knowledge is true belief.

The requirement that the beliefs be true, however, not only disqualifies the false beliefs but also those beliefs that lack truth values. For example, judgments of credal or subjective probability express beliefs – degrees of "partial" belief. Yet such beliefs lack truth-values as Frank Ramsey pointed out.

There is excellent reason for supporting Ramsey's view. To see this, it is useful first to consider some properties of beliefs that carry truth values.

If X assigns a degree of partial belief to the hypothesis that *h* (other than full belief that *h* or full belief that ~*h*), X is committed as a rational agent to judging both *h* and its negation to be serious possibilities. The same holds if X judges it probable that *h* (i.e., that *h* is highly probable or more probable than not). In all these cases, X is committed to suspending judgment with respect to the truth of *h*.

Full belief differs from these attitudes in the respects just indicated. When X is absolutely certain that *h* (~*h*) is true, X is committed to ruling ~*h* (*h*) out as a serious possibility. There is no suspense concerning the truth of *h*.[4]

Now if X fully believes (is absolutely certain) that *h*, X is committed to judging it true that *h*. So a question arises as to whether such full belief is true or false. X, of course, is committed to claiming that X's belief that *h* is true. Having ruled out ~*h* as impossible, to do otherwise would be incoherent.

Y (or X at some other time) may disagree with X. Y may fully believe that *h* is false and, hence, judge that it is false that *h*. Or Y may suspend judgment regarding h and remain in doubt as to whether X's belief is true or false. In that case, Y regards expanding Y's state of full belief by adding *h* (~*h*) as possibly importing false belief. Thus, in suspending judgment regarding the truth value of *h*, Y presupposes that *h* is truth valued.

Whether Y fully believes that *h* is false or is in suspense, as long as Y fully believes that X fully believes that *h*, Y agrees with X that X's full belief carries a truth value. That full beliefs carry truth values ought to be non controversial.

Matters are different in the case of partial belief. When X partially believes that *h* (i.e., judges it probable to some degree or other that *h*) and thereby suspends judgment as to the truth of *h*, X has made no assignment of truth value to *h*. Because X judges it seriously possible that *h* is true and also seriously possible that *h* is false, X does regard X's conjecturing that *h* to be truth-valued. However, X has not assigned the potential state that *h* a specific truth value. X is in suspense as to

[4] Notice that if X judges that *h* is probable to a degree other than 0 or 1, X is in suspense as to whether it is true or false that *h*. The converse, it should be emphasized, does not hold. X may judge that *h* carries probability 1 and yet be in suspense as to whether *h* is true or false. In that case, X has partial belief that *h*. When X is absolutely certain that *h*, X has full belief. Probability 1 in that event does not represent a degree of partial belief. X has ruled out as impossible the truth of ~*h*. If X assigns probability 1 and remains in suspense, X has not ruled out the possibility of the truth of ~*h* any more than X does if X assigns probability less than 1 and greater than 0. (Of course, probability 0 carries the same ambiguity between full and partial belief as probability 1 in dual form.) The difference between full belief and partial belief is no mere matter of degree.

whether it is true that *h*. To be sure X may assign a (numerically determinate or indeterminate) degree of credal probability to the hypothesis that *h*. Assigning such a probability is not, however, assigning a truth value to the potential state that *h*.

Moreover, in making the probability judgment, X is not judging the probability judgment to be true. Suppose Y agrees with X in suspending judgment concerning the truth of *h*. Y, however, assigns a different degree of partial belief to *h*. No doubt Y disagrees with X's judgment of probability. But the disagreement is not a disagreement concerning which degree of partial belief is true.

This conclusion is supported by the following consideration. If degrees of partial belief or credal probability have truth values, one should be in a position to be coherently uncertain or "unsure" of their truth and assign them (second order) credal probabilities. But, as L.J.Savage compellingly argued, one cannot coherently assign credal probabilities to credal probabilities.[5] One cannot suspend judgment concerning the rival credal probability judgments. Such rival partial beliefs, as a consequence, lack truth values.

Partial beliefs are generally taken to be a species of belief. Whether judgments of serious possibility are beliefs is more controversial. It is easy to see that judgments of serious possibility lack truth values (Levi 1979). If X judges it seriously possible that *h*, that *h* is consistent with X's full beliefs. If it is not consistent with X's full beliefs, X should rule out the serious possibility that *h*. As long it is entertainable that *h* according to X, it follows that X should either fully believe that *h*, fully believe

[5] Let different agents hold conflicting probability judgments, value judgments, modal judgments or conditional modal judgments. In order that any one of the parties to the dispute may maintain that there is a true answer to the issue under dispute and that beliefs contradicting this answer are false, that agent should presuppose that the following conditions hold:

(i) Suspending judgment concerning the issue under dispute should be a rationally coherent attitude to adopt.

(ii) Judging the alternative views to be serious possibilities of truth should be a rationally coherent attitude.

(iii) Assigning credal probabilities to the alternative views should be rationally coherent.

In the case of probability judgment, Savage's argument (Savage 1974, p. 58) (as reconstructed in Levi, 1979 to relate to the truth-value bearing status of probability judgments satisfying these three requirements) may be sketched as follows:

Let X suspend judgment between the probability that *h* being r and being s. Suppose X judges it probable to degree *x* that the value is *r* and $1-x$ that it is s. $p(h/p(h) = r)$ should equal *r*. where *p* represents X's current probability. (This is an instance of Van Fraassen's (1984) "Reflection Principle" in the "synchronic" case, the only case where the principle has any merit as a norm of rationality. See Levi 1987.) It then follows from the calculus of probabilities that $p(h) = xr + (1 - x)s$. If $x = 1$, $p(h) = r$ counter to assumption. If $x = 0$, $p(h) = s$ again counter to assumption. If *x* is positive but less than 1, *x* is some real value distinct from *r* and *s* – again counter to assumption. If we allow indeterminacy in probability judgment so that we recognize a range of values for x to be permissible, there will be a corresponding range of permissible values for $p(h)$. But no matter what the interval may be it does not cohere with suspending judgment just between *r* and *s*. The supposition that X has suspended judgment concerning the true probability has led to absurdity in every case and should be rejected. So credal probability judgments cannot be truth-valued.

that ~*h* or be in suspense regarding the truth of *h*. In the first and the third cases, X judges it seriously possible that *h*. In the second case, X is committed to judging it impossible that *h*. There is no context where X may coherently be in suspense as to whether it is seriously possible that *h*. This is so even in cases where rational agents may disagree as to whether *h* is possible or impossible. The judgment that *h* is seriously possible lacks, therefore, one of the crucial features of truth value bearing hypotheses.[6]

Judgments of possibility and impossibility conditional on a supposition whether they are expressed in the form of pure indicative conditionals or as subjunctive or future indicatives exhibit the same failure to satisfy this requirement for carrying a truth – value.[7]

Whether judgments of value are beliefs or not remains another bone of contention. Regardless of the verdict, these judgments lack truth values. This can be shown by an extension of Savage's arguments (see Levi 1979, 1980, 1996, 1997).

If this line of reasoning is correct, value judgments, judgments of serious possibility, credal probability and conditional modality cannot be the kinds of belief (if they are taken to be beliefs at all) that sometimes qualify as knowledge in the sense in which a necessary condition for knowledge is that it be true belief.

12.3 Acceptance as True and Plain Belief

Full beliefs are not the only species of belief that exhibit the marks of carrying truth value. Those who wish to resist the claim that true full belief is a necessary condition for knowledge might insist that true "plain belief" in the sense of Spohn (1990) or true, "acceptance as true" in the sense of Levi (1967a) can qualify.

[6] We should be careful to distinguish between providing necessary and sufficient truth conditions for "it is possible that *h* according to X" and "it is possible that *h*" in terms of consistency with the information available to X (who may be a single inquirer or community agent.) In his excellent paper on possibility (Hacking, 1967, p. 148), Hacking defines what he calls epistemic possibility according to X. (Strictly speaking, Hacking speaks of "epistemic possibility within a community of speakers". I replace "within a community of speakers" by "according to X".) But it becomes fairly clear that his target is to supply truth conditions for "It is possible that *h*"). This shift has provided an excuse for subsequent commentators to explore contextualist or relativist accounts of truth conditions for so-called "epistemic modals". A good example is De Rose (1992). On the view I favor, however, when we drop the relativity to the agent X, there are no truth conditions to account for.

[7] I classify if-sentences according to V.H. Dudman's (1983, 1984, 1985) proposals and maintain that both "pure indicatives" (Dudman's hypotheticals) and subjunctive and future indicative conditionals (Dudman's conditionals) lack truth values. Gibbard's classification (1981) more or less resembles Dudman's but Gibbard regards the subjunctive and future conditionals to have truth-values and to be capable of being judged probabilistically – which I deny (see Levi 1997, 2.7.) This disagreement is closely related to Gärdenfors's argument for abandoning the Ramsey Test for subjunctive and future conditionals (Gärdenfors 1986, 1988) and my response to that argument (Levi 1988, 1997, ch. 3–4).

I do not think such views are convincing. It is true that inquirer X who is in doubt as to whether *h* is true or false can coherently plainly believe that *h* at the same time. The same holds for acceptance as true. However, to regard either notion as an adequate "qualitative" replacement for full belief as a condition of knowledge seems to be based on the assumption that the inquirer who plainly believes that *h* but remains in doubt as to whether it is true that *h* and, hence, judges it seriously possible that *h* is false should not seek to remove doubt by coming to full belief that *h*.

On this assumption, it becomes pointless to make judgments of plain belief or acceptance. Both plain belief and acceptance as true provide useful appraisals of hypotheses that are candidates for addition to inquirer X's stock of full beliefs *prior* to such expansion of X's state of full belief. They are evaluations relevant to efforts to remove doubt. If expansion were illegitimate and if it were inappropriate ever to remove doubt by expanding to full belief, such evaluations would be without a purpose. There would be no need for plain belief or acceptance as true.

To sustain this claim requires explaining the notions of acceptance as true and plain belief and appreciating the features that account for why full belief alone is suited as a necessary condition for knowledge.

Suppose that X is in doubt as to whether one of a set U_{K} of rival hypotheses is true but seeks to find out which of the rivals is true. On this view, X's initial state of full belief K entails that exactly one element of U_{K} is true. Moreover, each conjecture in U_{K} is consistent with K. Finally, X is not, in the context of the inquiry X is conducting, interested in ascertaining the truth of any hypothesis more specific than a hypothesis in U_{K}. The elements of the *Ultimate Partition* U_{K} (Levi 1967a) are the maximally informative potential answers to the question under study as far as X is concerned. A potential answer to the question under investigation is then the rejection of members of a subset of U_{K} and the *expansion* of the initial state K by adding the join or disjunction *h* of the subset |h| of the surviving elements of U_{K} to the stock of full beliefs in K together with the consequences of doing so.

The elements of the Ultimate Partition U_{K} are all serious possibilities according to X in state of full belief K. X does not rule any of the members of U_{K} out with absolute certainty. But X might be prepared to rule out a potential answer *h* equivalent to some subset $|h|$ were X bold enough. That is to say, if X were bold enough, X would be prepared to *change* X's state of full belief by no longer regarding it to be a serious possibility that any member of U_{K} in $|h|$ is true. *Prior to doing so*, X may recognize that q(*h*) is the highest level of boldness at which X would not rule *h* out as a serious possibility were X to fix on a degree of boldness and expand X's state of full belief accordingly (Levi 1967a, p. 8.6). If there is no such highest level because *h* would be rejected for every degree of boldness between 0 and 1, $q(h) = 0$.

Assuming that degrees of boldness may be normalized between 0 and 1, we may then take $1 - q(h)$ as a measure of X's degree of disbelief that *h*.[8] X's degree

[8] d(h) has the formal properties of potential surprise in the sense of G.L.S. Shackle (1949, 1961). Shackle himself also characterized potential surprise as a measure of degree of disbelief. The

of disbelief that *h* is also called X's degree of belief (or degree of confidence of acceptance) that ~*h*.

Consider any candidate answer to the question under investigation representable as the join of some subset of U_K. According to the way potential answers are determined by the ultimate partition, the function q(x) to apply to any such potential answer to the question.

The empty set corresponds to the inconsistent potential state $K_{\perp}$. $q(K_{\perp}) = 0$. Otherwise, $q(K^*)$ where K^* is the join of a subset |K*| of U_K is the maximum of the values for members for the subset $|K^*|$. At least one member of U_K should carry maximum degree of possibility 1. (All elements of U_K could do so.) If X were to fix on a level q of boldness, X should expand from K to a state of full belief K** that has joins of all subsets of U_K for which the degree of belief is at least q and the degree of disbelief or potential surprise in its negation is also at least q. $q(h)$ is a measure of what Dubois and Prade have called "degree of possibility". (Dubois and Prade, 1992).

X's judgment of the degree of possibility q(*h*), degree of potential surprise d(*h*) and degree of belief b(~*h*) are all equivalent ways of assessing prior to expansion of K the highest degree of boldness at which X judges it appropriate to expand K in a way that fails to reject *h*.

Plain belief may now be interpreted as follows: X plainly believes that *h* if and only if ~*h* would be rejected at every positive level of boldness. That is to say, X would reject ~*h* were X maximally bold. The set of all plain beliefs relative to state of full belief K is another potential state of full belief – one that has K as a consequence.

This does not mean that X does fully believe that *h* prior to expansion. That would be absurd. X only believes that *h* to some positive degree greater than 0. Were X to expand, X would then *come* to fully believe that *h*.

Given level of boldness q, one can identify a potential state K_q of full belief having as consequences all potential states of full belief that carry degree of belief greater than q. For any positive level q, we can, therefore, construct a qualitative notion of belief (or acceptance) which, like plain belief, allows X to believe that *h* while failing to fully believe that *h*. This was pointed out in Levi (1967a).

So there is a family of qualitative conceptions of belief or acceptance that can carry truth-value. Let us now return to the question: Should conceptions of belief belonging to this family figure in the belief-condition when knowledge is characterized as presupposing true belief?

It should by now be clear that a negative answer is required. Plain belief and the other notions belonging to the family of notions of qualitative belief as I have interpreted them are useful to have in a context where the inquirer is in doubt as to which of the rival potential answers generated by the ultimate partition U_K is true. Degrees of confidence of acceptance or degrees of belief are appraisals of the

proposal to interpret degree of disbelief in terms of boldness in inductive expansion is in Levi (1967).

potential answers relevant to determining which of them to add to form the new expanded state of full belief.

The point can be restated in dual form: Degrees of confidence of rejection, disbelief or potential surprise are appraisals of potential answers relevant to determining which elements of the ultimate partition U_{K} to reject in forming a new expanded state of full belief.

Those who object to considering full belief as necessary for knowledge should not find these notions any more acceptable as necessary conditions. They are intended for use in evaluating changes from one state of full belief to another. If they are to be prohibited from being used in this way, their use in deliberation and inquiry becomes unclear.

Wolfgang Spohn (2006) denies my allegation that plain belief is useless unless used in justifying changes in full belief by expansion. He suggests that plain belief in the sense of belief to a positive degree can be explicated as obtaining when the balance of reasons for and against a hypothesis is positive and where this balance of reasons is defined in terms of Spohn's ranking functions that exhibit the formal structure of Shackle's measures of degree of potential surprise of disbelief.

How are these ranking functions to be understood? They cannot be used to determine betting rates as credal probabilities can. They can be interpreted in terms of boldness dependent inductive expansion rules as I have done. But this cannot be acceptable to Spohn because it presupposes an account of justifying changes in full belief. Perhaps ranking functions should be taken as primitive and motivated by some presystematic understanding of grades of belief. This is congenial with the version of dualism concerning the conceptions of probability and belief Spohn calls "separatism" and endorses in disagreement with the version of dualism he calls "interactionism" and attributes to me.

I appreciate the idea of taking ranking functions to be primitives. But primitives need explication somehow. Spohn constructs an account of such grades of belief and disbelief accompanied by an account of updating such degrees that parallels R.C. Jeffrey's (1965) account of updating probabilities without being reducible to that account and which comes with an account of iterated updating that elaborates the structure. But the notion of updating ranking functions without adding new information to evidence (full belief) is just as mysterious as the idea of updating probability without adding new information to evidence.[9] And even if Spohn could

[9] For the record, I still stand by Levi (1967b, 1969) adjusted to accommodate my adopting a clear commitment to full belief as the standard for serious possibility as explaining acceptance as evidence. R.C. Jeffrey's alleged refutation of my views in (Jeffrey, 1970) seems to me to concede the main points I was making and illustrating in the examples I discussed that he so roundly criticized. As Jeffrey writes: "To judge the soundness of a shift from *p* to *p*' we must not only look at the two belief functions and their differences; we must also inquire into the forces which prompted the change – the dynamics of the change (Jeffrey 1970, p. 178). I take it that to judge a change sound, one would need to know the forces that prompt the change. But to ascertain this, one would have to acquire full beliefs. And this undercuts the probabilist program that Jeffrey was promoting. It seems to me that Spohn has to address a similar issue.

offer a response to this point, the method of updating of ranking functions does not really answer the question: what are ranking functions for? My difficulty with Spohn's seriously intended and ingenious efforts is that they fail explain the role of the notions of belief and disbelief involved in deliberation and inquiry. What is the point of updating attitudes that make no contribution to deliberation and inquiry? Spohn would no doubt insist that his ranking functions are important in this connection. I do not get it.

My contention is that degrees of belief and disbelief of the sort under consideration are, indeed, useful in an account of rationalizing changes in states of full belief or absolute certainty. As such, their use *presupposes* that states of full belief are subject to legitimate change. More specifically, when X evaluates the degree of belief that *h* and judges either that *h* is positively believed to some degree or other (that is to say, is plainly believed) or that *h* is believed to some degree greater than threshold, X remains in doubt as to whether *h* is true or false. There is a serious possibility that it is true and also that it is false. Under these circumstances, X does not know that *h*.

The clear implication of this is that plain belief or degree of belief above some other level cannot replace full belief as a necessary condition for knowledge. If it did, we could have a case where X knows that *h* and yet remains in doubt whether *h* is true in the sense in which relief from doubt is a motive for inquiry. This is absurd!

In Section 12.2, I argued that propositional knowledge as a species of belief cannot be partial belief in the sense of degree of credal probability. Nor can it be a qualitative notion of belief like "highly probable" or "more probable than not". These notions of belief lack truth values. Knowledge cannot be true belief if belief is understood in one of these senses. Full belief, on the other hand, does carry truth value.

In this section, I conceded that there are other qualitative notions of belief besides full belief that do carry truth values: Spohn's notion of plain belief and my notion of mere acceptance at a given level of confidence. Shackle's notion of degree of belief that *h* as the degree of disbelief that ~*h* gives rise to a family of qualitative notions of degree of belief that can be used to define notions of belief as degree of belief above a certain threshold. These qualitative notions can be generated also from L.J. Cohen's (1977) Baconian probability, the possibility measures of Dubois and Prade (1992), and Spohn's ordinal conditional functions. Spohn's notion of "plain belief" and my notion of mere acceptance with a certain degree of confidence are then definable as carrying a degree of belief above a specific threshold.

I have argued that these notions are useful as modes of evaluating potential answers to questions when they are still held in doubt prior to obtaining an answer. But when plain belief or acceptance with a given degree of confidence are used in this way, even when what is thus believed is true, the beliefs remain possibly false and, hence, in doubt. They are being appraised *prior* to addition to being adopted as answers to the question under study when they are mere conjectures. They cannot count as knowledge.

There are no doubt senses of belief that I have not reviewed.[10] Others are entitled to continue the search for an alternative conception of belief to full belief as a necessary condition for knowledge. I, for one, do not think that the prospects of such efforts will be rewarding. In the subsequent discussion, I take for granted that full belief or absolute certainty is a necessary condition for knowledge.

12.4 Absolute Certainty, Fallibilism and Corrigibilism

At least one contemporary philosopher, Peter Unger, agrees with me that knowledge presupposes absolute certainty (1975, ch. 3). Unger, however, conjoins this thesis with another one from which I strongly dissent:

> It is never all right to be absolutely certain that anything is so.

From this conjunction it is but a short step to Unger's conclusion:

> Nobody ever knows that anything is so.

As I understand full belief or absolute certainty, X's state of full belief is X's standard for serious possibility. That is the standard for possibility within the framework of which X's judgments of credal probability are defined. X's judgments of unconditional probability are fine grained refinements of the set of serious possibilities according to X's standard. X's judgments of conditional probability are restricted to conditions that are serious possibilities according to X. X is absolutely certain that *h* if and only if X rules ~*h* out as a serious possibility. *Pace* R.C. Jeffrey and his many probabilist disciples, there can be no coherent judgment of uncertainty or of probability without a framework of judgments of absolute certainty. Probability presupposes certainty and, indeed, absolute certainty. Full belief in the sense in which knowledge is a species of full belief is absolute certainty as I have already argued.

That *h* is a serious possibility according to X at *t* if and only if ~*h* is not a member of X's set K of full beliefs at *t*. On the assumption that rational X is committed to a set K that is closed under a classical consequence relation, that *h* is a serious possibility according to X at *t* if and only if *h* is consistent with X's state of full belief K (X's state of doxastic commitment K or the corpus of sentences in a regimented language that represents them). X's state of full belief (i.e., the set of full beliefs to

[10] Prominent among them is L.J. Cohen's distinction between belief and acceptance that focuses attention on belief as a disposition to what I should call a fit of doxastic conviction and acceptance as a decision (Cohen 1992). The distinction he makes may be compared to the distinction I draw between beliefs as performances and beliefs as commitments. That is to say, some of the concerns Cohen has in making the distinction are common to my concerns in adopting the contrast I propose. These concerns are not directly relevant to the discussion in section 3.

which X is committed) is, according to that state, infallibly true just in this sense: There can be no serious possibility that any item in it is in error. If Y at *t*′ should disagree with X's view at *t*, Y is, as far as X is concerned, certainly in error. This holds even if Y is X provided that *t*′ is some time different from (earlier than or later than *t*). It would be inconsistent for X at *t* to concede to Y that X might be mistaken at *t* if by this X acknowledged the falsity of *h* as a serious possibility.

To require that knowledge is a species of full belief is, therefore, to imply the infallibility of knowledge in the sense that, according to the inquirer at *t*, there is no serious possibility that what the inquirer knows at *t* is false.

Unger's objection to the propriety of the attitude of absolute certainty or full belief is based on the complaint that requiring absolute certainty fosters a dogmatic attitude where such an attitude involves a refusal to consider modification of one's views. The charge is baseless.

Unger contends rightly that full belief that *h* or absolute certainty that *h* requires the total absence of all doubt. I prefer to say that what is absent is the presence of "real and living doubt" in the sense of Peirce. According to Peirce, there is no serious possibility that *h* is false. Unger, however, understands the absence of doubt differently. It is the absence of "any *openness* on the part of the man to consider new experience or information as seriously relevant to the truth or falsity of the thing". (Unger 1975, p. 116).

If X is absolutely certain that there is an ink bottle before him, he rules out the logical possibility that there is no ink bottle before him as a serious possibility. Let us not worry now how X became so convinced. Unger's charge that X's absolute certainty fosters a dogmatic attitude makes no claim about how X acquired the conviction. Unger claims that once X is certain, X is committed to refusing "to consider any new experience or information as seriously relevant to the truth or falsity of the thing." But X can consistently rule out the logical possibility that there is no ink bottle before him as a serious possibility while, at the same time, recognizing as a serious possibility that future experience may warrant X's withdrawing this judgment. In this latter respect, being absolutely certain is consonant with being prepared "to consider new experience or information as seriously relevant to the truth or falsity of the thing"

Suppose with Norman Malcolm "when *I*(X) next reach for this ink-bottle my hand should seem to pass through it and I should not feel the contact of any object". Prior to this happening, X can consistently regard this episode as a serious possibility. Indeed, X can also take as a serious possibility that X responds to the stimulus by forming the belief that the ink bottle is not present. Given X's initial absolutely certainty that the ink bottle is present, X recognizes as a serious possibility in that state that X might come to form a false belief. Because X would then be certain that the belief is false as well as true, X acknowledges that the result of the experience would be that X is in an inconsistent or incoherent state of full belief. There is nothing incoherent in X recognizing this as a serious possibility while being absolutely certain of the presence of the ink bottle.

If X does in point of fact enter into the conflicted inconsistent state, X will need to retreat from the inconsistency. But X will not be in a position to deliberate coherently while in an inconsistent state of full belief.[11] To avoid incoherent deliberation, X, before the fact, may adopt a policy of how to respond to such serious possibilities of doxastic conflict. The deliberating agent X can anticipate what the available options for retreating from inconsistency would be. As I have argued elsewhere, X might (1) ignore the recalcitrant experience and retain the original conviction, (2) replace the conviction that the object is an inkbottle with the conviction that an inkbottle is not present and (3) suspend judgment between (1) and (2).

In addition, X can also anticipate from the prior point of view what the risk of incurring error should be in retreating from the inconsistent state. There should be no risk of error incurred in retreating from inconsistency since none of the three options imports any belief at all and, hence, cannot import false belief. To be sure, from the point of view prior to expansion into inconsistency, X may assess what the risk incurred is of adding the information contained in standing by (1), (2) and (3). But the alternatives to be evaluated are not expansions or additions of information to the initial state but contractions of the inconsistent state $K_{\perp}$. Even if X is initially absolutely certain that the object is an ink bottle, X cannot use this conviction to argue for (1) over the other two alternatives as Unger seems to think. It should not matter whether X initially was certain that the ink bottle is present, or that it is absent or was in suspense. X is supposing for the sake of the argument that X has expanded into inconsistency and needs to give up one or the other claims (or both). In doing so, X does not import any false belief. X is not importing anything. X is *giving up* information. Avoidance of error is irrelevant as a concern.[12]

In contracting a state of belief by giving up information X would prefer, everything else being equal, to minimize the value of loss of the information X is going to incur. (1) or (2) should be favored depending upon which carries more valuable information according to X. If X regards both to be more or less equal in value, (3) should be favored. From the prior point of view, the inquirer can assess what the losses in informational value should be in adopting (1), (2) and (3) as retreats from inconsistency.

[11] Erik Olsson (2003) rightly objected that X cannot evaluate the informational values of the three options and take the decision between (1), (2) and (3) in the state of inconsistent conflict $K_{\perp}$ that results after the recalcitrant experience takes place. For this reason, I suggested (2003) that X anticipate the serious possibility of such conflict by providing a procedure for handling such issues before they arise.

[12] There is an exception to this observation. If one is a Messianic Realist like Popper or like Peirce sometimes appears to be, one might argue that X should be concerned to minimize risk of error as assessed according to X's prior point of view. If the ink bottle is fully believed to be present, X might then think that precommitting to the view that it is absent is deliberately courting error. Messianic Realism would support Unger's claim that if one is absolutely certain, one should never give up the conviction. I have discussed and argued against Messianic Realism in Levi (1980, 1991). In any case, the fault for dogmatism could in that case be placed on the shoulders of Messianic Realism rather than absolute certainty.

Whether X opts for (1) or not is independent of whether X was absolutely certain initially that an inkwell was present. It depends entirely on the value of the information that emerges from opting for (1) rather than (2). This evaluation could favor option (2) without in any way rendering it incoherent for X to have been absolutely certain beforehand that an inkwell was present. The same holds for (3).

Of course, some inquirers might assess losses of informational value in a manner that guarantees that the convictions in belief state K prior to expansion into inconsistent state $K_{\perp}$ will remain. If it became general practice to proceed in this manner, absolute certainty would lead to a dogmatic resistance to change. I cannot demonstrate the incoherence of such a view. I think the values such inquirers pursue are bad values even if they are coherent.

But the refusal to give up initial absolute certainty is predicated on a commitment to evaluating losses of informational values of options for retreating from inconsistency in a way according to which option (1) is preferred to options (2) and (3). It does not follow that such losses must be evaluated in this fashion. It is Unger's tacit commitment to a way of evaluating retreats from inconsistency that, to my way of thinking, fosters the dogmatic attitude he rightly deplores. Absolute certainty is not the culprit! The endorsement of questionable epistemic values is.

I have just offered an account of how agent X may reasonably cease to be absolutely certain that *h* and how X might recognize this as a serious possibility while being absolutely certain that *h*. There are other conditions under which X might legitimately contract X's initial state of full belief (see Levi 1980, 1991). One possible scenario ought to be enough, however, to undermine the dogma that absolute certainty fosters a dogmatic attitude. X may fully believe that *h* and yet recognize as a serious possibility that X will be justified in ceasing to be absolutely certain that *h* subsequently. Moreover, X may do this while remaining coherent in X's beliefs.

The thesis of *doxastic infallibilism* (I called it "epistemological infallibilism" in Levi, 1980) is entailed by the thesis that X's state of full belief or absolute certainty is X's standard for serious possibility. I do not wish to insist on my linguistic practice of calling it "infallibilism". Perhaps, it would be better to call it "infallibilism of the present point of view." The thesis of doxastic infallibilism is, so I have argued, to be distinguished from the thesis of *incorrigibilism* according to which X's state of full belief ought to be immune to legitimate change. Incorrigibilism has also been called infallibilism. Perhaps, we may call it "infallibilism of the future point of view." Fallibilism of the present and of the future alike should be distinguished from categorical fallibilism according to which for every X and every time *t*, all logical possibilities should be serious possibilities.

Categorical fallibilism is compatible with infallibilism of the present. The conjunction of the two, however, entails the thesis of incorrigibilism or infallibilism of the future. X's standard for serious possibility should be restricted to logical truths and whatever else may count as fixed *a priori* or conceptual necessity. This conjunction seems to capture accurately the views of radical probabilists who follow R.C. Jeffrey (1965). It also characterizes the views of those skeptics who maintain that we may fully believe and know logical (and *a priori* and conceptual) truths but nothing else. It also comes close to capturing the view of Peter Unger.

I favor rejecting categorical fallibilism while endorsing the thesis that full belief is the standard for serious possibility (which entails the thesis of infallibilism of the present) and the thesis of corrigibilism (or fallibilism of the future).[13]

12.5 Rationality and Commitment

As a rational agent, X not only undertakes to conform to a minimal standard of rational deliberation but is committed to policing X's full beliefs, partial beliefs, plain beliefs, value judgments and judgments concerning what is to be done to insure that X's performance meets standards of a minimal canon of rational deliberation. What those standards are or should be in detail is not the present concern.[14]

No one can, of course, satisfy the requirements specified even approximately or "by and large". One reaction to this is to weaken the demands on minimally rational full belief even more than I have indicated. Such a dumbing down of the standards of rationality is not going to be helpful. Either the standards will be completely trivialized or they will continue to be beyond the capacities of inquirers to satisfy in complex enough situations. The most decisive objection, however, seems to me to be that by weakening the standards of rationality, there is no rationally motivated incentive for agents to make any effort to reach the original standards of minimal rationality. There is no incentive to attempt to improve on one's doxastic performance by acquiring full belief in the consequences of full beliefs one already has.

I contend that all rational agents are rational because they are subject to critical scrutiny to the extent that they fail to meet standards for minimally rational full belief. It is not that X at *t* ceases to be rational if X fails as fail X must. But X should improve X's doxastic performance by improved training in logic and mathematics, by therapy to eliminate the distractions of psychological instabilities and by the use of automata and other prosthetic devices to enhance computational capacity when the need arises and costs and feasibility permit. And when the costs are

[13] All of these ideas are to be distinguished from claims of infallibility for persons or for their testimony as in the idea that the Pope is infallible when speaking *ex cathedra*. This conception of infallibility relates to sources of information. A source of information need not be infallible even on topics concerning which the source is approved in order to be a source of knowledge.

[14] For illustrative purposes, I take it that minimal rationality requires that X's state of full belief or the set of X's full beliefs at a given time should be consistent, it should contain all the consequences of X's state of full belief, that X should fully believe that X fully believes that *h* if and only if X fully believes that h, that X should fully believe that X does not fully believe that *h* if and only if X does not fully believe that *h* and that X should be opinionated as to whether X fully believes that *h* or does not do so.. Insofar as X's state of full belief is representable by a set of sentences in a regimented language DML with a belief operator B_{xt}, every such set is a deductively closed set satisfying the axioms of an S5 modal propositional system (see Levi 1997, ch. 5 for further elaboration).

overwhelming or satisfying the requirements is beyond X's capabilities, efforts should be made (costs and opportunities permitting) to overcome these obstacles.

To attribute belief that *h* to X at time *t* has sometimes been understood as claiming that X is undergoing a fit of doxastic conviction. For the past century and a half, it has become more fashionable to hold that believing that *h* is having a disposition or condition manifested by linguistic or by bodily behavior or by the aforesaid fits of doxastic conviction. I think we should be tolerant of both usages of "believes". But there is another important interpretation of "belief that *h"*. Indeed, it seems to me to be the central one.

Although Donald Davidson was obviously wrong in maintaining that agents "by and large" conform to the requirements of rational belief, valuation and decision making in their dispositions and behavior, I think he was quite right to maintain that the principles of rationality (at least of a minimal rationality) are "constitutive" of attitudes of full belief, partial belief and other attitudes relevant to deliberation (Davidson 1980, Levi 1997, 1999). By "constitutive"I mean "axiomatic". But the axioms of rationality that regulate, say, full belief are false when full belief is understood dispositionally or phenomenologically.

If the doxastic dispositions are interpreted "naturalistically" it is unclear how to provide intentional contents to the dispositions. And if we do by some legerdemain manage the trick, belief so construed will often and egregiously fail to satisfy the requirements of rationality. But if we construe full belief in what I take to be the central sense, X believes that *h* if and only if X has undertaken a commitment to behave doxastically in a manner that conforms to the demands of minimal rationality for full belief combined with a specific commitment within that framework to full belief that *h*. It then becomes appropriate to examine X's doxastic performances to determine how well they measure up to X's doxastic commitments.[15]

According to that interpretation, when X fully believes that *h*, X is committed to being disposed to the behaviors associated with the dispositional sense of belief and to having doxastic commitments to believe the logical consequences of belief that *h*. And in this sense, it may be said that X's set of full beliefs at a time *t* are closed under logical consequence and meet the requirements of the S5 logic mentioned above (see Levi 1997, ch. 5). This set represents X's state of full belief at *t* or X's state of doxastic commitment. It is in the commitment sense that X's beliefs have intentionality.

X's beliefs in the dispositional sense or their manifestations in linguistic and overt behavior or in fits of doxastic conviction are to be understood as successful or unsuccessful attempts to fulfill doxastic commitments. Their intentionality is identified with the commitments they attempt to fulfill. Of course, there may be behaviors and phenomenal episodes that are not attempts to fulfill doxastic commitments. X may utter "The cat is on the mat" without saying anything. Such behaviors lack intentionality and are not relevant to the present discussion.

[15] These remarks summarize my reaction to Cohen's ideas concerning his distinction between belief and acceptance (Cohen 1992).

12.6 Justifying Change in Doxastic Commitment and Doxastic Performance

The primary task of epistemology ought to be to give an account of how inquiring agents ought rationally to justify changes in their states of full belief. That is to say, the task is to give an account of conditions for justifying changes in doxastic *commitment*.

There is no doubt another task pertaining to changes in full beliefs. It pertains to improving an agent's performance as an inquirer by enhancing the agent's capacity to fulfill his or her commitments. Improving performance involves changing doxastic dispositions or doxastic manifestations. This calls for skills grounded in clinical psychology, logic and theories of computability and other empirically grounded disciplines. Of course, the normative standards of interest in epistemology ought to guide the activities of the clinicians and the understanding of what is feasible uncovered by the clinicians ought to be taken into account in the prescriptions formulated by epistemologists. And it would be silly and pernicious to draw professional boundaries between epistemology, psychology, the social sciences and logic. Acknowledging this ought not to prevent us from insisting that there are important distinctions between an agent's doxastic (and other attitudinal) commitments and performance.

Let X be committed to full belief in a set of premises $P_1, \ldots, P_n$. Let C be a deductive consequence of these premises. As long as X remains committed to full belief in each of the set of premises, X is also committed to full belief that C whether or not X recognizes that C is a consequence of these premises.

Suppose that prior to recognizing the validity of the deductive argument, X recognized in doxastic performance X's commitment to fully believe $P_1,\ P_2, \ldots, P_n$ in the sense that X was disposed upon interrogation to assent appropriately. But X was not prepared to assent to C.

X's doxastic performance is improved when X recognizes that C is a deductive consequence of the premises and adjusts X's behavior accordingly.

The premises together with the deductive inference do not justify full belief that C. There is no need to justify such full belief. X is already committed to full belief that C. The deductive argument does justify X's adjusting X's doxastic performances so as to fulfill X's commitment to full belief that C (see Levi 1980, ch. 1, 1991, ch. 2, 1997, ch. 1).

Adjustments in doxastic performance are justified when they bring X's doxastic behaviors into conformity with X's doxastic commitments.

Changes in doxastic commitment are warranted in one of the following ways:[16]

(a) *Routine expansions*; Agent X implements a program for utilizing inputs to form new full beliefs to be added to X's state of full belief K – that is X's initial state

[16] These four ways of changing doxastic commitment are considered in Levi (1980, 1991, 2004).

of doxastic commitment. Observation and relying on the testimony of witnesses and other sources of information are the main ways this is done. Typically such programs are not deliberately chosen except when the programs that have been used are judged inadequate in one way or another and some modifications need to be made. The modifications to be made to the programs can become then a matter of deliberate choice. But the new beliefs added in expansion via implementation of a program of routine expansion to the state of full belief are not chosen by the agent. They are selected in response to the inputs by the program. As long as the inquirer X is convinced that the program produces full beliefs that are likely to be true and are informative, X is justified in expanding in accordance with the program.

Changes in doxastic commitment of type (a) are responses to inputs in conformity with programs for routine expansion. They are not to be confused with the changes that bring doxastic performance into conformity with doxastic commitment. Prior to the inputs and the response to them, there is no doxastic commitment with which to conform. The programs for routine expansion are not rules of inference. The inputs are not beliefs or premises from which the output is inferred but are events like sensory stimuli or testimony of others to which the inquirer responds by forming new doxastic commitments. Hence, the change in doxastic commitment that is the result of routine expansion is not justified inferentially. On the other hand, it is not justified immediately either for the legitimacy of the belief acquisition presupposes the reliability of the program for routine expansion and this presupposition is at least tacitly part of X's initial doxastic commitment. As long as the program is one that X is committed to fully believing in advance to be a reliable and informative way of harvesting information from X's environment, X's change in doxastic commitment is a justified response to the implementation of the program..

(b) *Deliberate expansions*: Whereas in routine expansion, the inputs determine what answer to a question is adopted, in deliberate expansion, the answer chosen is justified by showing it is the best option among those available given the cognitive goals of the agent.

Thus a justified change in full belief by deliberate expansion depends not only on the initial state K of full belief (X's evidence) but also on the ultimate partition U_{K} (constructed by abduction) and the algebra of potential answers generated by U_{K}, the set of permissible credal probability distributions over U_{K} according to X's confirmational commitment C and state of full belief K, X's assessment of the informational values of potential answers as determined by a set of informational value determining probability distributions over U_{K} and an index of boldness that represents the relative weight assigned to risk of error in expansion and informational value acquired. The justified change is the one that is best among the available options (relevant alternatives) according to the goal of seeking new error free and valuable information.

(c) *Contraction* in response to inadvertent expansion into inconsistency through routine expansion. The contraction is determined by precommitment to a plan for retreat from inconsistency that recommends the best contraction strategy that minimizes loss of informational value. That contraction is justified as the best option for minimizing loss of informational value.
(d) *Contraction* to give a hearing to a proposition currently judged certainly false which, however, is recognized as having merits as a worthwhile explanatory hypothesis.

Whether such contraction is warranted depends upon the expected benefits in subsequent inquiry of giving the new hypothesis a hearing as compared to remaining with the status quo.

In all four types of changes, what is justified or legitimate is a change in doxastic commitment. *None* of the justifications is an inference from premises to a new belief. In case (a), a change in doxastic commitment that is a response to inputs is legitimated. In cases (b–d) a choice of a change in doxastic commitment is justified by showing that it best promotes the goals of the problem of expansion or contraction among the options available. In none of the cases does it make sense to speak of a belief being justified.

Thus, there are two sorts of changes that may be justified changes in belief:

a. Changes in doxastic commitment.
b. Changes in doxastic performance that bring the performances into conformity with the inquirer's commitments.

Those who insist that to be knowledge current beliefs should be justified tend to deny the significance of the distinction between changes in doxastic commitment and changes in doxastic performance. Let the current state of full belief be represented by a set of sentences in some language L closed under logical consequence. According to foundationalists, the demand that current beliefs be justified is satisfied when the set S_K of sentences may be organized so that there is basis B_K and all other sentences in the set are derivable either deductively or by some legitimate non deductive form of reasoning from B_K. The elements of B_K are then alleged to be justified in some non inferential fashion. According to this foundationalist model, there may be a distinction between inferential justification and non inferential justification. But there is no distinction between commitment and performance. For example, induction which is a change in doxastic commitment on the account I am proposing is, according to the foundationalists, an inference from B_K and items logically entailed by B_K in accordance with inductive rules of inference of some kind. Conformity with these rules is no more a change in doxastic commitment than conformity with the principles of deductive logic.

There are, of course, authors who question the point of such a justificational structure as I do but who require the total set S_K to meet some test of coherence in order to be justified. Once again justification is of belief or at least of systems of beliefs but not of changes in belief and while the difference between inferential and non inferential justification may be called into question, no distinction between commitment and performance is recognized.

Needless to say those who would naturalize epistemology have no use for justifying changes in doxastic commitments and, indeed, for the distinction between commitment and performance. Nor do the various species of skeptics of these epistemological programs.

There is an approach that may appear similar to the one I favor but ought I think to be distinguished from it. Michael Williams advocates justifying current beliefs but maintains that justification exhibits a "default and challenge structure" of the sort discussed by Robert Brandom (1994) (Williams 2001, p. 149) As I understand this view, the inquirer X may at a given time have no justification for some and, perhaps, all of X's current beliefs. But X may, nonetheless, be granted a default justification for these beliefs.

However, when X's beliefs are challenged, X is obliged to come up with a justification for the beliefs.

When I claim, as I do, that X need not justify current beliefs but only changes in beliefs, this may seem to differ only verbally from the default and challenge model. That is not so. As Peirce pointed out, raising a question about a current belief is not sufficient to provoke an inquiry aimed at justifying the belief or giving it up. The inquirer must be given a good reason for giving up the belief that he or she has. It may be just as hard to justify ceasing to believe that *h* when one believes it as it is to justify coming to believe it when one initially does not. The default and challenge model, by way of contrast, presupposes that the inquirer should be responsive to any challenge.

What all views of these types share in common is that it is beliefs rather than changes in belief that are the primary target of justification (or explanation) and that the distinction between commitment and performance as I have drawn it is an untenable dualism.

To repeat, the view I favor focuses on justifying *changes in doxastic commitments* and thus offers a different view of what epistemology should be about than any of the alternatives mentioned.

Thus, insisting upon a distinction between doxastic commitment and doxastic performance is not optional for those who suggest that justification is required for changes in doxastic commitments but not for current doxastic commitments. Deductive arguments are not justifications of changes in doxastic commitments at all. And they are not justifications of current doxastic commitments. Deductive arguments are deployed in efforts to improve doxastic performance by showing how current commitments are to be fulfilled.

12.7 Knowledge as True Belief

When knowledge is a species of belief, the verb "to know" is an honorific. X knows that *h* if and only if X fully believes that *h* and X's belief that *h* is worthy. It is pointless to pretend that there is a uniquely correct definition of worthiness of belief except by reference to an ideal for distinguishing between beliefs to be prized and beliefs that are to be despised. I maintain that knowledge is true belief not because I think that I thereby capture the "correct" meaning or linguistic usage. I advocate an epistemic value commitment according to which beliefs are prized when they are true and carry valuable information (in virtue of being fully believed). Others may wish to work out different epistemic ideals that support different characterizations of knowledge as a species of belief. What makes Craig's approach so interesting is that he undertakes to do just that. Were it not for the fact that sources of information that are sources of knowledge need not be informants – or so I think, Craig's program would constitute a serious rival to the dualist position that acknowledges a difference between knowledge that entails belief and is a standard for serious possibility and sources of knowledge (i.e., certified sources of information) that may but need not be sources of information that are believers.

The appraisal of the value of information is a matter about which there can be a wide range of different views. Information may be counted as valuable because it carries explanatory power according to the standards of some research program or other, it may display some other allegedly cognitive virtue or it may fit in with a given vision of the moral order. The minimal requirement that must be met is a condition of weak positive monotonicity: if X's current state of full belief K and *h* entail *g* but K and *g* do not entail *h*, *h* carries more information than *g* and carries at least as much informational value. According to this vision of informational value, expanding K by adding *h* can never bring X into a state of full belief carrying less informational value than K.

Let X be in state of full belief K and contemplate adding *h* to K. In addition, let X suppose that *h* is true. Under that supposition, X can do no better than to add *h* to K. To be sure, X might increase informational value by adding other propositions as well. But X is not seeking to maximize informational value but to maximize informational value while avoiding the importation of false belief. No matter what X does, X should not pass up the opportunity to obtain a new belief error free. It is in this sense that true full belief is to be prized rather than despised. It is in this sense, that true full belief deserves the honorific title of knowledge.

Prior to expansion, the inquirer ought to identify and evaluate the options for expansion available as determined by the ultimate partition. These are the "relevant alternatives" critical to determining what is justified as a change in doxastic commitment. These alternatives are evaluated with respect to two desiderata: risk of importing false belief and the value of the new information afforded by the expansion.

After expansion by, let us say, adding *h* and the consequences of K and *h*, the comparison of the answer adopted with alternatives is no longer on the agenda. The inquirer X has the full belief that *h*. Perhaps, additional inquiry will yield even more informative conclusions; but as far as X is concerned, the information conveyed by *h* settled. Moreover, X is committed to judging *h* and all its consequence to be true with absolute certainty. From X's point of view, X has true full belief that *h*. Although, perhaps, looking for additional information may be worthwhile, the information X has is, given X's goals, as good as it gets. X knows that *h*.

The bulk of epistemologists adopt some variant of an alternative epistemological ideal. They may acknowledge the desirability of obtaining new error free information or they may not. But even if they do, they insist that ideally the new error free information must be acquired in a legitimate or non accidental manner. Otherwise it is not knowledge.

The dispute is not or ought not to be over the semantics or pragmatics of the verb "to know". The debate as to how to construe "X knows that *h*" ought to be a dispute over the goals that ought to be pursued or the values that ought to be promoted in inquiry and not a dreary dispute over dubious linguistic intuitions.

In seeking new error free information via deliberate expansion, X should choose that potential answer from those available that best promotes that aim where the criteria for what best promotes that aim are constrained by principles of rational choice. I also think that any proximate aim that can be characterized as seeking new error free information promotes cognitive goals. On these assumptions, X's choice of an answer should be rationally justified by showing that that choice best promotes the aim of obtaining new error free and valuable information given the information available to X according to the initial state of full belief K. If by this standard X is justified in expanding K by adding *h* and all the consequences of K and *h* according to X's initial state of full belief, X's *changing* from K to K^{+}_{h} is justified.

On this account, the justification is a justification of a change in state of full belief. It is not a justification of a full belief or of a state of full belief.

Consider, however, a change in state of full belief by expansion from K to K^{+}_{e} that is justified or warranted or legitimated. Suppose *h* is a consequence of K^{+}_{e} but not of K. If the expansion has been implemented, from X's new point of view, X fully believes or is committed to full belief that *h* and it is true that *h*. X may also judge that X's acquisition of belief that *h* was the product of justifiable change in belief. From X's new point of view, X has come to know that *h*.

The justification of the change in belief represented by the expansion of K by adding e and all deductive consequences is not a justification of X's full belief that *h* according to the demands of pedigree theories of knowledge. To do that requires justification of all the beliefs in K and the new information that e. And the circumstance that the change in X's state of full belief was justified is not relevant after X has implemented the change to whether from X's new point of view, X knows that *h*.

Critics of the Knowledge as True Full Belief thesis worry that X's new belief that *h* is a lucky guess or is an accident. These are legitimate concerns if one wishes to use X as an authority and trust X's utterances as reliable testimony. But as already noted, the reliability of X's utterances as testimony has little bearing on whether X's belief that *h* is to be prized or not by X as knowledge.

Of course, most epistemologists maintain that something additional to true belief is required for knowledge. They maintain that even when the focus is not on X's status as an authority, lucky guesses and accidentally correct belief formations ought not to count as knowledge. They contend that true belief that is not justified or otherwise certified ought not to count as knowledge.

I take this to mean that if Y recognizes that X has true belief *h* that is not justified, Y should urge X to remove *h* from X's full beliefs pending certification.

Y cannot take the charitable view that X may retain the belief that Y is certain is true. How can X look for justification or be provided with justification while X is absolutely certain that *h* is true? Y should insist that X cease believing that *h* pending further inquiry. In my opinion, this attitude is mean spirited and should be the object of scorn.

Nonetheless, the intuitions about knowledge invoked in discussions of the Gettier predicament provide ample evidence for how widespread this kind epistemic nastiness can be.

Let X, at initial stage *t* when X's state of full belief is K, recognize that expanding by adding e is justified. Whether or not it is justified, X will then also be justified in adding to e∨f. At the new stage *t′* when X's state of full belief is $K^{+}{}_{e}$, X is committed to full belief that e is true and so is e∨f. As far as X is concerned, no error was imported. X not only fully believes that e but also fully believes that e∨f is true. As far as X is concerned, X has true belief that e and that e∨f. X also has justification for the expansion. Whether X demands justification as a condition for knowledge or not, from X's point of view, X knows both that e and that e∨f.

Let Y at the second stage be certain that e is false. Y is also certain that f is true and, hence, that e∨f is true. Y agrees with X that the expansion was justified. According to the Gettierites, Y should judge that X does not know that e and also that X does not know that e∨f.

Y's contention that X does not know that e seems appropriate. But Gettierite Y denies that X knows that e∨f even though Y believes that X has true belief. Y should urge X to give up this belief pending X's finding a certification for e∨f. Gettierite Y recommends that X cease believing that e∨f because X obtained this conviction via derivation from a false belief even though both X and Y are convinced that e∨f is true.

Agent Y who agrees that knowledge is true full belief is not committed to such epistemological mean spiritedness. The non Gettierite Y can say that X failed to avoid error in expanding by adding e to K and, hence, failed to come to know that e. But if Y is convinced that e∨f is true anyhow, Y can have the generosity of spirit to urge X to retain e∨f. According to Y, X is entitled to claim knowledge that e∨f.

To my way of thinking, this use of the Gettier problem to display the grudging nature of conceptions of knowledge as requiring an extra condition beyond true belief is the most compelling reason one could have to endorse the Knowledge as True Full Belief Thesis.

12.8 Is Knowing an Attitude?

The Knowledge as True Full Belief Thesis states:

> *Knowledge as True Full Belief*: All rational agents ought to fully believe and, hence, agree that X knows that *h* if and only if X fully believes that *h* and it is true that *h*.

Timothy Williamson maintains that knowing that *h* is a propositional attitude. So is believing that *h*. On the other hand, Williamson maintains that truly believing that *h* is not a propositional attitude. Williamson contends that there is no mental state being in which is necessary and sufficient for believing truly that it is raining (Williamson 2000, p. 27). Propositional attitudes are mental states. So true belief cannot be a propositional attitude. Given that knowing that *h* is, according to Williamson, a propositional attitude, knowing that *h* cannot be equated with believing that *h*.

Williamson offers a general characterization of his understanding of mental states:

> If S is a mental state and C a non-mental condition, there need be no mental state S* such that, necessarily, one is in S* if and only if one is in S and C obtains (Williamson 2000, p. 28).

Williamson includes more than propositional attitudes among mental states; but propositional attitudes are, according to him, mental states. So are pleasure and pain which are not propositional attitudes and need not have any intentionality in them.

Williamson's taxonomy is not appropriate to the study of justifying changes in states of full belief.

X's full belief that *h* is not a state of full belief that X is in at a given time in the sense in which a state of full belief is a state of doxastic commitment. An agent X is in a state K of full belief at time *t* which is one of a set of potential states of full belief – states that are conceptually accessible to X. Each potential state of full belief is a coherent state of doxastic commitment. I assume that the set of such states constitute a Boolean algebra. Motivation for this technical condition is given in terms of a conception of the goals of inquiry in Levi (1991). According to this view, potential states of full belief are partially ordered with respect to the information they carry or the amount of doubt they remove. Potential state K_2 is a *consequence* of state K_1 if and only if being in state K_1 removes more doubt than being in state K_2.

To say that X fully believes at *t* that *h* is equivalent to saying that X is in a potential state of full belief that has the potential state that *h* as a consequence. Full belief is not a relation between the subject and a proposition. It is a relation between the state of full belief X is in and a potential state of full belief that X is not in (unless the potential state that *h* is X's state of full belief) but which is a consequence of the state X is in. Different beliefs that X has at *t* are not different mental states of X at that time. There is only one state of full belief at *t*. The different beliefs are different consequences of that state.

Attitudes such as fully believing that *h* are ineradicably normative. This is so whether believing that *h* is taken as a doxastic undertaking or commitment of an agent at a given time or as a performance that attempts to fulfill the commitment. Pleasure and pain, itches, etc. may be mental in some sense or other but they lack this crucial normativity.

On this account the primary bearers of truth values are not sentences, propositions or sets of possible worlds but potential states of full belief. X truly and fully believes at *t* that *h* if and only if X is in a state of full belief at *t* that has that *h* as a consequence and where it is true that *h*. The state of full belief X is in remains the same whether *h* is true or is false.

To say that X knows at *t* that *h* is, according to the view taken here, equivalent to saying that X truly and fully believes that *h*. And attributing knowledge that *h* to X does not alter the attribution of a state of full belief to X.

What are the truth conditions for the potential state of full belief that *h*? Each inquirer is in some state of full belief. While in state $K_{X,t}$, X is committed to judging all (and only) consequences of $K_{x,t}$ true and judging the Boolean complements (or negations) of these consequences false. If K* is not assigned a truth value in this manner, potential states whose meet with $K_{X,t}$ have K* as a consequence provide sufficient truth conditions for K* and potential states that are consequences of the meet of $K_{X,t}$ and K* are necessary truth conditions for K*.

It may, perhaps, be objected that the truth conditions so constructed are relative to X's belief state at *t*. They will be different for Y's belief state. If the "meaning" of a sentence that expresses K* is given by such truth conditions, the meaning will be too unstable to be useful in communication between X and Y in joint inquiry.

Communication between two or more inquirers is often challenging. But when the difficulties are overcome it is because the inquirers agree to use the join of their belief states as the common ground for spelling out truth conditions. It is not because of some standard usage in the OED or theoretical semantical principles. The point I am belaboring is that insofar as we find the specification of truth conditions urgent, these conditions are constrained by states of full belief just as standards for serious possibility are.

According to X at *t*, all consequences of $K_{X,\mathrm{t}}$ are true. Hence, according to X at *t*, the set of X's full beliefs at *t* coincides with the set potential states X knows at *t*. From X's point of view at *t*, X's state of full belief at *t* coincides with X's state of knowledge at *t*.

Y at t' with state of full belief $K_{Y,t}$, will pass a different verdict on the consequences of $K_{X,t}$. Even if Y agrees with X concerning the identification of X's state of full belief, Y will, in general, assess the consequences of $K_{X,t}$ differently. Y will concede that X has some knowledge but insist also that X has false beliefs. And Y may be in suspense as to whether some of X's full beliefs are knowledge.

Both X and Y can agree, however, that X at *t* knows that *h* if and only if X fully and truly believes that *h*.

The approach to the characterization of knowledge I have taken is an offshoot of my focus on the question of justifying change in full belief. I think there is more to an account of inquiry than justifying belief change – much more. But obtaining a grip on changing states of full belief, which I contend should be understood as changing states of doxastic commitment is central to a study of change in probability judgment, value judgment and the growth of knowledge. Williamson's states of mind in general and the attitudes of belief and knowledge in particular do not even recognize a distinction between doxastic commitment and doxastic performance. Such states of mind are not the sorts of states that seem relevant to epistemology as I understand it.

Throughout this discussion, I have avoided invoking either an externalist or an internalist view of full beliefs or their contents (whatever “externalist” or “internalist” may mean). I have taken the position that the truth values of beliefs and truth conditions for them are judged from the points of views of the inquirers doing the judging. This does not mean that truth is relative but only that judgments of truth are expressions of the viewpoints of those who make the judgment.

Williamson takes different stands on the question of internalism and externalism about contents and the attitudes and the “first personal” approach to judging truth than I do. For this reason, I cannot claim to have refuted Williamson. But I have explained why I am not caught up in the toils of Williamson’s argument for resisting the identification of knowledge with true full belief which purports to show that the former but not the latter are propositional attitudes. If a propositional attitude is a relation between subject and proposition then neither knowledge that *h* nor true full belief that *h* is a propositional attitude. The bearer of truth value is the potential state of full belief that *h* and not the proposition that *h* whatever that may be. And although full belief that *h* is a potential state of full belief, the very same potential state as knowledge that *h* is in case the full belief that *h* is true, the subject’s state of full belief as distinct from these potential states. And, in any case, states of full belief are states of doxastic commitment and, hence, normative.

12.9 Depositing Paychecks

Consider the by now overused bank examples discussed by Keith De Rose (1992).

> *Case A*: Mr and Mrs. X agree that depositing their paychecks today (Friday) would be an inconvenience due to the long lines. Mr. X suggests that they drive home and deposit the checks on Saturday. Mrs X points out that many banks are closed on Saturday. X says, “I know the bank will be open.” X reports that he was at the bank a couple of weeks ago on Saturday and it was open until noon.
>
> *Case B*: The scenario is the same as before except that the couple have written a very large and important check that might bounce if funds are not deposited in their checking account before Monday. After X declares his conviction that the bank will be open on Saturday and gives his testimony that it was open two weeks ago, Mrs. X asks: “Banks after all do change their hours. Do you know the bank will be open tomorrow?” Remaining as confidant as he was that the bank will be open, X replies “Well no. I’d better go in and make sure.”

In both cases, the bank is open on Saturday so that X’s conviction was correct.

De Rose sees a prima facie discrepancy between these two cases that can be explained away by maintaining that X’s declaration that he does know in case A and admission that he does not know in case B are uttered in different contexts of knowledge attribution. X’s claim to know in case A is a different proposition than the proposition denied when X concedes he does not know in case B. In both cases, De Rose maintains that X speaks truly. De Rose offers these scenarios as intuition pumps in support of contextualist conceptions of knowledge that have acquired a widespread vogue.

Jason Stanley recognizes that the contextualism on offer here is primarily a linguistic thesis, explores its merits as such and finds it wanting. Yet he takes the cases seriously and suggests an alternative account according to which the general truth conditions for knowledge attribution are the same in both cases but the truth of a knowledge attribution is relative to several factors including one controlled by practical interests. In case A, the practical interests make the risks of being wrong in waiting until Saturday to deposit the checks fairly small. So it is true that X knows that the bank will be open. But the stakes are higher in case B. The risk of being wrong in waiting until Saturday to make the deposit is too great to take. So X does not know that the bank will be open.

Both Stanley and De Rose have come up with different rationalizations of what they take to be real phenomenon typified by cases A and B.

I have my doubts as to whether there is apparent inconsistency to rationalize. Recall in both cases, X and Mrs.X are making a joint decision so that the deliberation as to whether to go to the bank on Friday or Saturday is one that, if possible, should terminate in a consensus as to what to do. The knowledge that matters for decision making is that of the joint agent constituted by X and Mrs. X together. Moreover, a clear headed X would not claim to know in case A and admit to not knowing in case B. In both scenarios, X is certain that the bank will be open on Saturday. Mrs. X is not. In case A, X, in spite of his full belief, which he expresses in the declaration that he knows that the bank will be open, offers a reason for believing that the bank will be open not because he needs one (he does not) but to offer considerations that might persuade Mrs. X to agree with him. Apparently in case A, X succeeds.

In case B, Mrs. X is not mollified by X's initial argument seeking to persuade her to agree with him. She mentions that banks sometimes change their hours of business. X sees that he is not going to persuade Mrs. X. So he suggests that they check and go out whether the bank will be open on Saturday. Both Stanley and De Rose claim X confesses that he does not know that the bank will be open tomorrow. Mr.X must be extremely browbeaten to confess to that. A far more likely scenario is that he agrees to check things out rather than engage in a hopeless debate with his wife. He does not say: "I guess I did not know after all. Let us check it out." He says simply, "Let us check it out." X is just as certain in case B as in case A. From X's point of view, X knows that the bank will be open tomorrow in both cases.

In short, I think the two scenarios fail to provide the kind of data for pumping intuitions that De Rose, Stanley and many others think they do. The dispute between contextualists and interest relative invariantists is a tempest in a teapot built on appeal to understandings of the verb "to know" that many of us do not share and do not pretend to comprehend.

I do acknowledge that fully articulate criteria for justifying expansions of states of full belief depend not only on the initial state of full belief (the "evidence) of the inquirer and the probability judgments supported by the initial state. In addition one appeals to the set of potential answers to some question generated by an ultimate partition that provides a set of relevant alternatives, an assessment of the informational values of these potential answers and a level of boldness representing the relative importance of avoiding error and obtaining new valuable information.

These contextual parameters control the recommendation of a potential answer to be added to the initial belief state K. They are integral to the models of inductive expansion I proposed in Levi (1967a), modified in Levi (1967b) and modified and elaborated further in subsequent work. But they do not constitute conditions to which the attribution of knowledge prior to expansion or subsequent to expansion must be responsive.

References

Brandom, R. 1994. *Making it explicit*. Cambridge, MA: Harvard University Press.

Cohen, L.J. 1977. *The probable and the provable*. Oxford: The Clarendon Press.

Cohen, L.J. 1992. *An essay on probability and acceptance*. Oxford: The Clarendon Press.

Craig, E. 1990. *Knowledge and the state of nature*. Oxford: Oxford University Press.

Davidson, D. 1980. *Essays on actions and events*. Oxford: Oxford University Press.

De Rose, K. 1992. Contextualism and knowledge attribution. *Philosophy and Phenomenological Research* 52:913–929.

Dubois, D., and H. Prade. 1992. Belief change and possibility theory. In *Belief revision*, ed. P. Gärdenfors, 142–182. Cambridge: Cambridge University Press.

Dudman, V.H. 1983. Tense and time in English verb clusters of the primary pattern. *Australasian Journal of Linguistics* 3:25–44.

Dudman, V.H. 1984. Conditional interpretations of if-sentences in English. *Australasian Journal of Linguistics* 4:143–204.

Dudman, V.H. 1985. Towards a theory of predication in English. *Australasian Journal of Linguistics* 5:143–193.

Gärdenfors, P. 1986. Belief revision and the ramsey test for conditionals. *Philosophical Review* 95:81–93.

Gärdenfors, P. 1988. *Knowledge in flux*. Cambridge: Cambridge University Press.

Gibbard, A. 1981. Two recent theories of conditionals. In *Ifs*: *Conditionals, belief, decision, chance and time*, eds. W.J. Harper et al., 211–247. Dordrecht: Reidel.

Hacking, I. 1967. A slightly more realistic personalist probability. *Philosophy of Science* 34: 311–325.

Jeffrey, R.C. 1965. *The logic of decision*. New York, NY: McGraw Hill.

Kitcher, P. 2006. The knowledge business. In *Knowledge and inquiry*, ed. E. Olsson, 50–64. Cambridge: Cambridge University Press.

Levi, I. 1967a. *Gambling with truth*. New York, NY: A. Knopf, paperback edition in 1973 by MIT Press.

Levi, I. 1967b. Probability kinematics. *British Journal for the Philosophy of Science* 18:197–209.

Levi, I. 1969. If Jones only knew more. *British Journal for the Philosophy of Science* 20:153–159.

Levi, I. 1976. Acceptance revisited. In *Local induction*, ed. R. Bogdan, 1–71. Dordrecht: Reidel.

Levi, I. 1979. Serious possibility. In *Essays in honour of Jaakko Hintikkai*, 219–236. Dordrecht: Reidel, reprinted in Levi (1984), 147–161.

Levi, I. 1980. *The enterprise of knowledge*. Cambridge, MA: MIT Press. 2nd paperback edition, 1983.

Levi, I. 1983. Truth, fallibility and the growth of knowledge. In *Language, logic and method*, eds. R.S. Cohen, and M.W. Wartofsky, 153–174. Dordrecht: Reidel, Reprinted as chapter 8 of Levi (1984).

Levi, I. 1984. *Decisions and revisions*. Cambridge: Cambridge University Press.

Levi, I. 1988. The iteration of conditionals and the Ramsey test. *Synthese* 76:49–81.

Levi, I. 1991. *The fixation of belief and its undoing*. Cambridge: Cambridge University Press.

Levi, I. 1996. *For the sake of the argument*. Cambridge: Cambridge University Press.

Levi, I. 1997. *The covenant of reason*. Cambridge: Cambridge University Press.

Levi, I. 1999. Representing preferences, Davidson on rational choice. In *The philosophy of Donald Davidson, library of living philosophers* XXII, chapter 23, 531–570. Chicago and LaSalle, IL: Open Court.

Levi, I. 2003. Contracting from epistemic hell is routine. *Synthese* 135:141–164.

Levi, I. 2004. *Mild contraction: Evaluating loss of information due to loss of belief*. Oxford: Oxford University Press.

Olsson. E.J. 2003. Avoiding epistemic hell: Levi on pragmatism and inconsistency. *Synthese* 135:119–140.

Shackle, G.L.S. 1949. *Expectations in economics*. Cambridge: Cambridge University Press. 2nd edn, 1952.

Shackle, G.L.S. 1961. *Decision, order and time in human affairs*. Cambridge: Cambridge University Press. 2nd edn, 1969.

Spohn, W. 1990. A general non-probabilistic theory of inductive reasoning. In *Uncertainty in artificial intelligence*, vol. 4, eds. R.D. Shachter et al. Amsterdam: North Holland.

Spohn, W. 2006. Isaac Levi's potentially surprising epistemological picture. In *Knowledge and inquiry,* ed. Erik Olsson, 125–142. Cambridge: Cambridge University Press.

Stanley, J. 2003. *Knowledge and practical interests*. Oxford: Oxford University Press.

Unger, P. 1975. *Ignorance*: *A case for skepticism*. Oxford: Oxford University Press. Paperback edition 2002.

Van Fraassen, B. 1984. Belief and the will. *The Journal of Philosophy* 81:235–256.

Williams, M. 2001. *Problems of knowledge: A critical introduction to epistemology*. Oxford: Oxford University Press.

Williamson, T. 1996. Knowing and asserting. *The Philosophical Review* 105:489–523.

Williamson, T. 2000. *Knowledge and its limits*. Oxford: Oxford University Press.

Chapter 13
Reasoning About Belief Revision

Caroline Semmling and Heinrich Wansing

13.1 Introduction

The theory of belief revision developed by Carlos Alchourrón, Peter Gärdenfors and David Makinson (AGM) is one of the most influential and well-investigated theories of rational belief change; for a comprehensive presentation and references see Hansson (1999) and Rott (2001). This highly successful research program co-exists with another major research program concerned with the belief and knowledge of rational agents, namely doxastic and epistemic logic. With respect to epistemic logic, in *Knowledge in Flux* (1988), Peter Gärdenfors remarked:

> [M]y strategy is to "epistemize" the whole semantics, in the sense that I locate the epistemological machinery in the belief systems rather than in the object language. This does not mean that I have any aversion to epistemic logic—on the contrary. However, because I believe that the study of epistemic operators in a formal or natural language is not of primary concern for understanding the dynamics of knowledge and belief, I have chosen to keep the object language as simple as possible.

This choice certainly has its rationale and merits, but it might nevertheless be worth while also to explore a way of ascribing belief changes of (rational) agents and to state conditions on such changes within a suitable object language. Indeed, in the literature one can find various suggestions for reasoning about belief revision in a formal object language. Giacomo Bonanno (2005, 2007) develops a propositional modal logic of belief revision for a single agent using the following three modal operators:

$B_0\varphi$ at time 0 (initially) the agent believes that φ,
$I\varphi$ (between time 0 and time 1) the agent is informed that φ,
$B_1\varphi$ at time 1 (after revising her belief in view of the information received) the agent believes that φ.

C. Semmling (✉)
Institute of Philosophy, Dresden University of Technology, Dresden, Germany
e-mail: Caroline.Semmling@gmx.de

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_13,

The operators B_0 and B_1 have a standard relational possible worlds semantics, whereas a formula $I\varphi$ is true at a state (world) w iff the set of states related to w by the relation associated with I is exactly the set of states at which φ is true. Bonanno introduces three axioms that, together, characterize the *Qualitative Bayes Rule* (QBR). If $\mathcal{B}_0$, $\mathcal{B}_1$, and $\mathcal{I}$ are the binary relations associated with the three operators, $\mathcal{B}_i(w) = \{w' \mid \mathcal{B}_i(w, w')\}$ $(i = 0, 1)$, and $\mathcal{I}(w) = \{w' \mid \mathcal{I}(w, w')\}$, then the QBR says:

$$\forall w, \text{ if } \mathcal{B}_0(w) \cap \mathcal{I}(w) \neq \varnothing \text{ then } \mathcal{B}_1(w) = \mathcal{B}_0(w) \cap \mathcal{I}(w).$$

The axioms are:

Qualified acceptance $(I\varphi \wedge \neg B_0 \neg\varphi) \supset B_1\varphi$
Persistence $(I\varphi \wedge \neg B_0 \neg\varphi) \supset (B_0\psi \supset B_1\psi)$
Minimality $(I\varphi \wedge B_1\psi) \supset B_0(\varphi \supset \psi)$.

It is thus possible to express belief changes over time as a result of incoming information. Bonanno (2005, p. 219) emphasizes that "[p]revious modal axiomatizations of belief revision required an infinite number of modal operators", whereas just three operators are enough to characterize the QBR. However, in order to deal with receiving sequences of pieces of information, Bonanno introduces countably many belief and information operators. For every $t \in \mathbb{N}$, there are the following three modal operators:

$B_t\varphi$ at time t the agent believes that φ,
$I_{t,t+1}$ between time t and time $t+1$ the agent is informed that φ,
$B_{t+1}\varphi$ at time $t+1$ (in view of the the information received between t and $t+1$) the agent believes that φ,

and the generalized QBR is:

$$\forall w, \text{ if } \mathcal{B}_t(w) \cap \mathcal{I}_{t,t+1}(w) \neq \varnothing \text{ then } \mathcal{B}_{1+1}(w) = \mathcal{B}_t(w) \cap \mathcal{I}_{t,t+1}(w).$$

Whereas in Bonanno's approach belief change is expressed by using time indices for static belief operators, other approaches are based on *dynamic logic*.

In van Benthem's (1995) and de Rijke's (1994) *dynamic modal logic*, DML, for every formula φ, modal operators $[+(\varphi)]$ ("after every expansion by φ it is the case that") and $[*(\varphi)]$ ("after every revision by φ it is the case that") are defined. Valid principles of this logic are, for example, $[*\varphi]\varphi$ and $[*\varphi]\psi \supset [+\varphi]\psi$. The first formula may be interpreted as expressing that if a system of beliefs is revised by φ, then φ is believed.

In Krister Segerberg's *dynamic doxastic logic*, DDL (Leitgeb and Segerberg 2007; Segerberg 1999), the same reading is associated with $[*\varphi]B\varphi$, where B is a modal belief operator. In the dynamic approaches of van Benthem, de Rijke, and Segerberg, belief changes are treated as *generic actions*, i.e., action types. The idea is to be able to talk about the outcome of all or some executions of certain action types. The formula $[*\varphi]B\psi$, for instance, states that after every performance of a revision by φ it is the case that ψ is believed.[1]

In the present chapter, we will suggest another object language for reasoning about belief revision, namely a language of a propositional doxastic logic of concrete agency. This is a language not for describing the *outcome* of belief changes but for ascribing performances of certain actions, namely *changes of beliefs*. We will use a logic of concrete agency combined with belief, desire, intention, and other modal operators. One remarkable point is that in this language it is possible to express intentions and desires to form beliefs. We include static belief operators, but the addition of temporal modalities could be used to describe beliefs over time, in a similar way to what is done with Bonanno's time-indexed belief operators. In the first place, however, the formal object language of the present paper allows one to ascribe and reason about deliberate revisions of beliefs. This is a fundamental difference to previous suggestions for reasoning about belief revision in a logical object language.

The language of the logic to be presented extends the language of the deliberatively seeing-to-it-that operator from Stit Theory by belief operators for agents $\alpha_1, \ldots, \alpha_n$. If an agent α *expands* her beliefs by the proposition (expressed by) φ, this may be stated as $\alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi$ ("α sees to it that α believes that φ"). If α withdraws the proposition φ from her beliefs (in other words, *contracts* her beliefs by φ), this may be expressed as $\alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{bel}\colon \varphi$ ("α sees to it that α does not believe that φ"). Ascriptions of belief *revision* may then be obtained by a sort of symmetric Levi Identity: "α revises her beliefs by accepting φ" is expressed as $\alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{bel}\colon \neg\varphi \wedge \alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi$. Note that in the language which we shall deal with, also expansions, contractions, and revisions of intentions and desires can be ascribed:

$\alpha\, \mathit{dstit}\colon \alpha\, \mathit{int}\colon \varphi$ ("α sees to it that α intends that φ")
$\alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{int}\colon \varphi$ ("α withdraws the intention that φ")
$\alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{int}\colon \neg\varphi \wedge \alpha\, \mathit{dstit}\colon \alpha\, \mathit{int}\colon \varphi$ ("α revises her intentions by φ")
$\alpha\, \mathit{dstit}\colon \alpha\, \mathit{des}\colon \varphi$ ("α sees to it that α desires that φ")
$\alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{des}\colon \varphi$ ("α withdraws the desire that φ")
$\alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{des}\colon \neg\varphi \wedge \alpha\, \mathit{dstit}\colon \alpha\, \mathit{des}\colon \varphi$ ("α revises her desires by φ")

The language under consideration is the language of the *bdi-stit* logic developed in Semmling and Wansing (2008). It is interpreted in models based on the branching-time frames from Stit Theory. In this chapter we shall first present the semantic definition of *bdi-stit* logic (Section 13.2) and then develop a sound and complete tableau calculus for *bdi-stit* logic (Section 13.3). In view of the above suggested readings, the tableau calculus for *bdi-stit* logic may then be seen as a proof system for reasoning about belief revision. In Section 13.4 we shall consider a translation from the language of the AGM theory into the language of *bdi-stit* logic. It

[1] Another framework of interest in this connection is *dynamic epistemic logic*, DEL, see van Ditmarsch et al. (2005).

will turn out that not all of the AGM postulates are translatable, but some of the suggested translations of AGM postulates for rational belief change emerge as provable in *bdi-stit* logic.[2]

13.2 *bdi-stit* Logic

A motivating discussion of *bdi-stit* logic can be found in Semmling and Wansing (2008). In this section we shall just briefly recall the syntax and semantics of this logic.

13.2.1 The Syntax of bdi-stit *Logic*

The language of *bdi-stit* logic comprises denumerably many atomic formulas $(p_1, p_2, p_3, \ldots)$, the connectives of classical propositional logic $(\neg, \wedge, \vee, \supset, \equiv)$, and the modal necessity and possibility operators $\Box$ and $\Diamond$. We assume that $\Diamond$ is defined as $\neg\Box\neg$. This vocabulary is supplemented by action modalities and operators used to express the beliefs, desires, and intentions of arbitrary (rational) agents. Additionally there is a possibility operator $\Diamond\!\!\!\cdot\;$ taken over from Semmling and Wansing (2008). We also assume a set of agent variables $(\alpha_1, \alpha_2, \ldots, \alpha_n)$.

Definition 1 (*bdi-stit* **syntax**)

1. Every atomic formula p_1, p_2, $\ldots$ is a formula.
2. If α, β are agent variables, then $\alpha = \beta$ is a formula.
3. If φ, ψ are formulas and α is an agent variable, then $\neg\varphi$, $(\varphi \wedge \psi)$, $\Box\varphi$, $\Diamond\!\!\!\cdot\;\varphi$, α dstit: φ, α *bel*: φ, α *des*: φ and α *int*: φ are formulas.
4. Nothing else is a formula.

The reading of a formula α_i *dstit*: φ where $1 \leq i \leq n$ is "agent α_i deliberatively sees to it that φ". The formula α_i *bel*: φ is read as "agent α_i believes that φ" or "agent α_i has the belief that φ". In this vein also the readings of the desire operators α_i *des*: and the intention operators α_i *int*: are conceived.

What is particularly interesting about the language of *bdi-stit* logic is the combination of the action modalities α_i *dstit*: with the cognitive modalities α_i *bel*:, α_i *des*:, and α_i *int*:. Considering the expressive power of this language can be tied up with the discussion of several important philosophical problems related to the deontological conception of epistemic justification and the notions of responsibility for and blameworthyness of beliefs. Whereas the view that human agents actively form intentions is probably not extremely contentious, it is a matter of considerable debate, whether human agents are indeed capable of seeing to it that they believe certain propositions and capable of actively forming desires, and if so, whether this is a question

[2] We would like to thank Erik Olsson for inviting us to contribute to this volume and an anonymous commentator for her/his useful remarks on an earlier version of this chapter.

of direct or rather indirect voluntary control. For discussions and references to the literature, see, for example, Nottelmann (2007) and Wansing (2006). Note that the language of *bdi-stit* logic does not come with any commitments to specific views of the *psychology* of belief and desire formation. The semantics of *bdi-stit* logic, however, is such that ascriptions of belief, desire, and intention acquisition (alias expansion) are satisfiable.[3]

13.2.2 The Semantics of **bdi-stit** *Logic*

The semantics of *bdi-stit* logic is based on the indeterministic framework of branching temporal structures assumed in Stit Theory, see Belnap et al. (2001).

A *bdi-stit* model, used to interpret the formulas from Definition 1, consists of a frame $\mathcal{F} = (M, \leq, \mathcal{A}, N, C, B, D, I)$ together with a valuation v. The set M is a non-empty set understood as a set of moments of time, and the relation $\leq$ is a partial order on M. This relation $\leq$ of temporal precedence is reflexive, transitive but acyclic. Every moment in M has a unique $\leq$-predecessor; in other words, $(M, \leq)$ is a branching-time structure. Every maximal linearly $\leq$-ordered subset of M is said to be a history in $(M, \leq)$, and a pair $s = (m, h)$, where h is a history in $(M, \leq)$ and $m \in h$, is called a situation. If $m \in h$, the history h is also said to be "passing through" moment m. The set of all histories of (a given frame) $\mathcal{F}$ is denoted by H, and the set of all histories h such that $m \in h$ is denoted by H_m.

Let S be the set of all situations of $(M, \leq)$. The function N is a mapping from S to $\mathcal{P}(\mathcal{P}(S) \backslash \{\emptyset\})$, and $N(s) = N_s$ is called a *neighbourhood system* of situation s. In the following we will also denote by N the union of all sets N_s with $s \in S$, $N = \cup\{\, U \mid U \in N_s,\ s \in S\}$, i.e., the set of all neighbourhoods of $\mathcal{F}$. The set $\mathcal{A}$ is a non-empty, finite set of agents, and the functions B and D are mappings from $\mathcal{A} \times S$ to $\mathcal{P}(N)$. Then $B(\alpha, s) = B_s^{\alpha}$ $(D\,(\alpha, s) = D_s^{\alpha})$ is a set of sets of situations, where each set $U \in B_s^{\alpha}$ $(U \in D_s^{\alpha})$ is called a neighbourhood of situation s endorsing certain beliefs (desires) of agent α. The function I used to interpret intention ascriptions is a function from $\mathcal{A} \times S$ into N. Intuitively, $I(\alpha, s) = I_s^{\alpha}$ is the set of all situations compatible with what α intends at situation s.

Finally, the function C is a mapping from $\mathcal{A} \times M$ into $\mathcal{P}(\mathcal{P}(H))$ that assigns to every agent at each moment a family of sets of histories such that for every agent $\alpha \in \mathcal{A}$, the set $C(\alpha, m) = C_m^{\alpha}$ is an equivalence relation on the set H_m. The equivalence class $C_m^{\alpha}(h)$ contains the histories which are *choice-equivalent* to h for agent α at moment m. The idea is that agent α cannot distinguish at moment m by her or his actions between the histories from $C_m^{\alpha}(h)$. The equivalence classes $\{C_m^{\alpha}(h) \mid h \in H_m\}$ on H_m are also said to be the choice cells (or the actions available) for α at m. In particular, histories which share a moment later than m are choice-equivalent at m for any agent. If $s = (m, h)$, instead of $C_m^{\alpha}(h)$ we also write C_s^{α}.

[3] In a frequently cited paper, Bernard Williams (1973) claimed to have shown that deciding to believe is logically impossible. See also Winters (1979).

In the present branching-time framework, the agents are assumed to be independent of each other in the sense that at every moment, every agent must be able to realize any of her or his actions, no matter what choices are available to the other agents. Let $\mathcal{F} = (M, \leq, \mathcal{A}, N, C, B, D, I)$ be a frame and let $Select_m$ be the set of all functions σ from $\mathcal{A}$ into subsets of H_m, such that $\sigma(\alpha) \in C_m^{\alpha}$. $\mathcal{F}$ satisfies the *independence of agents* condition iff for every $m \in M$,

$$\bigcap_{\alpha \in Agent} \sigma(\alpha) \neq \emptyset$$

for every $\sigma \in Select_m$.

A valuation v on $\mathcal{F}$ is an arbitrary function mapping the set of atomic formulas into the power set of S and the set of agents variables into $\mathcal{A}$. A pair $(\mathcal{F}, v)$ is then said to be a *bdi-stit* model based on the frame $\mathcal{F}$. Satisfiability of a formula in a *bdi-stit* model is defined as follows.

Definition 2 (*bdi-stit* **semantics**)

Let $s = (m, h)$ be a situation in model $\mathcal{M} = (\mathcal{F}, v)$, let α be an agent variable, and let φ, ψ be formulas according to Definition 1. Then:

$\mathcal{M}, s \models \varphi$ iff $s \in v(\varphi)$, if φ is an atomic formula.
$\mathcal{M}, s \models \neg\varphi$ iff $\mathcal{M}, s \not\models \varphi$.
$\mathcal{M}, s \models \varphi \wedge \psi$ iff $\mathcal{M}, s \models \varphi$ and $\mathcal{M}, s \models \psi$.
$\mathcal{M}, s \models \Box\varphi$ iff $\mathcal{M}, (m, h') \models \varphi$ for all $h' \in H_m$.
$\mathcal{M}, s \models \Diamond\varphi$ iff there exists $U \in N_s$ with $U \subseteq \{ s' \mid \mathcal{M}, s' \models \varphi \}$.
$\mathcal{M}, s \models \alpha\, dstit\colon\ \varphi$ iff (i) $\{ (m, h') \mid h' \in C_s^{v(\alpha)} \} \subseteq \{ (m, h') \mid \mathcal{M}, (m, h') \models \varphi \}$,
(ii) $\mathcal{M}, s \models \neg\Box\varphi$.
$\mathcal{M}, s \models \alpha\, int\colon\ \varphi$ iff $I_s^{v(\alpha)} \subseteq \{ s' \mid \mathcal{M}, s' \models \varphi \}$.
$\mathcal{M}, s \models \alpha\, des\colon\ \varphi$ iff there exists $U \in D_s^{v(\alpha)}$ with $U \subseteq \{ s' \mid \mathcal{M}, s' \models \varphi \}$.
$\mathcal{M}, s \models \alpha\, bel\colon\ \varphi$ iff there exists $U \in B_s^{v(\alpha)}$ with $U \subseteq \{ s' \mid \mathcal{M}, s' \models \varphi \}$.

A *bdi-stit* formula φ is valid in a model $\mathcal{M} = (\mathcal{F}, v)$, $\mathcal{M} \models \varphi$, iff $\mathcal{M}, s \models \varphi$ for every situation s from the frame $\mathcal{F}$, and φ is valid on a frame $\mathcal{F}$, $\mathcal{F} \models \varphi$, iff φ is valid in every model based on $\mathcal{F}$. A set of *bdi-stit* formulas Δ is valid in a model $\mathcal{M}$, $\mathcal{M} \models \Delta$, (valid on a frame $\mathcal{F}$, $\mathcal{F} \models \Delta$) iff every element of Δ is valid in $\mathcal{M}$ (on $\mathcal{F}$). If $\Delta \cup \{\varphi\}$ is a set of *bdi-stit* formulas, then Δ entails φ, $\Delta \models \varphi$, iff for every model $\mathcal{M}$ it holds that $\mathcal{M} \models \varphi$, if $\mathcal{M} \models \Delta$. The formula φ is valid (simpliciter) iff $\emptyset \models \varphi$.

13.3 Tableaux for *bdi-stit* Logic

We want to present a tableau calculus in the style of Priest (2001) and to use it as a proof system such that it is possible to show soundness and completeness of *bdi-stit* logic in a very easy and transparent way.[4] Tableau calculi for normal modal logics have first been defined by Kripke (1963) and have been adapted to monotonic neighbourhood modalities in Allen (2005). By means of the tableau rules we can construct for a given formula φ a tableau, which has an appropriate tree structure. If a branch of the tableau for φ is open and complete, then φ has a model that can be defined from this branch.

Since we are working with branching-time structures, the tableaux have to provide information about the histories passing through each moment. Furthermore, they have to disclose in a suitable manner the available choices for agents to acquire, give up or revise certain beliefs, desires, and intentions but also the choices of agents to see to it that something is the case. Also to keep in mind is that the tableaux (or rather the frames of counter models defined from them) must satisfy the independence of agents condition.

13.3.1 Tableau Rules

If $\varDelta$ is a set of formulas, then $\varDelta^0$ is defined as a set of certain compound expressions, namely $\Delta^0 := \{\varphi, (m, h_0) \mid \varphi \in \Delta\}$. A tableau is a rooted tree. If $\varDelta$ is the set of premises of a derivation, and ψ its conclusion, then the root of the tableau for derivation $\Delta \vdash \psi$ is $\Delta^0 \cup \{m \in h_0, m \lhd m_0, m_0 \in h_0\} \cup \{\neg\psi, (m, h_0)\}$. To this root identity rules, decomposition rules, and structural tableau rules may (or may not) be applied to complete the tableau. A tableau is said to be *complete* iff each of its branches is complete. A branch is complete if there is no possibility to apply one more rule to expand this branch. A tableau branch is said to be *closed* iff there are expressions of the form φ, s and $\neg\varphi, s$ on the branch. A closed branch is considered complete. A tableau is called *closed* if and only if all of its branches are closed, and it is called *open*, if it is not closed.

We impose some conditions on rule applications. The indices $i, k, l, \ldots$ used in the tableau rules are natural numbers, and a *new* index is the smallest natural number not used in the tableau. If a rule application yields an expression which is already on the branch, this expression is not created again, unless the branch is split up by that rule and the expression occurs only in one subtree. Then both subtrees are created. Moreover, we shall not apply the rules `ref`, `REF`, or `SER` to introduce new agent variables which are not on the tableau. Note that considering the identities to which the identity rules may be applied requires interpreting agent variables by agents in a model. One way of doing this is to suitably extend the domain and the range of the assignment function v. In models constructed from tableaux, we shall just interpret an agent variable α by α itself: $v(\alpha) = \alpha$. Note also that it may happen that a rule is applied to an expression at a tableau node more than once if the rule requires additional input, because suitable additional input may be introduced at a later node.

[4] A sound and complete axiomatization of *bdi-stit* logic is presented in Semmling and Wansing (2009).

If, for instance, the □-rule is applied to the expressions $\Box\varphi, (m, h_i),\ m \in h_k$, and later on the branch a new expression $m \in h_l$ is introduced, then the □-rule has also to be applied to $\Box\varphi, (m, h_i),\ m \in h_l$.

The tableau calculus for *bdi-stit* logic consists of the tableau rules presented in the Tables 13.1–13.4. Syntactic consequence is then defined as follows.

Table 13.1 Identity rules

sub	ref	sym	tran
$\varphi, (m, h_i)$	$\cdot$	$\alpha = \beta$	$\alpha = \beta$
$\alpha = \beta$			$\beta = \gamma$
↓	↓	↓	↓
$\varphi(\alpha/\beta), (m, h_i)$	$\alpha = \alpha$	$\beta = \alpha$	$\alpha = \gamma$

Table 13.2 Structural tableau rules

REF	SYM	TRAN	IND	SER
(m, h_i)	$h_i \lhd^{\alpha}_{m} h_k$	$h_i \lhd^{\alpha}_{m} h_k$ $h_k \lhd^{\alpha}_{m} h_l$	$h_{l_1} \lhd^{\alpha_1}_{m} h_{l_1}$ $\ldots h_{l_k} \lhd^{\alpha_k}_{m} h_{l_k}$	s
↓	↓	↓	↓	↓
$h_i \lhd^{\alpha}_{m} h_i$	$h_k \lhd^{\alpha}_{m} h_i$	$h_i \lhd^{\alpha}_{m} h_l$	$m \lhd m_n$, n new $m \in h_n, m_n \in h_n,$ $h_{l_1} \lhd^{\alpha_1}_{m} h_n \ldots h_{l_k} \lhd^{\alpha_k}_{m} h_n$	$(m_l, h_l) \in I^{\alpha}_{s},$ $m_l \in h_l, I^{\alpha}_{s} \in N_{s'},$ $m \in h_l,\ m \lhd m_l,$ for some $m \in h,\ s'$ on the branch, l new

Table 13.3 Decomposition rules for deliberative-stit logic, cf. Wansing (2006a)

$\neg\neg\varphi, s$ ↓ φ, s	$(\varphi \wedge \psi), s$ ↓ $\varphi, s,\ \psi, s$	$\neg(\varphi \wedge \psi), s$ ↙ ↘ $\neg\varphi, s \qquad \neg\psi, s$
$\Box\varphi, (m, h_i),\ m \in h_k$ ↓ $\varphi, (m, h_k)$	$\neg\Box\varphi, (m, h_i)$ ↓ $\neg\varphi, (m, h_k),$ $m \in h_k,\ m_k \in h_k,$ $m \lhd m_k,\ k$ new	
$\alpha\ dstit\colon \varphi, (m, h_i),$ $h_i \lhd^{\alpha}_{m} h_k$ ↓ $\varphi, (m, h_k),\ m \lhd m_l,$ $m \in h_l,\ m_l \in h_l,$ $\neg\varphi, (m, h_l),\ l$ new	$\neg\alpha\ dstit\colon \varphi, (m, h_i),\ m \in h_l$ ↙ ↘ $\varphi, (m, h_l) \qquad \neg\varphi, (m, h_k)$ $m \in h_k,\ m_k \in h_k,$ $h_i \lhd^{\alpha}_{m} h_k,$ $m \lhd m_k,\ k$ new	

Table 13.4 Decomposition rules for the new *bdi-stit*-operators

$\Diamond\varphi, s$ $\downarrow$ $\varphi, (m_l, h_l),\ m_l \in h_l,$ $\{(m_l, h_l)\} \in N_s,\ m \lhd m_l,\ m \in h_l,$ for some $m \in h$ on the branch, l new	$\neg\Diamond\varphi, s,\ \{s'\} \in N_s$ $\downarrow$ $\neg\varphi,\ s'$
$\alpha\, bel\colon \varphi, s$ $\downarrow$ $\varphi, (m_l, h_l),\ m_l \in h_l, \{(m_l, h_l)\} \in B^{\alpha}_{s},$ $\{(m_l, h_l)\} \in N_{s'},\ m \lhd m_l,\ m \in h_l,$ for some s', $m \in h$ on the branch, l new	$\neg\alpha\, bel\colon \varphi, s,\ \{s'\} \in B^{\alpha}_{s}$ $\downarrow$ $\neg\varphi,\ s'$
$\alpha\, des\colon \varphi, s$ $\downarrow$ $\varphi, (m_l, h_l),\ m_l \in h_l, \{(m_l, h_l)\} \in D^{\alpha}_{s},$ $\{(m_l, h_l)\} \in N_{s'},\ m \lhd m_l,\ m \in h_l,$ for some s', $m \in h$ on the branch, l new	$\neg\alpha\, des\colon \varphi, s,\ \{s'\} \in D^{\alpha}_{s}$ $\downarrow$ $\neg\varphi,\ s'$
$\alpha\, int\colon \varphi, s,\ s' \in I^{\alpha}_{s}$ $\downarrow$ φ, s'	$\neg\alpha\, int\colon \varphi, s$ $\downarrow$ $\neg\varphi, (m_k, h_k),\ m_k \in h_k,$ $(m_k, h_k) \in I^{\alpha}_{s},\ m \lhd m_k,\ m \in h_k$ for some $m \in h$ on the branch, k new

Definition 3 Let $\Delta \cup \{\varphi\}$ be a set of *bdi-stit*-formulas. $\Delta \vdash \varphi$ (φ is derivable from Δ) iff there exists a closed and complete tableau for $\Delta^0 \cup \{m \in h_0, m \lhd m_0, m_0 \in h_0\} \cup \{\neg\varphi, (m, h_0)\}$.

If there is a complete and non-closed branch on a tableau with root $\Delta^0 \cup \{m \in h_0, m \lhd m_0, m_0 \in h_0\} \cup \{\neg\varphi, (m, h_0)\}$, then, as we shall see, there exists a counter model to the derivation $\Delta \vdash \varphi$. The construction of a counter model is possible from the branch, since each tableau node contains some information about how to build up this model. The expression $m \lhd m'$ gives the information that m and m' are moments and that m is before m'. The choice-equivalence of histories h_i, h_k for α at a moment m is represented by $h_i \lhd^{\alpha}_{m} h_k$. The expression $m \in h$ means that (m, h) is a situation. The term $U \in B^{\alpha}_{s}$ expresses that U is a set of situations which is compatible with what α believes at situation s. In a similar way one can interpret the expression $U \in D^{\alpha}_{s}$. Finally, $U \in N_s$ means that U is a neighbourhood of s.

13.3.2 Examples of Tableaux

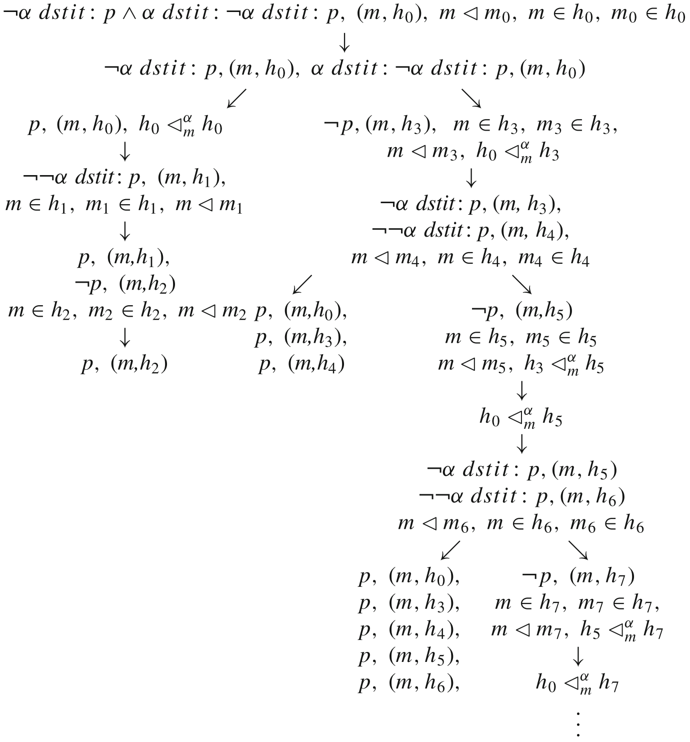
It can easily be seen that the rule SER leads to infinite tableaux in the tableau calculus for *bdi-stit* logic. However, also in the tableau calculus for deliberative-stit logic there are some rules with this effect. The rule TRAN already gives rise to infinite tableaux, cf. Example 3.4.7 in Priest (2001). The tableau in Table 13.5 can be extended ad infinitum.

It is well-known that if the independence of agents condition is imposed, no ‘other-agent-stit’-formula α *dstit*: β *dstit*: φ is satisfiable, where α and β are (interpreted by) distinct agents, see Wansing (2006b) and Belnap et al. (2001). Hence, in particular, any formula of the shape

$$\neg\alpha\ dstit\colon \beta\ dstit\colon \beta\ bel\colon \varphi$$

is a theorem of *bdi-stit* logic. In other words, if α and β are distinct agents which are *independent of each other*, then α cannot see to it that β expands her beliefs by φ. The following tableau for $\emptyset \vdash \neg\alpha\ dstit\colon \beta\ dstit\colon \beta\ bel\colon \varphi$ is closed.

Table 13.5 An infinite tableau


$\neg\alpha\ dstit\colon p \wedge \alpha\ dstit\colon \neg\alpha\ dstit\colon p,\ (m, h_0),\ m \lhd m_0,\ m \in h_0,\ m_0 \in h_0$
$\neg\alpha\ dstit\colon p, (m, h_0),\ \alpha\ dstit\colon \neg\alpha\ dstit\colon p, (m, h_0)$
$p,\ (m, h_0),\ h_0 \lhd^{\alpha}_{m} h_0$
$\neg\neg\alpha\ dstit\colon p,\ (m, h_1),$
$m \in h_1,\ m_1 \in h_1,\ m \lhd m_1$
$p,\ (m,h_1),$
$\neg p,\ (m,h_2)$
$m \in h_2,\ m_2 \in h_2,\ m \lhd m_2$
$p,\ (m,h_2)$
$\neg p, (m, h_3),\ m \in h_3,\ m_3 \in h_3,$
$m \lhd m_3,\ h_0 \lhd^{\alpha}_{m} h_3$
$\neg\alpha\ dstit\colon p, (m, h_3),$
$\neg\neg\alpha\ dstit\colon p, (m, h_4),$
$m \lhd m_4,\ m \in h_4,\ m_4 \in h_4$
$p,\ (m,h_0),$
$p,\ (m,h_3),$
$p,\ (m,h_4)$
$\neg p,\ (m,h_5)$
$m \in h_5,\ m_5 \in h_5$
$m \lhd m_5,\ h_3 \lhd^{\alpha}_{m} h_5$
$h_0 \lhd^{\alpha}_{m} h_5$
$\neg\alpha\ dstit\colon p, (m, h_5)$
$\neg\neg\alpha\ dstit\colon p, (m, h_6)$
$m \lhd m_6,\ m \in h_6,\ m_6 \in h_6$
$p,\ (m, h_0),$
$p,\ (m, h_3),$
$p,\ (m, h_4),$
$p,\ (m, h_5),$
$p,\ (m, h_6),$
$\neg p,\ (m, h_7)$
$m \in h_7,\ m_7 \in h_7,$
$m \lhd m_7,\ h_5 \lhd^{\alpha}_{m} h_7$
$h_0 \lhd^{\alpha}_{m} h_7$
$\vdots$

$$\neg\neg\alpha\ dstit\colon \beta\ dstit\colon \beta\ bel\colon p,\ (m,h_0),\ m \in h_0, m \lhd m_0$$
$$\downarrow$$
$$\alpha\ dstit\colon \beta\ dstit\colon \beta\ bel\colon p,\ (m,h_0)$$
$$\downarrow$$
$$h_0 \lhd_m^\alpha h_0$$
$$\downarrow$$
$$\beta\ dstit\colon \beta\ bel\colon p,\ (m,h_0),\ \neg\beta\ dstit\colon \beta\ bel\colon p,\ (m,h_1),$$
$$m \lhd m_1, m \in h_1, m_1 \in h_1$$
$$\downarrow$$
$$h_0 \lhd_m^\beta h_0$$
$$\downarrow$$
$$\beta\ bel\colon p,\ (m,h_0),\ \neg\beta\ bel\colon p,\ (m,h_2),$$
$$m \lhd m_2, m \in h_2, m_2 \in h_2,$$
$$\downarrow$$
$$h_2 \lhd_m^\beta h_2$$
$$\downarrow$$
$$h_0 \lhd_m^\alpha h_3,\ h_2 \lhd_m^\beta h_3$$
$$\downarrow$$
$$h_3 \lhd_m^\beta h_2$$
$$\downarrow$$
$$\beta\ dstit\colon \beta\ bel\colon p,\ (m,h_3)$$
$$\downarrow$$
$$\beta\ bel\colon p,\ (m,h_2),\ \neg\beta\ bel\colon p,\ (m,h_4)$$
$$m \lhd m_4, m \in h_4, m_4 \in h_4$$

Our next example shows that ascribing conflicting beliefs to an agent is consistent. Analogously, it may be shown that one can consistently ascribe conflicting desires. The following tableau can be further extended, but it will never become closed.

$$(\alpha\ bel\colon p \wedge \alpha\ bel\colon \neg p), (m,h_0)$$
$$\downarrow$$
$$\alpha\ bel\colon p,\ (m,h_0),\ \alpha\ bel\colon \neg p,\ (m,h_0)$$
$$\downarrow$$
$$p,\ (m_1,h_1),\ m_1 \in h_1,\ \{(m_1,h_1)\} \in B^\alpha_{(m,h_0)},$$
$$\{(m_1,h_1)\} \in N_{(m,h_0)},\ m \lhd m_1,\ m \in h_1$$
$$\neg p,\ (m_2,h_2),\ m_2 \in h_2,\ \{(m_2,h_2)\} \in B^\alpha_{(m,h_0)},$$
$$\{(m_2,h_2)\} \in N_{(m,h_0)},\ m \lhd m_2,\ m \in h_2$$

On the other hand, it is impossible to verify statements of conflicting intentions, which is obvious because no set I_s^α is empty. Thus, the following tableau is closed.

$$\begin{array}{c}
(\alpha\ int\colon p \wedge \alpha\ int\colon \neg p), (m, h_0) \\
\downarrow \\
\alpha\ int\colon p, (m, h_0),\ \alpha\ int\colon \neg p, (m, h_0) \\
\downarrow \\
(m_1, h_1) \in I_s^\alpha,\ m \lhd m_1, \\
m \in h_1,\ m_1 \in h_1,\ I_s^\alpha \in N_{(m_1,h_1)} \\
p, (m, h_1),\ \neg p, (m, h_1)
\end{array}$$

In Section 13.4 we shall consider some further examples.

13.3.3 *Soundness and Completeness of the Tableau Calculus for* bdi-stit *Logic*

We first show that for an arbitrary set $\Delta \cup \varphi$ of *bdi-stit* formulas it holds that $\Delta \vdash \varphi$ implies $\Delta \models \varphi$, i.e., our logic is sound. To this end we define what it means that a model $\mathcal{M}$ is faithful to a tableau branch b. Then we show that our tableau rules are such that a model $\mathcal{M}$ is still faithful to at least one branch b' obtained after applying one rule to a branch b, if $\mathcal{M}$ is faithful to b. In Wansing (2006a) the construction is already carried out for deliberative-stit logic, in particular for the rules of the historical necessity operator $\Box$ and the *dstit*-operator, which are defined as in Belnap and Perloff (1988), Belnap et al. (2001).

Definition 4 Let $\mathcal{M} = (\text{Tree}, \leq, \mathcal{A}, \bar{N}, \text{Choice}, \text{Bel}, \text{Des}, \text{Int}, v)$ be a model. Let History be the set of all histories of $\mathcal{M}$ and $\bar{S}$ be the set of all resultant situations, and b be a tableau branch. The model $\mathcal{M}$ is faithful to b iff there exists a function, $f : S \to \bar{S}$, where $S = \{\, (m_k, h_i) \mid m_k \in h_i \text{ occurs on } b \,\} \subseteq M \times H$ with $M = \bigcup\{\, m_k \mid m_k \text{ occurs on } b \,\}$, $H = \bigcup\{\, h_k \mid h_k \text{ occurs on } b \,\}$, such that the following conditions hold, where $f(U) = \{f(s) \mid s \in U\}$ if $U \subseteq S$:

1. For every expression φ, s on b, it holds that $\mathcal{M}, f(s) \models \varphi$.
2. If $f((m,h)) = f((m',h'))$, then for all $m'' \in M$, $h'' \in H$, if (m'',h), $(m'',h') \in S$, then $f((m'',h)) = f((m'',h'))$ and if (m,h''), $(m',h'') \in S$, then $f((m,h'')) = f((m',h''))$. Thus, it is possible to define two auxiliary functions related to f, $\pi_1 : M \to \text{Tree}$ and $\pi_2 : H \to \mathit{History}$ by requiring that for $m \in M$, $\pi_1(m) = \bar{m}$, if $f((m,\ldots)) = (\overline{m},\ldots)$, and for $h \in H$, $\pi_2(h) = \overline{h}$, if $f((\ldots,h)) = (\ldots,\overline{h})$.
3. If $m_i \in h_k$ occurs on b, then $\pi_2(h_k) \in H_{\pi_1(m_i)}$.
4. If $h_i \lhd_m^\alpha h_k$ occurs on b, then $\pi_2(h_k) \in \mathit{Choice}_{\pi_1(m)}^{v(\alpha)}(\pi_2(h_i))$.
5. If $U \in N_s$ occurs on b, then there is a $\bar{U} \in \bar{N}_{f(s)}$, such that $f(U) \subseteq \bar{U}$.
6. If $U \in B_s^\alpha$ occurs on b, then there is a $\bar{U} \in \mathit{Bel}_{f(s)}^{v(\alpha)}$, such that $f(U) \subseteq \bar{U}$.
7. If $U \in D_s^\alpha$ occurs on b, then there is a $\bar{U} \in \mathit{Des}_{f(s)}^{v(\alpha)}$, such that $f(U) \subseteq \bar{U}$.
8. If $s_k \in I_s^\alpha$ occurs on b, then $f(s_k) \in \mathit{Int}_{f(s)}^{v(\alpha)}$.
9. If s and α occur on b, then there exists an $(\bar{m}, \bar{h}) \in \mathit{Int}_{f(s)}^{v(\alpha)}$.

The function f is said to show that $\mathcal{M}$ is faithful to branch b. Now we show for every tableau rule, that a model which is faithful to a branch b is still faithful to at least one branch obtained from b by applying the rule.

Lemma 1 *Let $\mathcal{M} = (Tree, \leq, \mathcal{A}, \bar{N}, Choice, Bel, Des, Int, v)$ be a model, and b be a tableau branch. If $\mathcal{M}$ is faithful to b and a tableau rule is applied to b, then the application produces one extension b' of b, such that $\mathcal{M}$ is faithful to b'.*

Proof Assume that f is a function that shows $\mathcal{M}$ to be faithful to b. We have to consider each of the tableau rules. If the extended branch b' is obtained by applying an identity rule or one of the rules for $\neg\neg\varphi$, $(\varphi \wedge \psi)$ or $\neg(\varphi \wedge \psi)$, obviously f shows $\mathcal{M}$ to be faithful to b'.

Suppose the $\Box$-rule is applied to $\Box\varphi, (m, h_i)$. Then we obtain b' as extension of b by $\varphi, (m, h_k)$ for all $m \in h_k$ on b. Since f is faithful to b, we have $\mathcal{M}, f((m, h_i)) \models \Box\varphi$. By Definition 2 it holds that $\mathcal{M}, (\pi_1(m), \overline{h}) \models \varphi$ for all $\overline{h} \in H_{\pi_1(m)}$. Since $\pi_2(h_k) \in H_{\pi_1(m)}$, we have $\mathcal{M}, f((m, h_k)) \models \varphi$ for all h_k and f shows $\mathcal{M}$ to be faithful to b'.

Suppose now that the $\neg\Box$-rule is applied to $\neg\Box\varphi, (m, h_i)$, so that b is extended by $m \lhd m_k$, $m \in h_k$, $m_k \in h_k$, and $\neg\varphi, (m, h_k)$, for a new index k. Since f shows $\mathcal{M}$ to be faithful to b, we have $\mathcal{M}, f((m, h_i)) \models \neg\Box\varphi$. That means that there is $\overline{h} \in H_{\pi_1(m)}$ with $\mathcal{M}, (\pi_1(m), \overline{h}) \models \neg\varphi$. Define f' to be the same function as f and set for the new index k, $f'((m, h_k)) = (\pi_1(m), \overline{h})$ and $f'((m_k, h_k)) = (\overline{m}, \overline{h})$ for one $\overline{m} \in \overline{h}$ with $\pi_1(m) \leq \overline{m}$. The auxiliary functions π_1, π_2 of f' are appropriately expanded. Then $\mathcal{M}, f'((m, h_k)) \models \neg\varphi$, $\pi_2(h_k) \in H_{\pi_1(m)}$ and $\pi_2(h_k) \in H_{\pi_1(m_k)}$. The function f' shows $\mathcal{M}$ to be faithful to the extended branch b'.

Next, assume that the $\neg$*dstit*-rule is applied to $\neg\alpha\ dstit : \varphi, (m, h_i)$ to obtain a branch b' which is b extended by $\varphi, (m, h_l)$ for every expression $m \in h_l$ on b or a branch b'' as extension of b by $m \lhd m_k$, $m \in h_k$, $m_k \in h_k$, $h_i \lhd^{\alpha}_{m} h_k$ and $\neg\varphi, (m, h_k)$ for a new index k. Since f shows $\mathcal{M}$ to be faithful to b, we know that $\mathcal{M}, f((m, h_i)) \models \neg\alpha\ dstit : \varphi$. Hence either, case (i), there is $\overline{h} \in Choice^{v(\alpha)}_{\pi_1(m)}(\pi_2(h_i))$ with $\mathcal{M}, (\pi_1(m), \overline{h}) \models \neg\varphi$ or, case (ii), for all $h' \in H_{\pi_1(m)}$ it holds that $\mathcal{M}, (\pi_1(m), h') \models \varphi$.

Consider case (i). Let f' be the same function as f and set for the new index k, $f'((m, h_k)) = (\pi_1(m), \overline{h})$ and $f'((m_k, h_k)) = (\overline{m}, \overline{h})$ for one $\overline{m} \in \overline{h}$ with $\pi_1(m) \leq \overline{m}$. The auxiliary functions π_1, π_2 of f' are appropriately expanded. Then $\mathcal{M}, f'((m, h_k)) \models \neg\varphi$. Since $\pi_2(h_k) = \overline{h} \in Choice^{v(\alpha)}_{\pi_1(m)}(\pi_2(h_i)) \subseteq H_{\pi_1(m)}$, $\pi_2(h_k) \in H_{\pi_1(m)}$. The function f' shows $\mathcal{M}$ to be faithful to b''.

Consider case (ii). For every expression $m \in h_l$, $\pi_2(h_l) \in H_{\pi_1(m)}$. Therefore, $\mathcal{M}, f((m, h_l)) \models \varphi$ for all $m \in h_l$ on b. Therefore, f shows $\mathcal{M}$ to be faithful to b'.

Assume the *dstit*-decomposition rule is applied to $\alpha\ dstit : \varphi, (m, h_i)$. On the new branch b' there are $\varphi, (m, h_k)$ for all expression $h_i \lhd^{\alpha}_{m} h_k$ occurring on b as well as $m \lhd m_l$, $m \in h_l$, $m_l \in h_l$, and $\neg\varphi, (m, h_l)$, for a new index l. As f shows $\mathcal{M}$ to be faithful to b, we have $\mathcal{M}, f((m, h_i)) \models \alpha\ dstit : \varphi$. That means, for all $h' \in Choice^{v(\alpha)}_{\pi_1(m)}(\pi_2(h_i))$, $\mathcal{M}, (\pi_1(m), h') \models \varphi$ and there is $\overline{h} \in H_{\pi_1(m)}$ with $\mathcal{M}, (\pi_1(m), \overline{h}) \models \neg\varphi$. Let f' be the same function as f, and set for the new index l, $f'((m, h_l)) = (\pi_1(m), \overline{h})$ and

$f'((m_l, h_l)) = (\overline{m}, h')$ for one $\overline{m} \in h'$ with $\pi_1(m) \leq \overline{m}$. The auxiliary functions π_1, π_2 of f' are again appropriately expanded. Then $\mathcal{M}, f((m, h_l)) \models \neg\varphi$ and $\pi_2(h_l) \in H_{\pi_1(m)}$. Since, by assumption, $\pi_2(h_k) \in Choice^{v(\alpha)}_{\pi_1(m)}(\pi_2(h_i))$, also $\mathcal{M}, f'((m, h_k)) \models \varphi$, and thus f' shows $\mathcal{M}$ to be faithful to b'.

Suppose the $\Diamond$-rule is applied to an expression $\Diamond A,\ s$ on a branch b and the branch is extended to a branch b' by $\varphi, (m_l, h_l),\ m_l \in h_l,\ \{(m_l, h_l)\} \in N_s,\ m \lhd m_l,\ m \in h_l$, for a new index l and some $m \in h$ on b. Since f shows faithfulness, $\mathcal{M}, f(s) \models \Diamond\varphi$. By Definition 2, there is $\bar{U} \in \bar{N}_{f(s)}$ where $\bar{U} \neq \emptyset$ and for all $\bar{s} \in \bar{U}$, $\mathcal{M}, \bar{s} \models \varphi$. Choose f' as the same function as f for all s occurring on b and set $f'((m_l, h_l)) = \bar{s}$ for an arbitrary $\bar{s} = (\overline{m}, \overline{h}) \in \bar{U}$, such that $\pi_1(m_l) = \overline{m}$, $\pi_2(h_l) = \overline{h}$ and, thus, $\pi_2(h_l) \in H_{\pi_1(m_l)}$. Also by assumption, it holds that $\mathcal{M}, f'(s) \models \varphi$ for all expressions $\varphi,\ s$ which occur on b', and for all expressions $U \in N_s$, there is $\bar{U} \in \bar{N}_{f(s)}$, such that $f(U) \subseteq \bar{U}$, i.e., f' shows $\mathcal{M}$ to be faithful to b'.

If there are $\neg\Diamond\varphi,\ s$ on b, then the $\neg\Diamond$-rule effects that a branch b is extended to branch b' by $\neg\varphi,\ s'$ for all expressions $\{s'\} \in N_s$ on b. Since $\mathcal{M}, f(s) \models \neg\Diamond\varphi$, it holds for all $U \in \bar{N}_{f(s)}$ that there is $s_U \in U$ with $\mathcal{M}, s_U \models \neg\varphi$. By assumption, there is a $U_{s'} \in \bar{N}_{f(s)}$ with $f(\{s'\}) \subseteq U_{s'}$. Choose f' as the same function as f for all s occurring on b, except that $f'(s') = s_{U_{s'}}$ for all s' occurring in an expression $\{s'\} \in N_s$ on b. Then $\mathcal{M}, f'(s') \models \neg\varphi$ and f' shows the faithfulness of $\mathcal{M}$ to b'.

Assume that the *bel*-rule is applied to $\alpha\, bel : \varphi, s$. Then the branch b is extended to b' by $\varphi, (m_l, h_l), m_l \in h_l, \{(m_l, h_l)\} \in B^{\alpha}_s, m \lhd m_l$, and $m \in h_l$, for some new index l and $m \in h$ on b. $\mathcal{M}, f(s) \models \alpha\, bel : \varphi$, because f shows $\mathcal{M}$ to be faithful to b. That means that there exists $\bar{U} \in Bel^{v(\alpha)}_{f(s)}$ and for all $\bar{s} \in \bar{U}$ it holds that $\mathcal{M}, \bar{s} \models \varphi$. Set f' as f and choose one $\bar{s} \in \bar{U}$ with $f'((m_l, h_l)) = \bar{s}$, such that $\mathcal{M}, f'((m_l, h_l)) \models \varphi$. By assumption and since $f(\{(m_l, h_l)\}) = \{\bar{s}\} \subseteq \bar{U}$ it holds for all $U \in B^{\alpha}_s$ that $f'(U) \subseteq \bar{U}$ for $\bar{U} \in Bel^{v(\alpha)}_{f'(s)}$. Therefore, $\mathcal{M}$ is faithful to b'.

Let the $\neg$*bel*-rule be applied to $\neg\alpha\, bel : \varphi,\ s$. Then the branch b is extended to a branch b' by $\neg\varphi, s'$ for every expression $\{s'\} \in B^{\alpha}_s$. Since f shows $\mathcal{M}$ to be faithful to b, it follows that $\mathcal{M}, f(s) \models \neg\alpha\, bel : \varphi$, i.e., for all $U \in Bel^{v(\alpha)}_{f(s)}$ there is $s_U \in U$ with $\mathcal{M}, s_U \models \neg\varphi$. By assumption we know that there is $U_{s'} \in Bel^{v(\alpha)}_{f(s)}$ with $f(s') \in U_{s'}$. Choose f' as f, except that $f'(s') = s_{U_{s'}}$ for every $\{s'\} \in B^{\alpha}_s$ on b. Then $\mathcal{M}, f'(s') \models \neg\varphi$ and f' shows that $\mathcal{M}$ is faithful to b'.

The cases of the *des*-rule and the $\neg$*des*-rule can be dealt with analogously.

The semantics of the α *int*:-operator is relational. Therefore, the *int*-rule and the $\neg$*int*-rule are very similar to the $\Box$-rule and the $\neg\Box$-rule. Suppose that the *int*-rule is applied to $\alpha\, int : \varphi, s$. Then the branch b is extended to b' by φ, s' for all $s' \in I^{\alpha}_s$ on b. Since f shows $\mathcal{M}$ to be faithful to b, $\mathcal{M}, f(s) \models \alpha\, int : \varphi$. This means that for all $\bar{s} \in Int^{v(\alpha)}_{f(s)}$ it holds that $\mathcal{M}, \bar{s} \models \varphi$. Since $f(s') \in Int^{v(\alpha)}_{f(s)}$ for all $s' \in I^{\alpha}_s$, the function f shows $\mathcal{M}$ to be faithful to b'.

Suppose that the $\neg$*int*-rule is applied to $\neg\alpha\, int :\ \varphi, s$, so that b is extended by $m \lhd m_k$, $m \in h_k$, $m_k \in h_k$, and $\neg\varphi, (m_k, h_k)$, for some $m \in h$ on the branch and new index k. Since f shows $\mathcal{M}$ to be faithful to b, we have $\mathcal{M}, f(s) \models \neg\alpha\, int$: φ, which means that there is $(\overline{m}, \overline{h}) \in I^{\alpha}_{s}$ with $\mathcal{M}, (\overline{m}, \overline{h}) \models \neg\varphi$. Define f' to be the same function as f and set for the new index k, $f'((m_k, h_k)) = (\overline{m}, \overline{h})$ and $f'((m, h_k)) = (\pi_1(m), \overline{h})$. The auxiliary functions π_1, π_2 of f' are appropriately expanded. Then $\mathcal{M}, f'((m_k, h_k)) \models \neg\varphi$, $\pi_2(h_k) \in H_{\pi_1(m)}$, $\pi_2(h_k) \in H_{\pi_1(m_k)}$. The function f' shows $\mathcal{M}$ to be faithful to the extended branch b'.

It remains to consider the structural tableau rules. If f shows $\mathcal{M}$ to be faithful to b, and one of the first three structural rules is applied to obtain a branch b', then f shows $\mathcal{M}$ to be faithful to b' and Condition 4 is satisfied, because for every agent $\bar{\alpha} \in \mathcal{A}$ and every moment $m \in Tree$, the relation $\{(h, h') \mid h' \in Choice^{\bar{\alpha}}_{m}(h)\}$ is an equivalence relation.

For IND, we have $\pi_2(h_n) \in Choice^{v(\alpha_1)}_{\pi_1(m)}(\pi_2(h_{l_1})), \ldots, \pi_2(h_n) \in Choice^{v(\alpha_k)}_{\pi_1(m)}(\pi_2(h_{l_k}))$, because $\mathcal{M}$ satisfies the independence of agents condition. Moreover, since $\pi_2(h_n) \in Choice^{v(\alpha_1)}_{\pi_1(m)}(\pi_2(h_{l_1})) \subseteq H_{\pi_1(m)}$, it follows that $\pi_2(h_n) \in H_{\pi_1(m)}$. Thus, f itself shows $\mathcal{M}$ to be still faithful to a branch, if the IND-rule is applied to it.

Finally, if s occurs on b, and (m_l, h_l) is the situation newly introduced by applying SER, since f is faithful to b, there exists $(\bar{m}, \bar{h}) \in Int^{v(\alpha)}_{f(s)} \neq \emptyset$. Define f' like f and set $f(m_l, h_l) = (\bar{m}, \bar{h})$. ■

We want to use this lemma to show the soundness of *bdi-stit* logic by contraposition.

Theorem 1 *If* $\Delta \not\models \psi$*, then* $\Delta \nvdash \psi$*.*

Proof If $\Delta \not\models \psi$, then there is a model $\mathcal{M}$ and a situation $\overline{s} \in \mathcal{M}$, such that for all $\varphi \in \Delta$, $\mathcal{M}, \overline{s} \models \varphi$ but $\mathcal{M}, \overline{s} \not\models \psi$. We consider an arbitrary tableau for $\Delta^0 \cup \{\neg\psi, (m, h_0)\} \cup \{m \in h_0, m \lhd m_0, m_0 \in h_0\}$, such that every branch starts with the single-node branch b consisting of $\Delta^0 \cup \{\neg\psi, (m, h_0)\} \cup \{m \in h_0, m \lhd m_0, m_0 \in h_0\}$. Let $f((m, h_0)) = \overline{s} = f((m_0, h_0))$. Then f shows $\mathcal{M}$ to be faithful to b. According to the previous lemma, after applying a rule to branch b the model $\mathcal{M}$ is faithful to at least one branch which is an extension of b. So when we complete the tableau, there is still one complete branch b' which $\mathcal{M}$ is faithful to. If every branch of such a tableau is closed, then there are formulas $\chi, \neg\chi$, and a situation s' such that χ, s' and $\neg\chi, s'$ are on the branch b'. Since $\mathcal{M}$ is faithful to b', there is a function f', which maps the situations occurring on the branch into the set of moment/history-pairs of the model $\mathcal{M}$ and we have the contradiction $\mathcal{M}, f'(s') \models \chi$ and $\mathcal{M}, f'(s') \models \neg\chi$. Hence, it is not possible that a complete tableau for $\Delta^0 \cup \{\neg\psi, (m, h_0)\} \cup \{m \in h_0, m \lhd m_0, m_0 \in h_0\}$ exists, whose branches are all closed. By Definition 3, this means that ψ is not derivable from Δ, i.e., $\Delta \nvdash \psi$. ■

Now we want to show completeness: $\Delta \models \varphi$ implies $\Delta \vdash \varphi$. At first, we define for a given open branch of a complete tableau a model. Then we shall show that the model induced by the open branch satisfies the formulas occurring on this branch.

Definition 5 Let b be an open branch of a complete tableau. Then the model $\mathcal{M}_b = (Tree, \leq, \mathcal{A}, \bar{N}, Choice, Bel, Des, Int, v)$ induced by b is defined as follows:

1. Tree : $= \{ m \mid (m, h)$ occurs on $b \}$.
2. $\leq := cl$[5] $\left\{ (m_i, m_j) \mid m_i \lhd m_j \text{ occur on } b,\ m_i, m_j \in \text{Tree} \right\}$.
3. $\mathcal{A} := \{ \alpha \mid \alpha$ is an agent variable occurring on $b \}$.
4. $\text{Choice}^{\alpha}_{m}(h)$[6] $:= \left\{ h_l \mid h \lhd^{\alpha}_{m} h_l \text{ occurs on } b \right\}$
   for all $\alpha \in \mathcal{A}, m \in$ Tree, $m \in h$ occurring on b.
5. $\bar{N}(s) := \{ U \mid U \in N_s$ occurs on $b \}$ for all s occurring on b.
6. $\text{Bel}(\alpha, s) := \left\{ U \mid U \in B^{\alpha}_{s} \text{ occurs on } b \right\}$ for all s, α occurring on b.
7. $\text{Des}(\alpha, s) := \left\{ U \mid U \in D^{\alpha}_{s} \text{ occurs on } b \right\}$ for all s, α occurring on b.
8. $\text{Int}(\alpha, s) := \left\{ s' \mid s' \in I^{\alpha}_{s} \text{ occurs on } b \right\}$ for all s, α occurring on b.
9. (a) $v(\alpha) := \alpha$.
   (b) $v(p) := \{ s \mid p, s$ occurs on $b \}$.
   (c) $s \notin v(p)$, if $\neg p, s$ occurs on b.
   (d) The definition of v for any other atomic formulas p is arbitrary.[7]

Because of the transitive and reflexive closure and since every moment m_{k} introduced by a tableau rule is a $\lhd$-successor of the root moment of the first line in a tableau, the ordered set $(\mathit{Tree}, \leq)$ is a tree structure and so as well a branching time structure. According to the branching time structure we have induced sets of histories and situations. By the structural rules `REF`, `SYM`, and `TRAN`, and since the tableau is complete, it is obvious that $\lhd^{\alpha}_{m}$ is an equivalence relation defined on H_m, where H_m is the set of histories passing through moment m. Therefore, $\mathit{Choice}^{\alpha}_{m}(h)$ is the corresponding equivalence class of h. By rule `IND`, the independence of agents condition is assured, cf. [21]. For any element U of $\bar{N}(s)$ it is obvious that $U \neq \emptyset$. Similarly, no $U \in \mathit{Bel}(\alpha, s)$ and no $U \in \mathit{Des}(\alpha, s)$ is empty. By rule `SER` the sets $\mathit{Int}(\alpha, s)$ are not empty for arbitrary situation s and agent α.

Lemma 2 *Let b be an open branch of a complete tableau and let $\mathcal{M}_b = (\mathit{Tree}, \leq, \mathcal{A}, \bar{N}, \mathit{Choice}, \mathit{Bel}, \mathit{Des}, \mathit{Int}, v)$ be induced by b. Then it holds that*

$$\text{if } \varphi, s \text{ occurs on } b, \text{ then } \mathcal{M}_b, s \models \varphi.$$

Proof The proof is by induction not on the construction of a formula φ but on the number of connectives in φ. Suppose φ contains no connectives (i.e., φ is an atomic formula) and φ, s occurs on b. By definition of v, $s \in v(\varphi)$, so that $\mathcal{M}_b, s \models \varphi$.

Let $\varphi = \neg p$ be a negated atom and $\neg p, s$ occurs on b. Again by definition of v, $\mathcal{M}_b, s \not\models p$, and therefore $\mathcal{M}_b, s \models \varphi$.

If φ has the form $\neg\neg\psi$, $\psi \wedge \chi$, or $\neg(\psi \wedge \chi)$ just use the induction hypothesis and the completeness of the tableau.

Let $\varphi = \Box\psi$. If $\varphi, (m, h)$ occurs on b, then for every h_k with $m \in h_k$ occurring on b we have by the completeness of the tableau ψ, (m, h_k) on b and by the

[5] Here cl stands for the reflexive and transitive closure of a binary relation.

[6] Since we interpret α by α itself and since every situation over $(\mathit{Tree}, \leq)$ corresponds to a situation s on b, it is warrantable that we use the same letters in the tableaux and the notation of $\mathcal{M}_b$.

[7] Choosing v in this way is suitable, since b is open. There is no situation s on b, such that p, s and $\neg p, s$ occur on b.

induction hypothesis $\mathcal{M}_b, (m, h_k) \models \psi$. According to the definition of H_m it holds that $\mathcal{M}_b, (m, h) \models \Box\psi$.

Let $\varphi = \neg\Box\psi$. If $\varphi, (m, h)$ occurs on b, then, by completeness of the tableau, there is a situation (m, h_k) on b with $\neg\psi, (m, h_k)$ occurring on b, too. By induction hypothesis and definition of M_b we have $\mathcal{M}_b, (m, h_k) \models \neg\psi$, which entails $\mathcal{M}_b, (m, h) \models \neg\Box B$.

Let $\varphi = \alpha\, dstit\colon \psi$. If $\varphi, (m, h)$ occurs on b, then $\psi, (m, h_k)$ occurs on b for every h_k with $h \lhd_m^\alpha h_k$. Furthermore, there must be a history h_l with $m \in h_l$ and $\neg\psi, (m, h_l)$ occurring on b. By the induction hypothesis and the definition of $\mathcal{M}_b$ it follows that $\mathcal{M}_b, (m, h) \models \alpha\ dstit\colon \psi$.

Let $\varphi = \neg\alpha\, dstit\colon \psi$. If $\varphi, (m, h)$ occurs on b, then two cases are possible. Either there are some histories h_k with $h \lhd_m^\alpha h_k$ and $\neg\psi, (m, h_k)$ on b or for all h_l with $m \in h_l$ it holds that $\psi, (m, h_l)$ occurs on b. In the first case, it follows by definition of $\mathcal{M}_b$ that $h_k \in Choice_m^\alpha(h)$ and by hypothesis that $\mathcal{M}_b, (m, h) \models \neg\alpha\, dstit : \psi$. In the second case, the 'negative condition' is not satisfied and again $\mathcal{M}_b, (m, h) \models \neg\alpha\, dstit\colon \psi$.

Let $\varphi = \alpha\, int : \psi$ be. If φ, s occurs on b, then for all $s' \in I_s^\alpha$ it holds that ψ, s' occurs on b. By induction hypothesis for all such s' it holds that $\mathcal{M}_b, s' \models \psi$. Therefore, $\mathcal{M}_b, s \models \alpha\, int\colon \psi$

Let $\varphi = \neg\alpha\, int : \psi$. If φ, s occurs on b, there is an expression s_k on b with k new, such that $\neg\psi, s_k$ and $s_k \in I_s^\alpha$ occur on b. By induction hypothesis and definition of $\mathcal{M}_b$ there is thus $s_k \in Int_s^\alpha$ with $\mathcal{M}_b, s_k \models \neg\psi$, which means that $\mathcal{M}_b, s \models \neg\alpha\, int\colon \psi$.

Let $\varphi = \Diamond B$. If φ, s occurs on b, then there are expressions $\{s_l\} \in N_s$ and ψ, s_l on b with s_l new. By induction hypothesis we have $\mathcal{M}_b, s_l \models \psi$, and it follows that $\mathcal{M}_b, s \models \Diamond\psi$.

Let $\varphi = \neg\Diamond\psi$. If φ, s and $\{s'\} \in N_s$ occur on b, then $\neg\psi, s'$ occurs on b. By induction hypothesis we have $M_b, s' \models \neg\psi$ and since $N_s = \bar{N}_s$, it follows that $\mathcal{M}_b, s \models \neg\Diamond\psi$.

The cases that $\varphi = \alpha\, bel\colon \psi$, $\varphi = \neg\alpha\, bel\colon \psi$, $\varphi = \alpha\, des\colon \psi$, and $\varphi = \neg\alpha\, des\colon \psi$ are analogous to the corresponding cases $\varphi = \Diamond B$ and $\varphi = \neg\Diamond B$. ■

Theorem 2 *If* $\Delta \nvdash \psi$, *then* $\Delta \not\models \psi$.

Proof Let us assume that $\Delta \nvdash \psi$. According to Definition 3, this means that there is no complete and closed tableau for $\Delta^0 \cup \{\neg\psi, (m, h_0)\} \cup \{m \in h_0, m \lhd m_0, m_0 \in h_0\}$. Let b be an open branch of a complete tableau for this set and let $\mathcal{M}_b$ be the model induced by b. According to the previous Lemma it follows that $\mathcal{M}_b, (m, h_0) \models \varphi$ for every formula $\varphi \in \Delta$ and $\mathcal{M}_b, (m, h_0) \models \neg\psi$, thus $\mathcal{M}_b, (m, h_0) \not\models \psi$. Hence, $\Delta \not\models \psi$. ■

Corollary 1 *If one complete tableau for* $\Delta^0 \cup \{\neg\psi, (m, h_0)\} \cup \{m \in h_0, m \lhd m_0, m_0 \in h_0\}$ *corresponding to the derivation* $\Delta \vdash \psi$ *is open, then every complete tableau is open. If one complete tableau for* $\Delta^0 \cup \{\neg\psi, (m, h_0)\} \cup \{m \in h_0, m \lhd m_0, m_0 \in h_0\}$ *is closed, then so is every tableau for the corresponding derivation.*

So we have a sound and complete proof system. The validity of a derivation $\Delta \vdash \varphi$ has been shown, if we have constructed a complete and closed tableau for $\Delta \vdash \varphi$.

13.4 Translation of AGM Postulates

In the quotation from Gärdenfors (1988) in Section 13.1, Gärdenfors explains that he has chosen to keep the object language of the theory of belief changes as simple as possible. The language of the AGM theory, however, is not recursively defined as a formal language. The AGM postulates are stated in a language that contains an underlying language L in which the propositional content of belief states is expressed, schematic letters K, $K_1, K_2, \ldots$ for belief sets (which are subsets of L), for every $\varphi \in \mathsf{L}$ the function symbol $^*_\varphi$ (denoting the revision of a belief set by φ), the function symbol $^-_\varphi$ (denoting the contraction of a belief set by φ) and the function symbol $^+_\varphi$ (denoting the expansion of a belief set by φ), the provability predicate $\vdash$ for $\mathtt{L}$, the classical connectives[8] (expressed in English), the standard set theoretic vocabulary (the postulates display $\in$ and $\subseteq$), and the predicate "is a belief set". In a very simple setting, the language L may be assumed to be the language of classical propositional logic. We shall consider a translation of the AGM postulates for belief expansion, revision, and contraction into formulas from the language $\mathcal{L}$ of *bdi-stit* logic. Since the language of the AGM postulates is not recursively defined, its translation into $\mathcal{L}$-formulas is not recursively specified either and rests to some extent on plausibility considerations.

Gärdenfors (1988) defines a belief set in the following way:

> Sets of sentences that may be rationally held by an individual are called *belief sets*. In order to determine which sets of sentences constitute belief sets, I focus on two rationality criteria:
>
> 1. The set of accepted sentences should be consistent.
> 2. Logical consequences of what is accepted should also be accepted.

He also explains that it is convenient *for technical reasons* to regard the set of all sentences as a belief set, too. This set is called the *absurd belief set* and is denoted by $K_\perp$.

Agents in *bdi-stit* logic cannot have inconsistent beliefs and, moreover, their beliefs are closed under valid consequence in the sense that if $\alpha\, bel\colon \varphi$ and φ entails ψ, then $\alpha\, bel\colon \psi$. However, it is neither the case that $\{\alpha\, bel\colon \varphi, \alpha\, bel\colon \neg\varphi\}$ is unsatisfiable, nor that $\{\alpha\, bel\colon \varphi, \alpha\, bel\colon \psi\}$ entails $\alpha\, bel\colon (\varphi \wedge \psi)$. A motivation for this concept of logical closure of beliefs is presented in Semmling and Wansing (2008).

Therefore, the set of sentences believed by an agent at a situation s in general is not a belief set. But every neighbourhood $U \in Bel^\alpha_s$ of a situation s represents a belief set of agent α, since the set of formulas, which are satisfied on every situation of a neighbourhood, is a belief set. Moreover, the claim that a set K of $\mathcal{L}$-formulas is a belief set cannot be translated as an $\mathcal{L}$-formula.

[8] and the quantifiers, if we want to make explicit the implicit universal quantification over belief sets and formulas from $\mathtt{L}$.

13.4.1 Postulates for the Basic Changes of Belief Sets

As is well-known, there are three basic types of changes of belief in the AGM theory: expansion, contraction, and revision. Expanding a belief set by φ means that the agent changes her epistemic attitude to φ from indetermined to accepted, i.e., φ is just added to the belief set. Contracting a belief set by φ means that the agent gives up the belief that φ, i.e., the sentence φ is deleted from the belief set. If φ is removed from a belief set K to obtain a new belief set K', the requirement of deductive closure of K' may enforce the deletion of other formulas from K than just φ. The third type of belief change, revising a belief set by φ, amounts to first giving up the belief that $\neg\varphi$ and then adding the belief that φ. The belief $\neg\varphi$ thus has to be deleted from the belief set, so that another consistent and deductively closed belief set is obtained, to which φ is adjoined.

The AGM theory then puts forward a number of postulates that impose conditions on the rational expansion, revision, and contraction of belief sets. In the following we intend to express (as many as possible of) these postulates for the three kinds of changes of beliefs in the language $\mathcal{L}$ of *bdi-stit* logic. As mentioned in the introduction, the idea is to express that agent α expands her belief by φ as $\alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi$ and that α contracts her belief by φ as $\alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{bel}\colon \varphi$. Agent α's revision of her beliefs by φ is expressed as $\alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{bel}\colon \neg\varphi \wedge \alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi$. Let $\alpha\, \mathit{contra}\colon \varphi$ abbreviate $\alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{bel}\colon \varphi$, and let $\alpha\, \mathit{rev}\colon \varphi$ abbreviate $\alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{bel}\colon \neg\varphi \wedge \alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi$.

13.4.1.1 Postulates for the Expansion of a Belief Set

Let $\mathcal{K}$ be the set of all belief sets including the absurd belief set. Then an operation $+ : \mathcal{K} \times \mathsf{L} \to \mathcal{K}$ is said to be an expansion operation on $\mathcal{K}$ iff it fulfills all postulates (K$^+$1)–(K$^+$6) listed in Table 13.6. For $K \in \mathcal{K}$ and $\varphi \in \mathsf{L}$, the set $+(K, \varphi)$ is usually denoted as K^+_φ.

Postulate (K$^+$1) is evidently satisfied with respect to the codomain of operator +, and (K$^+$6) is a higher-level postulate referring to the other postulates. It is used to obtain an explicit representation of expansion functions. Indeed, a function $+ : \mathcal{K} \times \mathsf{L} \to \mathcal{K}$ satisfies the six postulates iff $+(K, \varphi)$ is the deductive closure of $K \cup \varphi$. In view of (K$^+$3), we may replace (K$^+$4) by (K$^+$4'): If $\varphi \in K$ then $K^+_\varphi \subseteq K$.

Translations of expansion postulates into formulas of *bdi-stit* logic are presented in Table 13.7. The translation (tK$^+$2) is obvious: α believes that φ, if α expands

Table 13.6 The AGM postulates for belief expansion

(K^+1)	K^+_φ is a belief set.
(K^+2)	$\varphi \in K^+_\varphi$.
(K^+3)	$K \subseteq K^+_\varphi$.
(K^+4)	If $\varphi \in K$ then $K^+_\varphi = K$.
(K^+5)	If $K \subseteq H$ then $K^+_\varphi \subseteq H^+_\varphi$.
(K^+6)	K^+_φ is the smallest belief set that satisfies $(K^+1) - (K^+5)$.

Table 13.7 Translation of (K$^+$) postulates into *bdi-stit* logic

(tK^+2)	$\alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi \supset \alpha\, \mathit{bel}\colon \varphi$
(tK^+3)	$\alpha\, \mathit{bel}\colon \psi \supset (\alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi \supset \neg\alpha\, \mathit{contra}\colon \psi)$
(tK^+4’)	$\alpha\, \mathit{bel}\colon (\varphi \wedge \psi) \supset (\alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi \supset \alpha\, \mathit{bel}\colon \psi)$
(tK^+5)	$(\alpha\, \mathit{bel}\colon \chi \supset \alpha\, \mathit{bel}\colon \psi) \supset$ $((\alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi \wedge \alpha\, \mathit{bel}\colon (\chi \wedge \varphi)) \supset (\alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi \wedge \alpha\, \mathit{bel}\colon (\psi \wedge \varphi)))$

her belief by φ. (tK$^+$3) is unproblematic, too: If α believes that ψ and expands her beliefs by φ, then it is not the case that α contracts her beliefs by (gives up the belief that) ψ. (tK$^+$4’) is a plausible translation as well. If α believes that $(\varphi \wedge \psi)$, then she still believes that ψ if she expands her beliefs by φ.[9] Also translation (tK$^+$5) appears to be plausible and clearly reflects the syntactic structure of the postulate it translates.

(tK$^+$2) – (tK$^+$5) are valid by the fact that $\alpha\, \mathit{dstit}\colon$ is veridical (truth-implying) and that we have the closure with respect to beliefs which is expressed by the valid schema $\alpha\, \mathit{bel}\colon (\psi \wedge \varphi) \supset \alpha\, \mathit{bel}\colon \psi$.[10]

13.4.1.2 Postulates for the Revision of a Belief Set

We now consider ascriptions of belief revision. The AGM postulates for belief revision are listed in Table 13.8. Since we consider postulates for an operator $*\colon\ \mathcal{K} \times \mathcal{L} \rightarrow \mathcal{K}$, we may again ignore the first postulate. One direction of the fifth postulate means that if an agent revises by an inconsistent formula φ, she believes everything, which is absurd (for rational agents). In our *bdi-stit* logic a formula $\alpha\, \mathit{dstit}\colon\alpha\, \mathit{bel}\colon \varphi$ is not satisfiable for inconsistent formulas φ, so we avoid this absurdity.

The other direction means that if an agent believes everything as a result of revising her beliefs by φ, then φ is inconsistent. Since in *bdi-stit* logic a belief set consists of formulas satisfied in all situations of a nonempty neighbourhood, it has to be consistent. So an agent in *bdi-stit* logic cannot believe everything, as it is unsatisfiable to believe an inconsistency. However, the fifth postulate does not render moot conflicting beliefs ascribed by formulas such as $\alpha\, \mathit{bel}\colon \varphi \wedge \alpha\, \mathit{bel}\colon \neg\varphi$. It rather expresses that it is only absurd to have an inconsistent belief expressed by, for example, $\alpha\, \mathit{bel}\colon (\varphi \wedge \neg\varphi)$.

In the AGM theory, belief revision is definable from belief expansion and belief contraction by the so-called Levi Identity: $K^*_\varphi = (K^-_{\neg\varphi})^+_\varphi$. Thus, K is *first* contracted by $\neg\varphi$ and *afterwards* expanded by φ. It is not clear, why revision by φ should not be a one-step process consisting of a simultaneous contraction by $\neg\varphi$ and expansion by φ. The suggested translations of the postulates (K*2) – (K*4) and (K*6) – (K*8) are stated in Table 13.9.

[9] Note that we do not translate (K$^+$4’) as $(\alpha\, \mathit{bel}\colon \varphi \wedge \alpha\, \mathit{bel}\colon \psi) \supset (\alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi \supset \alpha\, \mathit{bel}\colon \psi)$, because $(\alpha\, \mathit{bel}\colon \varphi \wedge \alpha\, \mathit{bel}\colon \psi) \supset \alpha\, \mathit{bel}\colon (\varphi \wedge \psi)$ fails to be valid, hence neighbourhoods $U \in B^\alpha_s$ represent belief sets.

[10] For a discussion of closure principles for belief, see, for instance, Fagin and Halpern (1988) and Semmling and Wansing (2008).

Table 13.8 The AGM postulates for belief revision

(K*1)	K^*_φ is a belief set.
(K^*2)	$\varphi \in K^*_\varphi$.
(K^*3)	$K^*_\varphi \subseteq K^+_\varphi$.
(K^*4)	if $\neg\varphi \notin K$ then $K^+_\varphi \subseteq K^*_\varphi$.
(K^*5)	$K^*_\varphi = K_\perp$ iff $\vdash \neg\varphi$.
(K^*6)	if $\vdash \varphi \equiv \psi$ then $K^*_\varphi = K^*_\psi$.
(K^*7)	$K^*_{\varphi\wedge\psi} \subseteq (K^*_\varphi)^+_\psi$.
(K^*8)	if $\neg\psi \notin K^*_\varphi$ then $(K^*_\varphi)^+_\psi \subseteq K^*_{\varphi\wedge\psi}$.

Table 13.9 Translation of (K*) postulates into *bdi-stit* logic

(tK^*2)	$\alpha\ rev\colon\varphi \supset \alpha\ bel\colon\varphi$
(tK^*3)	$\alpha\ rev\colon\varphi \supset \alpha\ dstit\colon\alpha\ bel\colon\varphi$
(tK^*4)	$\neg\alpha\ bel\colon\neg\varphi \supset (\alpha\ dstit\colon\alpha\ bel\colon\varphi \supset \alpha\ rev\colon\varphi)$
(tK^*6)	If $\vdash \varphi \equiv \psi$ then $\vdash \alpha\ rev\colon\varphi \equiv \alpha\ rev\colon\psi$.
(tK^*7)	$\alpha\ rev\colon(\varphi \wedge \psi) \supset (\alpha\ rev\colon\varphi \wedge \alpha\ dstit\colon\alpha\ bel\colon\psi)$
(tK^*8)	$\alpha\ rev\colon(\varphi \wedge \neg\psi) \supset (\alpha\ dstit\colon\alpha\ bel\colon\psi \supset \alpha\ rev\colon(\varphi \wedge \psi))$

The formulas (tK*2) and (tK*3) are evidently valid, and also (tK*6) holds. Moreover, (tK*4) is valid, since $\alpha\ rev\colon\varphi \supset \neg\alpha\ bel\colon\neg\varphi$ is valid. (tK*7) is also valid, since $\neg\alpha\ bel\colon\neg(\varphi \wedge \psi) \supset \neg\alpha\ bel\colon\neg\varphi$ holds, if it is possible at s that agent α believes the negation of φ. To show that (tK*8) is valid, one can use the tableau calculus of *bdi-stit* logic, cf. Appendix A.

13.4.1.3 Postulates for the Contraction of a Belief Set

We finally turn to ascriptions of belief contraction. The AGM postulates for belief contraction are listed in Table 13.10. Again we may ignore the first postulate, and, obviously, the fourth and the sixth postulate cannot be translated just as $\mathcal{L}$-formulas, but they are translatable into correct proof rules. In view of (K$^-$2) we may replace (K$^-$3) by (K$^-$3'): If $\varphi \notin K$ then $K \subseteq K^-_\varphi$. However, in the case of contraction the idea of a translation into *bdi-stit* logic runs into serious difficulties, and we shall not try to satisfactorily tackle these problems here. The translation of (K$^-$5) poses a problem, because an agent cannot consistently be ascribed to contract her beliefs by φ *and* (at the same time) expand her beliefs by φ. Nevertheless, plausible translations of some postulates for belief contraction are stated in Table 13.11.

The formula (tK$^-$2) is obviously valid and may be understood to express that if formulas φ and ψ are true in every situation in (the representation of) a belief set in *bdi-stit* logic, then a contraction by φ does not defeat the belief that ψ.[11] The translation (tK$^-$3') says that if an agent does not believe that φ, the agent sees to it that she does not believe that $\neg\varphi$. So the formula (tK$^-$3') is not valid, and it seems that this formula cannot be validated by a *purely structural* condition on models, a

[11] An alternative translation would be $\neg\alpha\ bel\colon\psi \supset (\alpha\ contra\colon\varphi \supset \neg\alpha\ dstit\colon\alpha\ bel\colon\psi)$. ("If an agent contracts her beliefs, she does not add new beliefs".) This formula is also valid.

Table 13.10 The AGM postulates for belief contraction

$(K^{-}1)$	K^{-}_{φ} is a belief set.
$(K^{-}2)$	$K^{-}_{\varphi} \subseteq K$.
$(K^{-}3)$	If $\varphi \notin K$ then $K = K^{-}_{\varphi}$.
$(K^{-}4)$	If $\nvdash \varphi$ then $\varphi \notin K^{-}_{\varphi}$.
$(K^{-}5)$	If $\varphi \in K$ then $K \subseteq (K^{-}_{\varphi})^{+}_{\varphi}$.
$(K^{-}6)$	If $\vdash \varphi \equiv \psi$ then $K^{-}_{\varphi} = K^{-}_{\psi}$.
$(K^{-}7)$	$K^{-}_{\varphi} \cap K^{-}_{\psi} \subseteq K^{-}_{\varphi\wedge\psi}$.
$(K^{-}8)$	If $\varphi \notin K^{-}_{\varphi\wedge\psi}$ then $K^{-}_{\varphi\wedge\psi} \subseteq K^{-}_{\varphi}$.

Table 13.11 Translation of (K^{-}) postulates into *bdi-stit* logic

($\mathrm{tK}^{-}2$)	$\alpha\ bel\colon (\varphi \wedge \psi) \supset (\alpha\ contra\colon \varphi \supset \alpha\ bel\colon \psi)$
($\mathrm{tK}^{-}3$’)	$(\neg\alpha\ bel\colon \varphi \wedge \alpha\ bel\colon \psi) \supset (\alpha\ contra\colon \varphi \supset \alpha\ bel\colon \psi)$
($\mathrm{tK}^{-}4$)	If $\nvdash \varphi$ then $\vdash \alpha\ contra\colon \varphi \supset \neg\alpha\ bel\colon \varphi$
($\mathrm{tK}^{-}6$)	If $\vdash \varphi \equiv \psi$ then $\vdash \alpha\ contra\colon \varphi \equiv \alpha\ contra\colon \psi$
($\mathrm{tK}^{-}7$)	$(\alpha\ contra\colon \varphi \wedge \alpha\ contra\colon \psi) \supset \alpha\ contra\colon (\varphi \wedge \psi)$
($\mathrm{tK}^{-}8$)	$(\alpha\ contra\colon (\varphi \wedge \psi) \supset \neg\alpha\ bel\colon \varphi) \supset (\alpha\ contra\colon (\varphi \wedge \psi) \supset \alpha\ contra\colon \varphi)$

condition which does not refer to valuations. Clearly, the implications ($\mathrm{tK}^{-}4$) and ($\mathrm{tK}^{-}6$), however, hold. The translation ($\mathrm{tK}^{-}7$) means that if an agent α contracts her beliefs by φ and contracts her beliefs by ψ, then she contracts her beliefs also by the conjunction of φ and ψ. This formula is true at a situation s, if it is not a necessary truth at s that agent α does not believe the conjunction of φ and ψ.[12] The translation ($\mathrm{tK}^{-}8$) means that if an agent contracts her beliefs by a conjunction and she does not believe one of the conjuncts, then she contracts her beliefs by this conjunct. The formula is obviously not valid, since not believing does not entail that somebody sees to it that he does not believe and, on the other hand, contracting beliefs by a conjunction does not imply contracting by one of the conjuncts.

13.4.2 More Examples of Tableaux

We conclude this section by two more examples of tableaux as simple examples for reasoning about belief revision. We start with the following closed tableau associated with $\emptyset \vdash \alpha\ rev\colon \neg\varphi \supset \alpha\ contra\colon \neg\neg\varphi$:

$$\neg(\alpha\ rev\colon \neg\varphi \supset \alpha\ contra\colon \neg\neg\varphi),\ (m, h_0), m \lhd m_0, m \in h_0, m_0 \in h_0$$
$$\downarrow$$
$$\alpha\ dstit\colon \neg\alpha\ bel\colon \neg\neg\varphi \wedge \alpha\ dstit\colon \alpha\ bel\colon \neg\varphi,\ (m, h_0)$$
$$\neg\alpha\ dstit\colon \neg\alpha\ bel\colon \neg\neg\varphi,\ (m, h_0)$$
$$\downarrow$$
$$\alpha\ dstit\colon \neg\alpha\ bel\colon \neg\neg\varphi, (m, h_0),\ \alpha\ dstit\colon \alpha\ bel\colon \neg\varphi,\ (m, h_0)$$

[12] This condition of a not necessary truth refers to the negative condition in the semantics of the *dstit*-operator.

Table 13.12 $\emptyset \vdash \alpha\ rev : \neg(\varphi \vee \psi) \supset \alpha\ bel : \neg\psi$

$$\begin{gathered}
\neg(\alpha\ rev : \neg(\varphi \vee \psi) \supset \alpha\ bel : \neg\psi),\ (m, h_0), m \lhd m_0, m \in h_0, m_0 \in h_0 \\
\downarrow \\
\alpha\ dstit : \alpha\ bel : \neg(\varphi \vee \psi),\ (m, h_0), \\
\alpha\ dstit : \neg\alpha\ bel : \neg\neg(\varphi \vee \psi),\ (m, h_0), \\
\neg\alpha\ bel : \neg\psi,\ (m, h_0), \\
\downarrow \\
h_0 \lhd^{\alpha}_{m} h_0 \\
\downarrow \\
\alpha\ bel : \neg(\varphi \vee \psi),\ (m, h_0), \\
\neg\alpha\ bel : \neg(\varphi \vee \psi),\ (m, h_1), \\
\neg\alpha\ bel : \neg\neg(\varphi \vee \psi),\ (m, h_0), \\
\neg\neg\alpha\ bel : \neg\neg(\varphi \vee \psi),\ (m, h_2), \\
m \lhd m_1,\ m \in h_1,\ m_1 \in h_1 \\
m \lhd m_2,\ m \in h_2,\ m_2 \in h_2 \\
\downarrow \\
\neg(\varphi \vee \psi),\ s_3 = (m_3, h_3), \{s_3\} \in B^{\alpha}_{(m,h_0)}, \\
m \lhd m_3,\ m \in h_3,\ m_3 \in h_3,\ \{s_3\} \in N_{(m,h_0)}, \\
\downarrow \\
\neg\neg\neg(\varphi \vee \psi),\ (m_3, h_3), \\
\neg\neg\psi,\ (m_3, h_3), \\
\downarrow \\
\neg(\varphi \vee \psi),\ (m_3, h_3), \\
\psi,\ (m_3, h_3), \\
\downarrow \\
\neg\varphi,\ (m_3, h_3), \\
\neg\psi,\ (m_3, h_3),
\end{gathered}$$

Another example is presented in Table 13.12. It establishes that the derivation $\emptyset \vdash \alpha\ rev : \neg(\varphi \vee \psi) \supset \alpha\ bel : \neg\psi$ is valid. If someone revises her beliefs by the negation of a disjunction, then she believes the negation of the disjuncts.

13.5 Summary and Outlook

The contribution of this chapter is twofold. Firstly, we have defined a tableau calculus for *bdi-stit* logic, an extension of both *dstit* logic (the logic of deliberatively seeing-to-it-that) and a fragment of *BDI* logic (the logic of beliefs, desires, and intentions), see Semmling and Wansing (2008); Belnap et al. (2001); Georgeff and Rao (1998); Wooldridge (2000). Since the language of *bdi-stit* logic can be used to express ascriptions of belief expansion, contraction, and revision, the tableau calculus for *bdi-stit* logic may serve as a proof system for reasoning about such belief changes. Secondly, we have suggested a translation of several of the AGM postulates for rational belief change into the language of *bdi-stit* logic. It turned out that some of the translated postulates are valid, that some are not, and that other postulates do not admit of a translation into the language of *bdi-stit* logic.

Several directions for further research suggest themselves. We here only mention the addition of temporal operators. In Xu's axiomatization (and definition) of *dstit* logic temporal operators have been excluded from consideration, but the authors of Belnap et al. (2001) emphasize that the non-deterministic branching-time framework obviously invites the use of temporal operators. Also the project of translating postulates (and theorems) of the AGM theory of belief revision into the language of a doxastic logic of agency may benefit from extending this latter language by temporal modalities. Expansion, contraction, and revision functions are applied in some linear order to a belief set, and the temporal succession of such applications cannot be expressed in the language of *bdi-stit* logic. It therefore seems natural to introduce, for example, a next-time operator O for histories, which are now required to be discretely ordered. We would then postulate that $\mathcal{M}, (m,h) \models \mathsf{O}\,\varphi$ iff $\mathcal{M}, (m',h) \models \varphi$, where m' is the successor of m on h. That agent α expands her beliefs by φ *immediately after* first contracting her beliefs by ψ, for example, can then be expressed by $\alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{bel}\colon \psi \wedge \mathsf{O}\; \alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi$.

Appendix: A Tableau Proof of (tK*8)

Before applying the tableau calculus, we formulate the negation of (tK*8):

$$\begin{aligned} &\neg(\alpha\, \mathit{rev}\colon (\varphi \wedge \neg\psi) \supset (\alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \psi \supset \alpha\, \mathit{rev}\colon (\varphi \wedge \psi))) \\ \equiv\;\; &\alpha\, \mathit{rev}\colon (\varphi \wedge \neg\psi) \wedge (\alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \psi \wedge \neg\alpha\, \mathit{rev}\colon (\varphi \wedge \psi)) \end{aligned}$$

Note that $\alpha\, \mathit{rev}\colon \varphi \equiv (\alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{bel}\colon \neg\varphi \wedge \alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi)$ and that $\neg\alpha\, \mathit{rev}\colon \varphi \equiv (\neg\alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{bel}\colon \neg\varphi \vee \neg\alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \varphi)$. Now, we show that it is not possible to

Table 13.13 Another tableau proof

$$\begin{gathered}
\alpha\, \mathit{rev}\colon (\varphi \wedge \neg\psi) \wedge (\alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \psi \wedge \neg\alpha\, \mathit{rev}\colon (\varphi \wedge \psi)),\ (m,h_0), \\
m \lhd m_0,\ m \in h_0,\ m_0 \in h_0 \\
\downarrow \\
\alpha\, \mathit{rev}\colon (\varphi \wedge \neg\psi),\ (m,h_0),\ \alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon \psi,\ (m,h_0),\ \neg\alpha\, \mathit{rev}\colon (\varphi \wedge \psi),\ (m,h_0), \\
\downarrow \\
h_0 \lhd^{\alpha}_{m} h_0 \\
\alpha\, \mathit{dstit}\colon \alpha\, \mathit{bel}\colon (\varphi \wedge \neg\psi),\ (m,h_0),\ \alpha\, \mathit{dstit}\colon \neg\alpha\, \mathit{bel}\colon \neg(\varphi \wedge \neg\psi),\ (m,h_0), \\
\downarrow \\
\alpha\, \mathit{bel}\colon \psi, (m,h_0), \neg\alpha\, \mathit{bel}\colon \psi, (m,h_1), m \lhd m_1, m \in h_1, m_1 \in h_1, \\
\alpha\, \mathit{bel}\colon (\varphi \wedge \neg\psi), (m,h_0), \neg\alpha\, \mathit{bel}\colon (\varphi \wedge \neg\psi), (m,h_2), m \lhd m_2,\ m \in h_2, m_2 \in h_2, \\
\neg\alpha\, \mathit{bel}\colon \neg(\varphi \wedge \neg\psi), (m,h_0), \neg\neg\alpha\, \mathit{bel}\colon \neg(\varphi \wedge \neg\psi), (m,h_3), m \lhd m_3, m \in h_3, m_3 \in h_3, \\
\downarrow \\
\psi,\ s_4 = (m_4,h_4),\ \{s_4\} \in \mathit{Bel}^{\alpha}_{(m,h_0)},\ m \in h_4,\ \{s_4\} \in N_{(m,h_1)}, \\
\downarrow \\
\neg\neg(\varphi \wedge \neg\psi),\ s_4 \\
\downarrow \\
\varphi \wedge \neg\psi,\ s_4 \\
\downarrow \\
\varphi,\ s_4,\ \neg\psi,\ s_4
\end{gathered}$$

construe a counter model for the negation of (tK*8) by demonstrating that the finite tableau in Table 13.13 is closed.[13]

References

Allen, M. 2005. Complexity results for logics of local reasoning and inconsistent belief. In *Theoretical aspects of rationality and knowledge*, ed. Ron van der Meyden, 92–108. *Proceedings of the 10th conference*, National University of Singapore, Singapore.

Belnap, N.D., and M. Perloff. 1988 Seeing to it that: a canonical form for agentives. *Theoria* 54: 175–199.

Belnap, N.D., M. Perloff, and M. Xu. 2001. *Facing the future: Agents and choices in our indeterminist world*, New York, NY: Oxford University Press.

van Benthem, J. 1995. Logic and the flow of information. In *Logic, methodology and philosophy of science IX*, eds. D. Prawitz et al., 693–724. Amsterdam: North Holland.

Bonanno, G. 2005. A simple modal logic for belief revision. *Synthese* 147:5–40.

Bonanno, G. 2007. Temporal interaction of information and belief. *Studia Logica* 86:375–401.

van Ditmarsch, H., W. van der Hoek, and B. Kooi. 2005. Playing cards with Hintikka. An introduction to dynamic epistemic logic. *Australasian Journal of Logic* 3:108–134.

Fagin, R., and J.Y. Halpern. 1988 Belief, awareness and limited reasoning. *Artificial Intelligence* 34:39–76.

Gärdenfors, P. 1988. *Knowledge in flux. Modeling the dynamics of epistemic states.* Cambridge: MIT Press.

Georgeff, M.P., and A.S. Rao. 1998. Decision procedures for BDI logics. *Journal of Logic and Computation* 8:293–342.

Hansson, S.O. 1999. *A textbook of belief dynamics: theory change and database updating*. Dordrecht: Kluwer.

Kripke, S.A. 1963. Semantical analysis of modal logic I: Normal propositional calculi. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik* 9:6796.

Leitgeb, H., and K. Segerberg. (2007). Dynamic doxastic logic: Why, how, and where to? *Synthese* 155:167–190.

Nottelmann, N. 2007. *Blameworthy belief: A study in epistemic deontologism*. Berlin: Springer.

Priest, G. 2001. *An introduction to non-classical logic*. Cambridge: Cambridge University Press.

de Rijke, M. 1994. Meeting some neighbours. In *Logic and information flow* eds. J. van Eijk, and A. Visser. Cambridge, MA: MIT.

Rott, H. 2001. *Change, choice and inference. A study of belief revision and non-monotonic reasoning*. Oxford: Oxford University Press.

Segerberg, K. 1999. Two traditions in the logic of belief: Bringing them together. In *Logic, language and reasoning. Essays in honour of Dov Gabbay*, eds. H.-J. Ohlbach, and U. Reyle, 135–147. Dordrecht: Kluwer.

Semmling, C., and H. Wansing. 2008. From *BDI* and *stit* to *bdi-stit* logic. *Logic and Logical Philosophy* 17:185–207.

Semmling, C., and H. Wansing. 2009. A sound and complete axiomatic system of *bdi-stit* logic, In *Logica Yearbook 2008*, ed. M. Pelis, 193–210. London: College Publications.

Wansing, H. 2006a. Tableaux for multi-agent deliberative-stit logic. In *Advances in modal logic, vol. 6*, eds. G. Governatori, I. Hodkinson, and Y. Venema, 503–520. London: College Publications.

Wansing, H. 2006b. Doxastic decisions, epistemic justification, and the logic of agency. *Philosophical Studies* 128:201–227.

[13] In the following tableau we represent by an arrow the application of exactly one tableau rule, but this rule can be applied to more than one subformula of a given formula.

Williams, B. 1973. Deciding to believe. In *Problems of the self*, chapter 9, 136–151. New York, NY: Cambridge University Press.

Winters, B. 1979. Believing at will. *Journal of Philosophy* 76:243–256.

Wooldridge, M. 2000. *Reasoning about rational agents*. Cambridge, MA: MIT Press.

Chapter 14
Changing Minds About Climate Change: Belief Revision, Coherence, and Emotion

Paul Thagard and Scott Findlay

14.1 Scientific Belief Revision

Scientists sometimes change their minds. A 2008 survey on the Edge Web site presented more than 100 self-reports of thinkers changing their minds about scientific and methodological issues (http://www.edge.org/q2008/q08_index.html). For example, Stephen Schneider, a Stanford biologist and climatologist, reported how new evidence in the 1970s led him to abandon his previously published belief that human atmospheric emissions would likely have a cooling rather than a warming effect. Instead, he came to believe – what is now widely accepted – that greenhouse gases such as carbon dioxide are contributing to the dramatic trend of global warming. Similarly, Laurence Smith, a UCLA geographer, reported how in 2007 he came to believe that major changes resulting from global warming will come much sooner than he had previously thought. Observations such as the major sea-ice collapse in Canada's Northwest Passage had not been predicted to occur so soon by available computational models, but indicated that climate change is happening much faster than expected. Evidence accumulated over the past three decades is widely taken to show that global warming will have major impacts on human life, and that policy changes such as reducing the production of greenhouse gases are urgently needed. However, such scientific and policy conclusions have received considerable resistance, for example from former American president George W. Bush and Canadian Prime Minister Stephen Harper.

A philosophical theory of belief revision should apply to the issue of global warming by explaining how most scientists have come to accept the following conclusions:

1. The earth is warming.
2. Warming will have devastating impacts on human society.

P. Thagard (✉)
Department of Philosophy, University of Waterloo, Waterloo, ON, Canada, N2L 3G1
e-mail: pthagard@uwaterloo.ca

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_14,

3. Greenhouse gas emissions are the main causes of warming.
4. Reduction in emissions is the best way to reduce the negative impacts of climate change.

In addition, the theory should provide insight not only into how scientists have come to adopt these beliefs, but also into why a few scientists and a larger number of leaders in business and politics have failed to adopt them.

We will show that belief revision about global warming can be modeled by a theory of explanatory coherence that has previously been applied to many cases of scientific belief change, including the major scientific revolutions (Thagard 1992). We will present a computer simulation of how current evidence supports acceptance of important conclusions about global warming on the basis of explanatory coherence. In addition, we will explain resistance to these conclusions using a computational model of emotional coherence, which shows how political and economic goals can bias the evaluation of evidence and produce irrational rejection of claims about global warming.

Theory evaluation in philosophy, as in science, is comparative, and we will argue that explanatory coherence gives a better account of belief revision than major alternatives. The main competitors are Bayesian theories based on probability theory and logicist theories that use formal logic to characterize the expansion and contraction of belief sets. We will argue that the theory of explanatory coherence is superior to these approaches on a number of major dimensions. Coherence theory provides a detailed account of the rational adoption of claims about climate change, and can also explain irrational resistance to these claims. Moreover, we will show that it is superior to alternatives with respect to computational complexity. This paper reviews the controversy about climate change, shows how explanatory coherence can model the acceptance of important hypotheses, and how emotional coherence can model resistance to belief revision. Finally, we will contrast the coherence account with Bayesian and logicist ones.

14.2 Climate Change

The modern history of beliefs about climate change began in 1896, when the Swedish scientist Svante Arrhenius discussed quantitatively the warming potential of carbon dioxide in the atmosphere (Arrhenius 1896, Weart 2003; see also http://www.aip.org/history/climate/index.html). The qualitative idea behind his calculations, now called the greenhouse effect, had been proposed by Joseph Fourier in 1824: the atmosphere lets through the light rays of the Sun but retains the heat from the ground. Arrhenius calculated that if carbon dioxide emissions doubled from their 1896 levels, the planet could face warming of 5–6°C. But such warming was thought to be far off and even beneficial.

In the 1960s, after Charles Keeling found that carbon dioxide levels in the atmosphere were rising annually, Syukuro Manabe and Richard Wetherland calculated that doubling the carbon dioxide in the atmosphere would raise global temperatures a couple of degrees. By 1977, scientific opinion was coming to accept global warming as the primary climate risk for the next century. Unusual weather patterns and devastating droughts in the 1970s and 1980s set the stage for Congressional testimony by James Hansen, head of the NASA Goddard Institute for Space Studies. He warned that storms, floods, and fatal heat waves would result from the long-term warming trend that humans were causing. In 1988, the Intergovernmental Panel on Climate Change (IPCC) was established and began to produce a series of influential reports (http://www.ipcc.ch/). They concluded on the basis of substantial evidence that humans are causing a greenhouse effect warming, and that serious warming is likely in the coming century. In 2006, former Congressman and presidential candidate Al Gore produced the influential documentary *An Inconvenient Truth*. This film, Gore's accompanying book, and the 2007 IPCC report helped solidify the view that major political action is needed to deal with the climate crisis (Gore 2006, IPCC 2007).

Nevertheless, there remains substantial resistance to the IPCC's conclusions, in three major forms. Some scientists claim that observed warming can be explained by natural fluctuations in energy output by the Sun, the Earth's orbital pattern, and natural aerosols from volcanoes. Others are skeptical that human-emitted carbon dioxide can actually enhance the greenhouse effect. However, the most popular opposition today accepts most of the scientific claims but contends that there is no imminent crisis and necessity of costly actions.

Corporations and special interest groups fund research skeptical of a human-caused global warming crisis. Such works usually appear in non-scientific publications such as the *Wall Street Journal*, often with the financial backing of the petroleum or automotive industries and links to political conservatism. Corporations such as ExxonMobil spend millions of dollars funding right-wing think tanks and supporting skeptical scientists. For example, the Competitive Enterprise Institute (CEI) is a libertarian think tank that received $2.2 million between 1998 and 2006 from ExxonMobil. CEI sponsors the website globalwarming.org which proclaims that policies being proposed to address global warming are not justified by current science and are a dangerous threat to freedom and prosperity.

The administration of U.S. President George W. Bush was highly influenced by the oil and energy industries consisting of corporations like ExxonMobil. The energy industry gave $48 million to Bush's 2000 campaign to become President, and has contributed $58 million in donations since then. Critics of the global warming crisis claim that there is a great deal of uncertainty associated with the findings of the IPCC: humans may not be the cause of recent warming. Moreover, government should play a very small role in any emission regulation, as property rights and the free market will foster environmental responsibility. The power of technology will naturally provide solutions to global warming, and any emission cuts would be harmful to the economy of the United States, which is not obligated to lead any fight against global warming. Values that serve as the backbone to these beliefs include: small government, individual liberty and property rights, global equality, technology, and economic stability.

In contrast, global warming activists such as Al Gore believe that scientific predictions of global warming are sound and that the planet faces an undeniable crisis. Evidence shows that humans are the cause of warming, and the world's major governments must play a crucial leadership role in the changes necessary to save the planet. Some individuals have been convinced by a combination of evidence and moral motivations to switch views from the skeptics' to the environmentalists' camp. For example, the Australian global warming activist Tim Flannery was once skeptical of the case scientists had made for action against global warming. Influenced by evidence collected in the form of the ice cap record, he gradually revised his beliefs to become a prominent proponent of drastic actions to fight global warming (Flannery 2006). Gore himself had to reject his childhood belief that the Earth is so vast and nature so powerful that nothing we do can have any major or lasting effect on the normal functioning of its natural systems. From the other direction, skeptics such as Bjørn Lomborg (2007) argue against the need for strong political actions to restrict carbon dioxide emissions. Let us now analyze the debate about global warming.

14.3 Coherence and Revision

The structure of the inferences that the Earth is warming because of production of greenhouse gases can be analyzed using the theory of explanatory coherence and the computer model ECHO. The theory and model have already been applied to a great many examples of inference in science, law, and everyday life (see Thagard 1989, 1992, 1999, 2000; Nowak and Thagard 1992a, 1992b). The theory of explanatory coherence consists of the following principles:

Principle 1. Symmetry. Explanatory coherence is a symmetric relation, unlike, say, conditional probability. That is, two propositions *p* and *q* cohere with each other equally.

Principle 2. Explanation. (a) A hypothesis coheres with what it explains, which can either be evidence or another hypothesis; (b) hypotheses that together explain some other proposition cohere with each other; and (c) the more hypotheses it takes to explain something, the lower the degree of coherence.

Principle 3. Analogy. Similar hypotheses that explain similar pieces of evidence cohere.

Principle 4. Data priority. Propositions that describe the results of observations have a degree of acceptability on their own.

Principle 5. Contradiction. Contradictory propositions are incoherent with each other.

Principle 6. Competition. If *P* and *Q* both explain a proposition, and if *P* and *Q* are not explanatorily connected, then *P* and *Q* are incoherent with each other. (*P* and *Q* are explanatorily connected if one explains the other or if together they explain something.)

Principle 7. Acceptance. The acceptability of a proposition in a system of propositions depends on its coherence with them.

These principles do not fully specify how to determine coherence-based acceptance, but algorithms are available that can compute acceptance and rejection of propositions on the basis of coherence relations. The most psychologically natural algorithms use artificial neural networks that represent propositions by artificial neurons or *units* and represent coherence and incoherence relations by excitatory and inhibitory links between the units that represent the propositions. Acceptance or rejection of a proposition is represented by the degree of activation of the unit. The program ECHO spreads activation among all units in a network until some units are activated and others are inactivated, in a way that maximizes the coherence of all the propositions represented by the units. I will not present the technical details here, as they are available elsewhere (Thagard 1992, 2000). Several different algorithms for computing coherence are analyzed in Thagard and Verbeurgt (1998).

The problem of scientific belief revision concerns how to deal with situations where new evidence or hypotheses generate the need to consider rejecting beliefs that have previously been accepted. According to the theory of explanatory coherence, belief revision should and often does proceed by evaluating all the relevant hypotheses with respect to all the evidence. A scientific data base consists primarily of a set of propositions describing evidence and hypotheses that explain them. There are coherence relations between pairs of propositions in accord with principle 1: when a hypothesis explains a piece of evidence, they cohere. There are also incoherence relations between pairs in accord with principles 5 and 6. When a new proposition comes along, representing either newly discovered evidence or a newly generated explanatory hypothesis, then this proposition is added to the overall set, along with positive and negative constraints based on the relations of coherence and incoherence that the new proposition has with the old ones. Then an assessment of coherence is performed in accord with principle 7, with the results telling you what to accept and what to reject. Belief revision takes place when a new proposition has sufficient coherence with the entire set of propositions that it becomes accepted and some proposition previously accepted becomes rejected.

Because a variety of algorithms are available for computing coherence, belief revision can be modeled in a highly effective and computationally efficient manner, involving substantial numbers of propositions. For example, Nowak and Thagard (1992a) simulated the acceptance of Copernicus' theory of the solar system and the rejection of Ptolemy's, with a total belief set of over 100 propositions. The LISP code for ECHO and various simulations is available on the Web at: http://cogsci.uwaterloo.ca/Index.html. This site also makes available a partial JAVA version of ECHO.

Explanatory coherence is not intended to be a logically complete theory of belief revision, because it does not take into account a full range of operators such as conjunction and disjunction. Most emphatically, when a new proposition is added to a belief system, no attempt is made to add all its logical consequences, an infinitely large set beyond the power of any human or computer. Nevertheless, explanatory coherence gives a good account of what Gärdenfors (1988, 1992) describes as the three main kinds of belief change: expansion, revision, and contraction. Expansion takes place when a new proposition is introduced into a belief system, becoming accepted if and only if doing so maximizes coherence. Revision occurs when a new proposition is introduced into a belief system and leads other previously accepted propositions to be rejected because maximizing coherence requires accepting the new proposition and rejecting one or more old ones. Contraction occurs when some proposition becomes rejected because it no longer helps to maximize coherence. Simulations of these processes are described in the next section.

We do not use "maximize coherence" as a vague metaphor like most coherentist epistemologists, but rather as a computationally precise notion whose details are available elsewhere (e.g. Thagard 1992, 2000). Logicist theories view belief revision as the result of expansion followed by contraction, but explanatory coherence computes expansion and contraction as happening at the same time in parallel.

Scientific belief revision comes in various degrees. A new proposition describing recently collected evidence may become accepted easily unless it does not fit well with accepted views. Such acceptance would be a simple case of expansion. However, if the new evidence is not easily explained by existing hypotheses, scientists may generate a new hypothesis to explain it. If the new hypothesis conflicts with existing hypotheses, either because it contradicts them or competes as an alternative hypothesis for other evidence, then major belief revision is required. Such revision may lead to theory change, in which one set of hypotheses is replaced by another set, as happens in scientific revolutions. The development of new ideas about climate change has not been revolutionary, since no major scientific theories have had to be rejected. But let us now look in more detail at how explanatory coherence can model the acceptance of global warming.

14.4 Simulating Belief Revision About Climate Change

To show how explanatory coherence can be used to simulate belief revision, we begin with a series of simple examples shown in Fig. 14.1. Simulation A shows the simplest possible case where there is just one piece of evidence and one hypothesis that explains it. In accord with principle 4 of explanatory coherence, evidence propositions have a degree of acceptability on their own, which in the program ECHO is implemented by their being a positive constraint between each of them and a special unit EVIDENCE that is always accepted. Hence in simulation A, the acceptance of E1 leads to the acceptance of H1 as well.

Simulation B depicts a simple case of expansion beyond simulation A, in which a new piece of evidence is added and accepted. The situation gets more interesting in simulation C, where the hypothesis explains a predicted evidence proposition PE3, which however contradicts the actually observed evidence E4. A Popperian would say that H1 has now been refuted and therefore should be rejected, but one failed prediction rarely leads to the rejection of a theory, for good reasons: perhaps

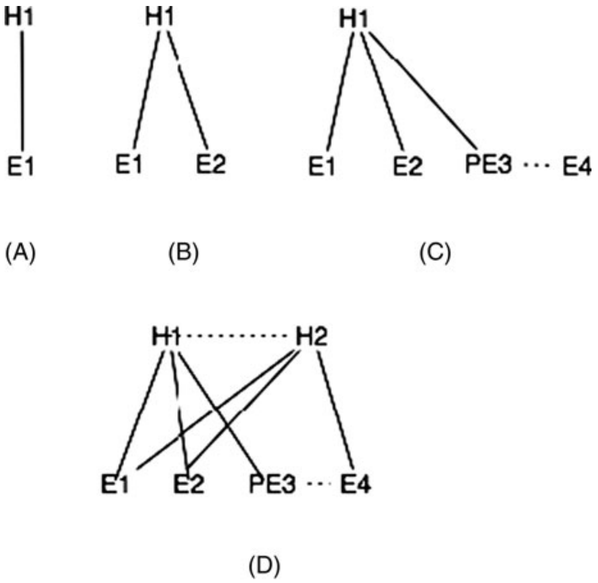


Fig. 14.1 The straight lines indicate coherence relations (positive constraints) established because a hypothesis explains a piece of evidence. The dotted lines indicate incoherence relations (negative constraints). Coherence relations between an evidence unit and E1, E2, and E4 are not shown

the experiment that produced E4 was flawed, or there were other intervening factors that made the prediction of PE3 incorrect. Hence ECHO retains H1 while accepting E4 and rejecting PE3. Refutation of H1 requires the availability of an alternative hypothesis, as shown in simulation D. Here the addition of H2 provides an alternative explanation of the evidence, leading to its acceptance by virtue of its explanation of E4 as well as E1 and E2. This is a classic case of inference to the best explanation, where belief revision is accomplished through simultaneous expansion (the addition of H2) and retraction (the deletion of H1).

Belief revision about climate change can naturally be understood as a case of inference to the best explanation based on explanatory coherence. Figure 14.2 displays a drastically simplified simulation of the conflict between proponents of the view that climate change is caused by human activities and their critics. The hypothesis that is most important because it has major policy implications is that humans cause global warming. This hypothesis explains many observed phenomena, but Fig. 14.2 shows only two crucial generalizations from evidence: global temperatures are rising and the recent rise has been rapid. The main current alternative explanation is that rising temperatures are just the result of natural fluctuations in temperature that have frequently occurred throughout the history of the Earth. Figure 14.2 also shows the favored explanation of how humans have caused global warming, through the greenhouse effect in which gases such as carbon dioxide and methane prevent energy radiation into space. Human industrial activity has produced huge increases in the amounts of such gases in the atmosphere over the past few hundred years.

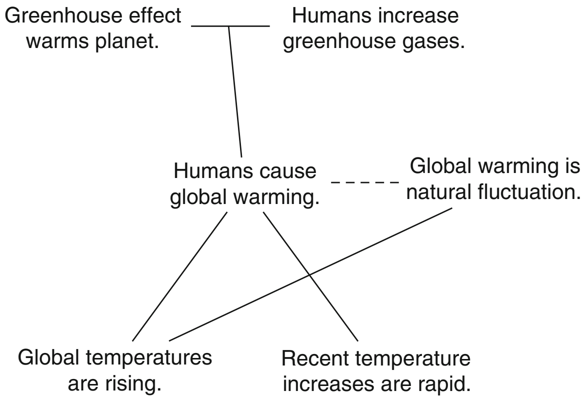

Greenhouse effect warms planet.
Humans increase greenhouse gases.
Humans cause global warming.
Global warming is natural fluctuation.
Global temperatures are rising.
Recent temperature increases are rapid.

Fig. 14.2 Highly simplified view of part of the controversy over climate change, with straight lines indicating coherence relations and dotted lines indicating incoherence ones

Figure 14.2 shows only a small part of the explanatory structure of the controversy over climate change, and our full analysis is presented in Fig. 14.3, with more pieces of evidence and a fuller range of hypotheses. The input to our simulation using the program ECHO can be found in the appendix. The key competing hypotheses are GH3, that global warming is caused by humans, and NH4, global warming is a natural fluctuation. As you would expect from the greater connectivity of hypothesis GH3 with the evidence, it wins out over NH4, which is rejected. The inputs to ECHO given in the appendix and the constraint structures shown in Fig. 14.3 capture much of the logical structure of the current debate over climate change. In accord with the current scientific consensus, ECHO accepts the basic claim that climate change is being caused by human activities that increase greenhouse gases.

ECHO can model belief revision in the previously skeptical by simulating what happens if only some of the evidence and explanations shown in Fig. 14.3 are available. For example, we have run a simulation that deletes most of the evidence for GH3 as well as the facts supporting GH1. In this case, ECHO finds NH1 more acceptable than GH3, rejecting the claim that humans are responsible for global warming. Adding back in the deleted evidence and the explanations of it by GH3 and the other global warming hypotheses produces a simulation of belief revision of the sort that would occur if a critic of human warming was presented with more and more evidence. The result eventually is a kind of gestalt switch, a tipping point in which the overall relations of explanatory coherence produce adoption of new views and rejection of old ones. Thus explanatory coherence can explain the move toward general acceptance of views currently dominant in the scientific community about climate change.

What explanatory coherence can *not* explain is why some political and business leaders remain highly skeptical about global warming caused by human activities and the need to take drastic measures to curtail it. To understand their resistance, we need to expand the explanatory coherence model by taking into account emotional attitudes.

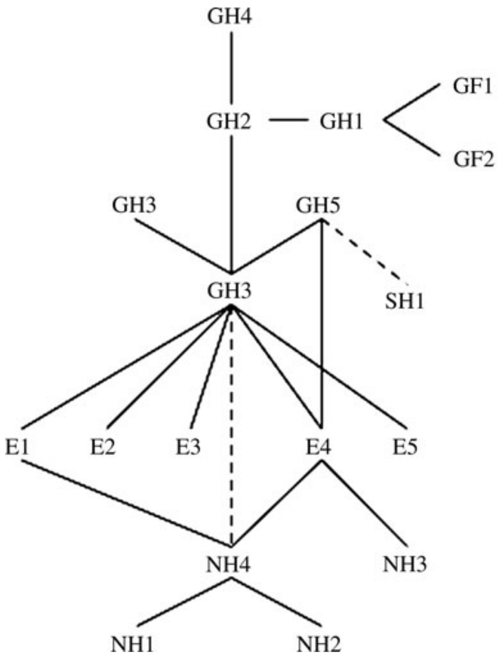


Fig. 14.3 More detailed analysis of the controversy over climate change, with straight lines indicating coherence relations and dotted lines indicating incoherence ones. See the appendix for full description of the propositions and their relations

14.5 Simulating Resistance to Belief Revision

Scientific theory choice has the same logical structure as juror decisions in criminal cases. Just as scientists need to decide whether a proposed theory is the best explanation of the experimental evidence, juries need to decide whether the prosecutor's claim that the accused committed a crime is the best explanation of the criminal evidence. Ideally, juries are supposed to take into account all the available evidence and consider alternatives explanations of it. Often they do, but juries are like all people including scientists in having emotional biases that can lead to different verdicts than the one that provides the best explanation of the evidence. Thagard (2003, 2006, ch. 8) analyzed the decision of the jury in the O. J. Simpson case: explanatory coherence with respect to the available evidence should have led to the conclusion that Simpson was guilty, but the jury nevertheless acquitted them. However, jurors' decision to acquit was simulated using the program HOTCO that simulates "hot coherence", which includes the contribution of emotional values to belief revision. Emotional values are a perfectly legitimate part of decision making as psychological indicators of the costs and benefits of expected outcomes. In the language of decision theory, deliberation requires utilities as well as probabilities. But normatively belief revision should depend on evidence, not utilities or emotional values.

We have already mentioned the motivations that lead some business and political leaders to be skeptical about claims about global warming. If climate change is a serious problem caused by human production of greenhouse gases, then measures need to be taken to curtail such production. Particularly affected by such measures will be oil companies, so it is not surprising that the research aimed at defusing alarm about global warming has been heavily supported by them. Moreover, some of the most powerful opposition to the Kyoto Protocol and other attempts to deal with global warming have come from politicians closely allied with the oil industry, such as former American president George W. Bush and Canadian Prime Minister Stephen Harper. In 2002, when he was leader of the Alberta-based Canadian Alliance which later merged with the Conservative Party that he now leads, Harper wrote:

> We're gearing up for the biggest struggle our party has faced since you entrusted me with the leadership. I'm talking about the "battle of Kyoto" – our campaign to block the job-killing, economy-destroying Kyoto Accord.
> It would take more than one letter to explain what's wrong with Kyoto, but here are a few facts about this so-called "Accord":
> – It's based on tentative and contradictory scientific evidence about climate trends.
> – It focuses on carbon dioxide, which is essential to life, rather than upon pollutants.
> – Canada is the only country in the world required to make significant cuts in emissions. Third World countries are exempt, the Europeans get credit for shutting down inefficient Soviet-era industries, and no country in the Western hemisphere except Canada is signing.
> – Implementing Kyoto will cripple the oil and gas industry, which is essential to the economies of Newfoundland, Nova Scotia, Saskatchewan, Alberta and British Columbia.
> – As the effects trickle through other industries, workers and consumers everywhere in Canada will lose. THERE ARE NO CANADIAN WINNERS UNDER THE KYOTO ACCORD.
> — The only winners will be countries such as Russia, India, and China, from which Canada will have to buy "emissions credits." Kyoto is essentially a socialist scheme to suck money out of wealth-producing nations. (http://www.thestar.com/article/176382)

Prime Minister Harper has since moderated his position on global warming, as has George W. Bush, but both have been slow to implement any practical changes.

We conjecture that at the root of opposition to claims about global warming are the following concerns. Dealing with climate change would require government intervention to restrict oil usage, which is doubly bad for a conservative politician with a preference for free market solutions and a long history of association with oil producing companies. Figure 14.4 expands Fig. 14.2 to include the strong emotional values of avoiding limiting oil use and production and avoiding government intervention in the economy. When the explanatory coherence relations shown in the appendix and Fig. 14.3 are supplanted with these values, belief revision is retarded, so that more evidence is required to shift from rejection to acceptance of the hypothesis that global warming is being caused by human activities.

In the ECHO simulation of just the hypotheses and evidence shown in Fig. 14.4, the obvious result is the acceptance of the hypothesis that global warming is caused by humans, implying that political action can and should be taken against it. But we have also used the extended program HOTCO to yield a different result. If, as Fig. 14.4 suggests, the hypothesis that humans caused global warming is taken to conflict with the values of avoiding oil limitations and government intervention, then the simulation results in the rejection of human causes for global warming and

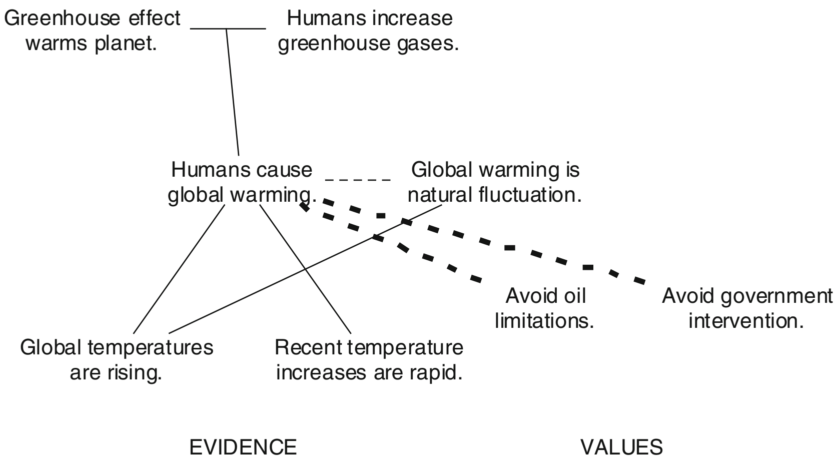

Greenhouse effect warms planet.
Humans increase greenhouse gases.
Humans cause global warming.
Global warming is natural fluctuation.
Avoid oil limitations.
Avoid government intervention.
Global temperatures are rising.
Recent temperature increases are rapid.
EVIDENCE
VALUES

Fig. 14.4 View of the controversy over climate change including emotional constraints as well as explanatory ones. As in previous figures, the solid lines indicate positive constraints based on explanatory relations and the thin dotted line indicates a negative constraint based on incompatibility. The thick dotted lines indicate negative emotional constraints

acceptance of the alternative hypothesis of natural fluctuation. Of, course, the actual psychological process of motivated inference in this case is much more complex than Fig. 14.4 portrays, leading skeptics to challenge evidence and explanations as well as hypotheses. But Fig. 14.4 and the HOTCO simulation of it show how values can interfere with belief revision by undermining hypotheses that are better supported by the evidence. Hence a psychologically natural extension to ECHO can explain resistance to belief revision.

14.6 Alternative Theories of Belief Revision

Explanatory coherence is not the only available account of scientific belief revision, which include at least the following: Popper's conjectures and refutations, Hempel's confirmation theory, Kuhn's paradigm shifts, Lakatos's methodology of research programs, and social constructionist claims that scientists revise their beliefs only to increase their own power. These are all much vaguer than the theory of explanatory coherence and its computational implementation in ECHO. Among formally exact characterization of belief revision, the two most powerful approaches are Bayesian ones that explain belief change using probability theory, and logicist ones that use ideas about logical consequence in deductive systems.

Detailed comparisons between explanatory coherence and Bayesian accounts of belief revision have been presented elsewhere (Thagard 2000, ch. 8; Eliasmith and Thagard 1997; Thagard 2004; Thagard and Litt 2008). It should be feasible to produce a Bayesian computer simulation of the full climate change case shown in Fig. 14.3 and the appendix. A simulator such as JavaBayes (http://www.cs.cmu.edu/~javabayes/Home/) could be used to produce a Bayesian alternative to our ECHO simulator, as was done for a rich legal example in Thagard (2004). But doing so would require a very large number of conditional probabilities whose interpretation and provenance are highly problematic. In the simplest case where you have two hypotheses, H1 and H2, explaining a piece of evidence E1, JavaBayes would require specification of eight conditional probabilities, such as P(E1/H1 and not-H2). For example, the simplified model shown in Fig. 14.2 would require specification of P(global temperatures are rising/humans cause global warming and global warming is not a natural fluctuation) as well as seven other conditional probabilities. In general, the nodes in an explanatory coherence network can be translated into nodes in a Bayesian network with the links translated into directional arrows. The price to be paid is that for each node that has *n* arrows coming into it, it is necessary to specify 2^{n+1} conditional probabilities. In our ECHO simulation specified in the appendix, a highly connected node such as E4 which has three global warming hypotheses and four alternative hypotheses explaining it would require specification of $2^8 = 256$ conditional probabilities, none of which can be estimated from available data. To produce a JavaBayes alternative to our full ECHO simulation, thousands of conditional probabilities would simply have to be made up. Bayesian models are very useful for cases where there are large amounts of statistical data, such as sensory processing in humans and robots; but they add nothing to understanding of cases of belief revision such as climate change where probabilities are unknown.

Writers such as Olsson (2005) have criticized coherence theories for not being compatible with probability theory, but probability theory seems to us irrelevant in qualitative cases of theory change. The comprehensive report of the Intergovernmental Panel on Climate Change sensibly relies on qualitative characterizations such as “extremely likely” and “very high confidence”. Probability theory should be used whenever appropriate for statistics-based inference, but applying it to qualitative cases of causal reasoning such as climate changes obscures more than it illuminates.

The other major formal approach to belief revision uses techniques of formal logic to characterize the expansion and contraction of belief sets (see e.g. Gärdenfors 1988, 1992; Tennant 1994, 2006). We will not attempt a full discussion, but the explanatory coherence approach seems to us superior to logicist approaches in several respects that we will briefly describe.

First, the explanatory coherence approach is both broad and deep, having been applied in detail to many important cases of scientific belief revision. In this paper, we have shown its applicability to a belief revision problem of great contemporary interest, climate change. To our knowledge, logicist approaches to belief revision have not served to model any historical or contemporary cases of belief change. Explanatory coherence has less generality than logicist approaches, because it does not attempt algorithms for computing revision and contraction functions for an *arbitrary* belief system. Rather, it focuses on revision in systems of hypotheses and evidence, which suffices for the most important kinds of belief change in science, criminal cases, and everyday life.

Second, the explanatory coherence approach has philosophical advantages over logicist approaches. It does not assume that that changes in belief systems should be *minimal*, retaining as much as possible from our old beliefs. The principle of minimal change has often been advocated (e.g. by Gärdenfors, Tennant, Harman, and Quine) but rarely defended, and Rott (2000) has argued that it is not even appropriate for logicist approaches. Aiming for minimal change in belief systems seems to us no more justifiable than aiming for minimal change in political and social systems. Just as political conservatism should not serve to block needed social changes, so epistemic minimalism should not get in the way of needed belief changes. Explanatory coherence theory shows how to make just the changes that are needed to maximize the coherence of evidence and hypotheses. As long as they are productive, both epistemic and social innovations are to be valued. Just as aiming for minimal change in production of greenhouse gases may prevent dealing adequately with forthcoming climate crises, so aiming for minimal change in belief revision may prevent arriving at maximally coherent and arguably true theories. Coherence approaches have often been chided for neglecting the importance of truth as an epistemic goal, but Thagard (2007) argues that applications of explanatory coherence in science do in fact lead to truth when they produce the adoption of theories that are both broad (explaining much evidence) and deep (possessing underlying mechanistic explanations).

Third, explanatory coherence has computational advantages over logicist approaches that assume that a belief set is logically closed. Then belief sets are infinite, and cannot be computed by any finite machine. The restricted problem of belief contraction in finite beliefs sets has been shown to be NP-complete (Tennant 2003). Problems in NP are usually taken by computer scientists to present computational difficulties, in that polynomial-time solutions are not available so they are characterized as intractable. It might seem that the coherence approach is in the same awkward boat, as Thagard and Verbeurgt (1998) showed that the coherence problem is NP-hard. However, many hard computational problems become tractable if a problem parameter is fixed or bounded by a fixed value (Gottlob et al. 2002; van Rooij 2008). Van Rooij has shown that coherence problems become fixed-parameter tractable if they have a high ratio of positive to negative constraints. Fortunately, all programmed examples of explanatory coherence have a high positive-negative ratio, as you would expect from the fact that most constraints are generated by explanatory coherence principle 2. In particular, our simulation of the climate change case using the input in the appendix generates 10 negative constraints and 53 positive ones. Hence explanatory coherence in realistic applications appears to be computationally tractable.

Fourth, explanatory coherence has a major psychological advantages over logistic approaches, which are not intended to model how people actually perform belief revision, but rather how belief revision should ideally take place. Explanatory coherence theory has affinities with a whole class of coherence models that have been applied to a wide range of psychological phenomena (Thagard 2000). The psychological plausibility of explanatory coherence over logicist approaches will not be appreciated by those who like their epistemology to take a Platonic or Fregean form, but is a clear advantage for those of us who prefer a naturalistic approach to epistemology. A related psychological advantage is that explanatory coherence meshes with psychological accounts that can explain what people are doing when they irrationally resist belief revision, as we saw in the case of political opposition to theories of climate change.

14.7 Conclusion

Our brief discussions of Bayesian and logicist accounts of belief revision hardly constitute a refutation of these powerful approaches, but should suffice to indicate how they differ from the explanatory coherence approach. We have shown how explanatory and emotional coherence can illuminate current debates about climate change. We can use it to understand the rational adoption of the hypothesis that global warming is caused by human activities. Moreover, deviations from rational belief revision in the form of emotion-induced rejection of the best explanation of a wide range of evidence can be understood in terms of intrusion of emotional political values into the assessment of the best explanation.

Scientific belief revision is not supposed to be impeded by emotional prejudices, but such impedance is common. The acceptance of Darwin's theory of evolution by natural selection was slowed in the nineteenth century by fears of its threat to established theological theories of the creation of species. Such resistance continues today, as many people – but very few scientists – continue to reject evolutionary theory as incompatible with their religious beliefs and desires. In practical settings, encouraging belief change about Darwin's theories, as about climate change, requires dealing with emotional constraints as well as cognitive ones generated by the available evidence, hypotheses, and explanations. Thus a broadly applicable, computationally feasible, and psychologically insightful account of belief revision such as that provided by the theory of explanatory coherence should be practically useful as well as philosophically informative.

References

Arrhenius, S. 1896. On the influence of carbonic acid in the air upon the temperature of the ground. *Philosophical Magazine and Journal of Science* 41:237–276.

Eliasmith, C., and P. Thagard. 1997. Waves, particles, and explanatory coherence. *British Journal for the Philosophy of Science* 48:1–19.

Flannery, T. 2006. *The weather makers*. Toronto: Harper Collins.

Gärdenfors, P. 1988. *Knowledge in flux*. Cambridge, MA: MIT Press/Bradford Books.

Gärdenfors, P. (ed.) (1992). *Belief revision*. Cambridge: Cambridge University Press.

Gore, A. 2006. *An inconvenient truth*. Emmaus, PA: Rodale.

Gottlob, G., F. Scarcello, and M. Sideri. 2002. Fixed-parameter complexity in AI and nonmonotonic reasoning. *Artificial Intelligence* 138:55–86.

IPCC. 2007. *IPCC Fourth assessment report*. Retrieved July 18, 2008, from http://www.ipcc.ch/ipccreports/ar4-syr.htm

Lomborg, B. 2007. *Cool it: The skeptical environmentalist's guide to global warming*. Toronto, ON: Random House.

Nowak, G., and Thagard, P. 1992a. Copernicus, ptolemy, and explanatory coherence. In *Cognitive models of science,* ed. R. Giere, vol. 15, 274–309. Minneapolis, MN: University of Minnesota Press.

Nowak, G., and Thagard, P. 1992b. Newton, descartes, and explanatory coherence. In *Philosophy of science, cognitive psychology and educational theory and practice*, eds. R. Duschl, and R. Hamilton, 69–115. Albany, NY: SUNY.

Olsson, E. 2005. *Against coherence: Truth, probability, and justification*. Oxford: Oxford University Press.

Rott, H. 2000. Two dogmas of belief revision. *Journal of Philosophy* 97:503–522.

Tennant, N. 1994. Changing the theory of theory change: Towards a computational approach. *British Journal for the Philosophy of Science* 45:865–897.

Tennant, N. 2003. Theory-contraction is NP-complete. *Logic Journal of the IGPL* 11:675–693.

Tennant, N. 2006. New foundations for a relational theory of theory-revision. *Journal of Philosophical Logic* 35:489–528.

Thagard, P. 1989. Explanatory coherence. *Behavioral and Brain Sciences* 12:435–467.

Thagard, P. 1992. *Conceptual revolutions*. Princeton, NJ: Princeton University Press.

Thagard, P. 1999. *How scientists explain disease*. Princeton, NJ: Princeton University Press.

Thagard, P. 2000. *Coherence in thought and action*. Cambridge, MA: MIT Press.

Thagard, P. 2003. Why wasn't O. J. convicted? Emotional coherence in legal inference. *Cognition and Emotion* 17:361–383.

Thagard, P. 2004. Causal inference in legal decision making: Explanatory coherence vs. Bayesian networks. *Applied Artificial Intelligence* 18:231–249.

Thagard, P. 2006. *Hot thought: Mechanisms and applications of emotional cognition*. Cambridge, MA: MIT Press.

Thagard, P. 2007. Coherence, truth, and the development of scientific knowledge. *Philosophy of Science* 74:28–47.

Thagard, P., and A. Litt. (2008). Models of scientific explanation. In *The Cambridge handbook of computational psychology*, ed. R. Sun, 549–564. Cambridge: Cambridge University Press.

Thagard, P., and K. Verbeurgt. 1998. Coherence as constraint satisfaction. *Cognitive Science* 22: 1–24.

van Rooij, I. 2008. The tractable cognition thesis. *Cognitive Science* 32:939–984.

Weart, S.R. 2003. *The discovery of global warming*. Cambridge, MA: Cambridge University Press.

Appendix

Input to the ECHO simulation of the acceptance of the claim that global warming is caused by humans.

Global warming: A simplified model of anthropogenic forcing vs. natural causes.

Evidence:

E1. Average global temperatures have risen significantly since 1880.
E2. The rate of warming is rapidly increasing.
E3. The recent warming is more extreme than any other warming period as far back as the record shows to 1000 AD.
E4. Arctic ice is rapidly melting and glaciers around the world are retreating.
E5. Global temperature shows strong correlation with carbon dioxide levels throughout history.

IPCC/Gore's facts

GF1. Carbon dioxide, methane gas, and water vapour are greenhouse gasses.
GF2. Greenhouse gasses absorb infrared radiation, some of which is reemitted back to the Earth's surface.
GF3. Carbon dioxide levels in the atmosphere have been increasing since the beginning of the industrial revolution.

IPCC/Gore's main hypotheses: "Anthropogenic forcing"

GH1. There is a greenhouse effect that warms the planet.
GH2. The greenhouse effect has the potential to be enhanced.
GH3. Global warming is a human caused crisis.

Secondary hypotheses

GH4. Increasing the concentration of greenhouse gasses in the atmosphere directly increases the warming of the Earth.
GH5. Small changes in global temperature have the potential to drastically upset a variety of climate systems through causal interactions.

Opposing hypotheses/beliefs

NH1. Long term cycling of Earth's orbital parameters, solar activity and volcanism and associated aerosols are natural causes that can warm the globe.
NH2. The impact of natural factors on global temperature dwarfs the enhanced greenhouse effect.
NH3. Climate systems will be affected by natural cycles and fluctuations.
NH4. Global warming is natural and not a concern.
SH1. Small changes in temperature will not have significant negative effects on global climate.

Explanations:

Gore's explanations:

of the enhanced greenhouse effect and anthropogenic forcing.
explain (GF1, GF2) GH1
explain (GH1, GH4) GH2
explain (GH2, GF3, GH5) GH3

of the evidence:
explain (GH2, GH3) E1
explain (GH2, GH3) E2
explain (GH2, GH3, GH4, GF3) E3
explain (GH2, GH3, GH5) E4
explain (GH2, GH3, GH4, GF3) E5

Natural explanations:

of a natural cause for global warming:
explain (NH1, NH2) NH4

of the evidence:
explain (NH1, NH2, NH4) E1
explain (NH1, NH2, NH3, NH4) E4

Contradictions:

contradict NH4 GH3
contradict NH2 GH2
contradict GH5 SH1

Chapter 15
Rationality in Flux – Formal Representations of Methodological Change

Jonas Nilsson and Sten Lindström

A central aim for philosophers of science has been to understand scientific theory change, or more specifically the rationality of theory change. Philosophers and historians of science have suggested that not only theories but also scientific *methods* and *standards* of rational inquiry have changed through the history of science. The topic here is methodological change, and what kind of theory of rational methodological change is appropriate. The modest ambition of this paper is to discuss in what ways results in formal theories of belief revision can throw light on the question of what an appropriate theory of methodological change would look like.

15.1 Methodological States

Let us start by introducing the term "methodological state". Apart from beliefs, theories and cognitive goals, an agent involved in scientific research has a number of methodological rules or standards of scientific rationality. These standards are of different kinds. Some standards are *heuristic*, prescribing that one should do, or try to do, certain things, such as "try to find causal explanations for observed phenomena", "test theories by making controlled experiments" or "avoid ad hoc hypotheses". Other standards are *evaluative*, telling us for example how we should choose between competing theories: "prefer theories which have been used to make novel predictions over theories which have merely been made to square with the observations made", "prefer simpler theories to less simple ones" or "a successor theory must retain all the corroborated empirical content of its predecessors". General principles of rationality, such as "act in such a way that you promote your goals" or "be prepared to listen to criticism of your beliefs" also function as standards. Logical laws and inference rules may also give rise to standards of rationality,

J. Nilsson (✉)
Department of Historical, Philosophical and Religious Studies, Umeå University, Sweden
e-mail: jonas.nilsson@philos.umu.se

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8_15,

or so it seems. For example: “You ought not to have contradictory beliefs” or “You ought to believe obvious logical consequences of what you believe”.

Furthermore, there are meta-standards pertaining to the evaluation of other standards (or methods), such as “a method which is more reliable should be preferred to an alternative method which is less reliable” or “general methods should be coherent with judgments about the rationality of particular episodes in the history of science”.

An agent, which may be an individual or a group of researchers, accepts a large number of normative methods or standards of these different types. All the standards accepted by an agent at a particular time constitute that agent’s methodological state. If an agent comes to accept a new standard, or reject one she previously accepted, she moves from one methodological state to a new such state.

15.2 Philosophical Theories of Methodological Change

Whereas some have seen previous changes of scientific standards as pervasive and have tried to formulate models in which all standards are seen as open, in principle, to revision (see Briskman 1977; Laudan 1984, 1996; Shapere 1984), others have instead argued that such changes as there have been are peripheral and that they can be explained as rational (or irrational) on the basis of some core set of standards that has remained constant (Worrall 1988; Newton-Smith 1981). We shall not try to take a stand on the issue of how extensive previous methodological changes have been, or discuss whether all standards are in principle revisable or if some standards must be treated as immune to revision. Instead we will merely assume that there has been methodological change in science, and that current standards are themselves open to future improvement. The question we want to consider here is what kind of philosophical account should be given of such methodological change.

Say that a certain methodological change is made in some scientific field: An old method is rejected, or a new one is added. If this is rational, there has to be some standard or method, which is used to evaluate the initial methodological state as problematic, and to evaluate the change to a new methodological state as rational. How, then, are standards evaluated? What standards or methods should be applied to determine the rationality of methodological change?

Philosophers of science who discuss methodological change often think of standards in an instrumental way, and their discussions depart from some *axiology*: some conception of the goal or goals of science. This axiology is presupposed as a background when discussing what the appropriate scientific standards are, and how methodological change should be evaluated. Among possible goals that are often mentioned are truth, true explanatory theories, maximizing predictive power, high verisimilitude, empirical adequacy, or problem-solving ability. A usual meta-methodological strategy is then to find some subgoal which is appropriately related to the ultimate goals of science.

Popper, for instance, took something like approach to general, true explanatory theories to be the goal of science, and proposed that a proper subgoal to aim at is to maximize the degree of falsifiability of theories and test them severely (Popper 1989). Newton-Smith proposes that the ultimate goal of science is theories with high verisimilitude, but that what we must aim at is theories with long-term observational success.

According to Popper and others, the considerations used to select standards (and thus to evaluate methodological change) are broadly logical and philosophical: which standards will ensure that increasingly falsifiable theories are proposed and tested?[1]

According to Newton-Smith, Laudan and others, the considerations are instead empirical: what standards have actually contributed to the selection of theories with such properties as long-term observational success (Newton-Smith 1981) or problem solving efficiency (Laudan 1984, 1996)? Empirical theories of methodological change have been rather popular, and are often associated with naturalistic conceptions of scientific method and methodological change.

15.3 Fixed Core Theories and Bootstrap Theories

What should a general theory of methodological change look like then? Suppose an agent revises her initial methodological state S_1 in such a way that she enters the new state S_2. In a theory of methodological change, one would like to answer questions like the following: When is it rational to revise a methodological state? And, when is it rational for an agent to make the transition from methodological state S_1 to a new state S_2? In answering these questions, it is relevant to consider two other questions.

(1) Is there a specific *core* of standards (meta-standards) for evaluating methodological change, or is the evaluation of standards and methodological change more varied and pluralistic, so that in principle any method or standard may be relevant to the evaluation of standards? Should a theory of methodological change be a "*core theory*" or a *pluralistic theory*?

The latter alternative strikes us as more plausible. It seems reasonable that different kinds of considerations – empirical, logical or broadly philosophical – may be relevant for evaluating standards. We think this provides part of the motivation for a bootstrap theory.

(2) The other question is this: Should standards for evaluating methodological change be regarded as necessarily *fixed*, or themselves open to change and improvement? For every pair of methodological states $<S_1, \ S_2>$, which is such that the transition from S_1 to S_2 is rational, is there a set of meta-standards according to which all such transitions are rational? That is, should a theory of methodological change be a *static* theory or a *dynamic* theory?

Again, the latter alternative strikes us as more plausible. If improvements in standards of empirical testing is possible, or if improvement of logical and formal

[1] This formal approach to the evaluation of methods is defended in Niiniluoto (1999, ch. 6).

standards is possible, it seems reasonable that such improvement could also benefit our resources for evaluating methodological change. This, we think, is a further motivation for exploring bootstrap theories of methodological change.

A *static core theory* would (if fully spelt out) contain a set of standards, which are used to evaluate other standards and changes from one methodological state to another. Outlines of such theories have been sketched by for example Newton-Smith (1981) and Worrall (1982, 1989), and we think that some such theory is presupposed by many philosophers of science who discuss methodological change.

The main motivations for a *bootstrap theory* of methodological change are instead that the evaluation of standards is likely to be a pluralistic matter – in different situations different standards or methods may be applicable – and that a theory of methodological change should itself be dynamic – one does not want to exclude the possibility that the standards used to evaluate methodological change may themselves be improved as science progresses. We shall not try to argue here that a theory of methodological change should take the form of a bootstrap theory rather than a fixed core theory, but rest content with indicating why it is an interesting alternative worthy of further development.

15.4 Outline of a Bootstrap Theory of Methodological Change

What does a bootstrap theory say? Let us say that as a science develops it goes through not only a sequence of theoretical states, but also a sequence of methodological states. A methodological state is here seen as the set of scientific standards or methods accepted at a certain point in time (in a field or by a group of scientists).

The main bootstrap idea is that some standards in such a methodological state are used to evaluate certain other standards or methods, or the state as a whole, as problematical. Therefore, what particular standards are used for the purpose of evaluating methodological change varies with the particular type of problem detected (say, a logical problem, or empirical evidence suggesting some method is unreliable). Furthermore, what standards are available for evaluating methodological change may change from one stage in scientific inquiry to another.

A bootstrap theory is neutral about which specific standards should be used to evaluate methodological states and methodological change. It is thus compatible with a pluralist view of the evaluation of methodological change, and with a dynamic view of standards for methodological change.

The bootstrap idea is instead to lay down requirements for how standards accepted at a particular point in time (making up a methodological state) may be used to evaluate other standards or a methodological state as a whole, as well as transitions between such states. These requirements we call “bootstrap standards”.

What is it that drives methodological change according to a bootstrap theory? Well, it may be different kinds of input which motivates scientists to revise their standards. The impetus may come from empirical information about the track record of some method which constitutes evidence that it is unreliable, or it may be new logical or philosophical arguments, or a perception of disequilibrium within a methodological state.

Versions of bootstrap theories of rationality have been proposed earlier by Briskman and Laudan. An early bootstrap theory of methodological change was proposed by Briskman already in 1977. His main idea is that in research certain kinds of problems arise ("problems of preference" and "problems of goal-pursuit") which cannot be solved by using existing methods or standards. A methodological change is rational to the extent that it solves such problems. The problems encountered function as standards for evaluating methodological changes.

Laudan has also proposed a bootstrap theory.[2] The central idea is that standards are to be seen as means for achieving scientific goals, and that standards and methodological changes can be evaluated in terms of how efficient they are as means for achieving the goals of scientific research.

In his dissertation Nilsson (2000) argued that previous bootstrap theories failed to account for the details of the bootstrap processes where standards are changed, and in a later paper (Nilsson 2005) he proposed a general bootstrap theory. In distinction from previous theories it is explicitly formulated in terms of how methods or standards operative at one scientific stage can be used to evaluate methodological change at that stage. The theory contains a number of bootstrap standards, which are held to govern rational changes of method. To illustrate the contents of such a theory, here is a tentative list of informal bootstrap standards Nilsson proposed (Nilsson 2005).

> Suppose an agent or group of agents accept a set of standards S_1 and revise some of these so that they come to accept a new methodological state S_2. For the transition from S_1 to S_2 to be rational, the following requirements should be met:
>
> *Conservatism*: It is rational to revise a methodological state S_1 only if there is some reason to regard S_1, or some part of S_1, as problematic.
>
> *Internal Conformance*: The standards used to evaluate S_1 or part of S_1 as problematic must themselves be part of S_1 (they must be standards accepted by the agent).
>
> *Problem Solving*: The particular problem identified in S_1 must be absent from S_2.
>
> *Stability*: S_2 must be better than S_1 according to the standards in S_2.
>
> *Prospective Acceptability*: S_2 must be better than S_1, according to the standards that are members of S_1, except for those standards in S_1 that are criticized and revised.
>
> *Goal-pursuit:* A change from S_1 to S_2 must not be such that it is judged to become more difficult – according to the standards in S_2 and those standards in S_1 that are not being criticized and revised – to achieve the scientific goals operative at that point in time.

The bootstrap theory presented in Nilsson (2005) simplifies matters in an important respect: it treats the standards accepted by an agent – a methodological state – as a pure set and specifies how one part of that set can be used for evaluating

[2] In Laudan (1984) it is called "the reticulated model of scientific rationality" whereas in Laudan (1996) it is called "normative naturalism".

another subset of standards. It does not take account of the different kinds of relations that hold between the different standards, thus treating methodological states as unstructured.

The further development of the theory would consist partly in describing these relations and formulating bootstrap standards, which prescribe how such relations are relevant to the rationality of methodological changes. For this purpose, constructing models of sets of standards or methods is likely to be fruitful as it may make it easier to discern and investigate patterns of relations holding between standards.

Should a bootstrap theory be formulated in such a way that the bootstrap standards belong to a metalevel which is separated from the object level of other standards? Philosophically it seems natural instead to formulate a bootstrap theory as a *one-level* theory. That would mean that the bootstrap standards themselves function on the object level, within the methodological states themselves.

When it comes to the question of how theories of methodological change should be formally represented, two questions arise in particular for bootstrap theories: Are there problems of formally representing bootstrap standards, over and above problems with representing other standards that can be applied to methodological change? And, are there obstacles to formally representing a bootstrap theory as a one-level theory?

We hope that bringing mathematical and other formal tools to bear on methodological states will make it easier to uncover and theorize about interesting structural features of such states. The mathematical models in question may be constructed along the lines of the BDI-model of rational agency.

15.5 The BDI-Model of Rational Agency

In this part we will discuss the possibility of studying methodological change from a formal or logical point of view. We start out by briefly describing some of the work that has been done in philosophy and artificial intelligence (AI) concerning the architecture of rational agents; and the cognitive dynamics of such agents. Much of the work has of course been concerned with the logic of belief change (belief revision and belief update), but researchers in AI have also created models of the dynamics of rational agents with goals, intentions, plans, etc. and the ability to act. The development of such agents is governed by very general laws of practical reasoning, roughly: If an agent has certain beliefs and certain goals, then he chooses some available course of action that he believes will favour his goals. A rational agent modifies his beliefs about the world on the basis of the information he receives. And he modifies his immediate goals (intentions) accordingly as his beliefs change. Thus AI researchers have not only studied rational belief change but also rational changes in goals, intentions and plans.

Here we want to discuss the possibility of adapting and extending the kind of models of rational attitude change developed within philosophy and AI to the modelling of rational change within science. The basic idea is to view a scientific research community as an agent with beliefs, goals, procedures, etc. We are not going to consider the interaction and communication between the members of such a community. In reality a research community may be far from homogeneous; there may be differences in opinions and goals between its members and it may be of great interest to study the dynamics within such groups. It is presumably also of great interest to study how different research communities with quite different research programmes may communicate and influence each other. Here, we will make, the no doubt, severe idealization that research communities can be treated as single agents that do not interact with other research communities.

Another question that we do not discuss is the one concerning the principles of individuation of agents in general, and research communities (or research traditions) in particular. In our special case, when is it correct to say that a community observed at time t is the very same research community as one that we observe at a later time $t^{'}$? Presumably there has to be a continuous development tying the two stages together in order to say that they are stages in the development of one research community (or belong to the same tradition). A question that may be even more fundamental is also ignored: what kinds of entities can be rational agents? Within AI the conception of a rational agent seems to be quite liberal: humans, robots, even entities living in "virtual reality" are described as being rational. Philosophers are usually more restrictive. We are only assuming that collectives of humans, in particular societies of researchers may be described as having beliefs, goals and plans, and being rational or irrational.

The BDI-model is an architecture for constructing software for intelligent machines inspired by the belief-desire-intention theory of human practical reasoning developed by Michael Bratman (Bratman 1987, 1999). According to this model an agent has at a given time a set $\boldsymbol{B}$ of *beliefs* and a set $\boldsymbol{G}$ of *goals* (or desires). The agent's beliefs correspond to information that she has about the world. We assume that the belief set $\boldsymbol{B}$ is a consistent set of propositions. $\boldsymbol{G}$ is a set of propositions representing states of affairs that the agent would like to see realized. We do not assume that $\boldsymbol{G}$ is consistent: the agent may very well have contradictory or opposing goals that cannot be realized simultaneously. However, there is at any given time a subset $\boldsymbol{I}$ of the agent's goals that she is committed to realizing. These are the agent's *intentions*. The set $\boldsymbol{I}$ is assumed to be consistent. At any time the agent's intentions are determined by her beliefs and goals at that time. The agent's intentions at any given time are the goals that are operational at that time in determining her actions. A natural assumption is that an agent gives up an intention only if she either believes that the intention has been achieved or that it cannot be achieved (with too much effort).

According to the BDI-model, the dynamics of a rational agent may be described as follows: Initially, the agent is in a certain mental state with beliefs $\boldsymbol{B}$, long term goals $\boldsymbol{G}$, intentions $\boldsymbol{I}$, and an active plan $\boldsymbol{P}$ for realizing her current intentions. Then the agent receives some new information or goes through some process of reasoning resulting in a new belief state $\boldsymbol{B'}$**.** The change in beliefs in turn leads the agent to reconsider her intentions. She then devises a plan $\boldsymbol{P'}$ for realizing the new intentions in light of her new beliefs, and so on.

15.6 Models of Rational Methodological Change

We may think of a scientific research program along the lines of the BDI-model. The agent is now a scientific research community:

Agent: A research community
Beliefs: A scientific corpus consisting of a theory, auxiliary hypotheses, data.
Goals: True explanatory theories, verisimilitude, empirical adequacy etc.
Intention: To test a certain hypothesis (research agenda)
Plan: To perform a series of experiments according to a well-established methodology.
Action: The tests are performed and the results are evaluated.

The results of the tests may then lead to changes in the corpus as well as in the research agenda and the methodological rules. Certain long-term goals may be constitutive of the scientific endeavour. Moreover certain structural (or logical) features, like the general BDI-model may also be characteristic of science. Perhaps one can speak a little vaguely of a logic of scientific reasoning, perhaps open to refinement and revision. However, in accordance with the bootstrap theory of rational scientific change, there are no theories, goals or methods of science that are beyond rational criticism.

15.7 Concluding Discussion

So what does it mean that a scientific agent accepts certain standards of rationality and what kind of entities are these standards? One idea that needs to be pursued is that accepting a standard of rationality is a mental state (a propositional attitude), namely a certain kind of belief about what we rationally-ought-to believe or do. Hence, on this view, rationality standards are *requirements* of rationality in the sense of Broome (2007). If we prefer to speak in terms of what we rationally-ought-to-do or rationally-ought-to-believe instead, rationality standards are beliefs about what we under the circumstances rationally-ought-to-do or rationally-ought-to-believe. For example, we may believe that rationality requires of us that our beliefs are logically consistent, or we may believe that rationality requires of us that we believe the (obvious) consequences of what we believe, or intend the necessary means for achieving our goals. If so, then these requirements are among the standards of rationality that we accept. Our rationality standards may, of course, also include beliefs about how we rationally-ought-to change our beliefs when we receive new information.

As has been pointed out by Broome and others, a belief that we rationally-ought-to *F*, need not be normative in the strong sense of entailing a belief that we, everything considered, ought to *F*. Rationality (as we conceive of it) may require of us that we *F*, although it is not the case that, everything considered, we ought to *F*. If there are objectively correct standards of rationality, then we may also be mistaken about what rationality requires of us.

If rationality standards are viewed as beliefs about what rationality requires of us, then scientists may deviate from their standards in their actual practice of science. It is natural to think that one function of our standards is precisely to enable us to criticize and correct our scientific practice. On the other hand, it appears that the direction of criticism could under some circumstances be reversed: if a certain scientific practice which we judge to be generally successful fails to meet our rationality standards, then at least prima facie that might constitute a case for considering the rationality standards themselves to be problematic.

Now, if standards of rationality are – or can be viewed as – beliefs of a certain kind, then the theory of methodological change becomes a special branch of a generalized theory of belief change. The formal methods of belief revision theory can then also be applied to methodological change. However, the standard AGM-axioms of belief revision (cf. Alchourrón et al. 1985) are not applicable without restriction to methodological change. For example, AGM-revision satisfies *Preservation:*

$$\text{If } A \text{ is consistent with the theory } \boldsymbol{T}, \text{ then } \boldsymbol{T} \subseteq \mathbf{T} * \mathrm{A},$$

where $\boldsymbol{T}{*}\mathrm{A}$ is the revision of the theory $\boldsymbol{T}$ with the statement A. However, let A be the statement "One ought to look out for dodos". Someone who does not know whether or not there are any dodos around may accept A, although he would give up the belief in A once he learned that the dodo is extinct. Hence, in the presence of deontic beliefs Preservation has to be abandoned.

In an extension of the BDI-model, which includes an agent's methodological states, the agent is situated in an environment (the external world). The agent has a total (internal) state consisting of (at least) the following components: A theoretical state $\boldsymbol{T}$ (the agent's current scientific theory about the world), a goal state $\boldsymbol{G}$, certain intentions $\boldsymbol{I}$ for action, and a methodologicial state $\boldsymbol{M}$. Moreover, there is for each total state S a preference relation $\leq_S$ over total states. $S_1 \leq_S S_2$ means that the state S_2 better satisfies the goals and standards that hold in S than does S_1. S is in all likelihood not optimal from its own perspective. This fact will move the agent to a state S' that is better than S from the perspective of the current state S. The question arises: Can we formulate any informative constraints on this process?

We can distinguish at least four different kinds of change:

(i) Changes in scientific theory in response to new research results.
(ii) Changes in actual scientific practice in order to make this practice conform better to current rationality standards (our beliefs about correct methodology).
(iii) Changes of current rationality standards as a result of a critical discussion of their appropriateness.
(iv) Changes of basic scientific goals and values.

It is changes of types (iii) and (iv) that primarily interest us here. Generally, inquiring agents will prefer to change their theories about the world rather than their rationality standards and their basic scientific goals. Under what circumstances is it instead rational to change, e.g., one's rationality standards rather than one's theories or one's actual practice?

The bootstrap theory discussed above is one attempt to answer this question, by proposing constraints on how different methodological states in a sequence of changes should be related to each other if the process is to be one of rational methodological change. One challenge is then to develop a suitable formal framework, with a language which allows one to represent rationality standards (including meta-standards such as the bootstrap standards) as well as theories, goals, intentions, plans and cognitive actions. A related challenge is to extend the BDI-model of rational agency in such a way that it also covers those rare but interesting occasions when inquiring agents come to the conclusion that what rationality requires of them is to reevaluate their beliefs about what rationality amounts to.

Acknowledgements The authors gratefully acknowledge financial support from the Swedish Research Council (VR), the Swedish Collegium for Advanced Study (SCAS) and Umeå University.

References

Alchourrón, C., P. Gärdenfors, and D. Makinson. 1985. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic* 50:510–530.

Bratman, M. 1987. *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Bratman, M. 1999. *Faces of intention: Selected essays on intention and agency*. Cambridge: Cambridge University Press.

Briskman, L. 1977. Historicist relativism and bootstrap rationality. *The Monist* 60:509–539.

Broome, J. 2007. Requirements. In *Hommage à Wlodek: Philosophical papers dedicated to Wlodek Rabinowicz*, eds. Toni Rønnow-Rasmussen, Björn Petersson, Jonas Josefsson, and Dan Egonsson. http//www.fil.lu.se/hommageawlodek/site/papper/BroomeJohn.pdf Lund: Lund University.

Laudan, L. 1984. *Science and values: The aims of science and their role in scientific debate*. Berkeley, CA: University of California Press.

Laudan, L. 1996. *Beyond positivism and relativism: Theory, method, and evidence*. Oxford: Westview.

Newton-Smith, W. 1981. *The rationality of science*. London: Routledge and Kegan Paul.

Niiniluoto, I. 1999. *Critical scientific realism*. Oxford: Oxford University Press.

Nilsson, J. 2000. *Rationality in inquiry: On the revisability of cognitive standards*. Ph.D. dissertation, Umeå University.

Nilsson, J. 2005. A bootstrap theory of rationality. *Theoria* 71:182–199.

Popper, K.R. 1989. *Conjectures and refutations: The growth of scientific knowledge*, 5th revised and corrected ed. London: Routledge.

Shapere, D. 1984. Reason and the search for knowledge: Investigations in the philosophy of science. Dordrecht: Reidel.

Worrall, J. 1982. Broken bootstraps. *Erkenntnis* 18:105–130.

Worrall, J. 1988. The value of a fixed methodology. *British Journal for the Philosophy of Science* 39:263–275.

Worrall, J. 1989. Fix it and be damned: A reply to Laudan. *British Journal for the Philosophy of Science* 40:376–388.

Index

E.J. Olsson, S. Enqvist (eds.), *Belief Revision Meets Philosophy of Science*, Logic, Epistemology, and the Unity of Science 21, DOI 10.1007/978-90-481-9609-8,