7

BELIEF REVISION THEORY

Hanti Lin

We often revise beliefs in response to new information. But which ways of revising beliefs are "OK" and which are not? A belief revision theory is meant to provide a general answer, with a sense of "OK" that it specifies. This article is an introduction to belief revision theory and its foundations, with a focus on some issues that have not received sufficient attention. First we will see what belief revision theories are, and examine their possible *normative or evaluative* interpretations. Second we will compare the standard belief theory called AGM with its alternatives, especially the alternatives that are motivated by nonmonotonic logic and formal learning theory. Third we will discuss counterexamples to some belief revision theories, and categorize how we might explain those counterexamples away. Fourth and finally we will examine a variety of motivated formal techniques for constructing belief revision theories, and discuss how those motivations might be transformed into explicit arguments.

1 INTRODUCTION

We often revise beliefs in response to new information. But which ways of revising beliefs are "OK" and which are not? A belief revision theory is meant to provide a general answer, with a sense of "OK" that it specifies.

This article is an introduction to some belief revision theories and their foundations. We will see what belief revision theories are, or could possibly be, *as normative or evaluative theories*, and discuss why most belief revision theories in the literature tend to claim to be only about idealized, perfect rationality (Section 2). We will survey a variety of motivated, formal techniques for constructing belief revision theories, and see how to use these techniques to construct the standard theory called AGM and its dissenters (Section 3–Section 4). We will discuss how we might argue against a belief revision theory (Section 5), and how we might argue for it (Section 6).

Articles surveying belief revision theories have been available, such as the excellent ones by Hansson (2017), Rodrigues, Gabbay, and Russo (2011), and Huber (2013a, 2013b). To help the reader make the most of the survey articles available, including the present one, let me explain what my emphases will be.

- Earlier surveys tend to focus on a particular normative or evaluative interpretation of formal theories of belief revision, taking those theories to say something about idealized, perfect rationality. This is the dominant interpretation in the literature. Other possible interpretations will be explored here as well. In fact, the choice among possible interpretations ultimately concerns the choice among very different research programs in belief revision theory—or so I will argue in Section 2.3.

- Earlier surveys tend to focus on the standard, AGM theory of belief revision, together with its add-ons and improvements. But I wish to spend more time on dissenters from the AGM theory. In Section 4.4, I will present belief revision theories that disagree with the content of the AGM theory in *permitting something that the AGM theory prohibits*. (These theories usually come from so-called *nonmonotonic logic*.) In Section 4.6, I will present belief revision theories that disagree with the spirit of the AGM theory in *taking the ultimate concern to be finding the truth* rather than conforming to what intuition says about rationality. (These theories usually come from so-called *formal learning theory*.)

- The use of intuitive counterexamples is important when we argue against a belief revision theory, and earlier surveys do cover that. But I will make a first step toward categorizing how counterexamples might be explained away. The reason is that the dialectic exchange between alleging-counterexamples and explaining-them-away turns out to raise very interesting issues about the goal and nature of belief revision theory. This will be the highlight of Section 5.

- Earlier surveys tend to focus on various *motivated* techniques for constructing theories of belief revision. But I will explore how those motivations could be reconstructed into *explicit arguments* for the intended normative claims. This will help us identify and formulate issues of utmost importance to the very foundations of belief revision theory—or so I will argue in Section 6.

Achieving these goals means that I will have to set aside, or just mention in passing, many other interesting topics in belief revision theory. But this is exactly why we need multiple survey articles to complement one another.

One last point of clarification before we get started, regarding the kind of belief that will concern us in this article. Compare the following examples.

(i) Ann is 95% confident that it will rain tomorrow.

(ii) Ann believes that it will rain tomorrow.

Sentence (i) attributes to Ann a *quantitative* doxastic attitude toward a certain proposition, called a *credence*. There are infinitely many such quantitative attitudes that she could have had toward that proposition. She could have had, say, credence 50%, 50.1%, or 50.17% in that proposition. By contrast, sentence (ii) attributes to Ann a *qualitative* doxastic attitude toward a certain proposition, call a *belief*. There are two qualitative doxastic attitudes she could have had toward that proposition: believing it, or not believing it.[1] The subject matter of this article is concerned with revision of beliefs (qualitative doxastic attitudes). For revision of credences (quantitative doxastic attitudes), please see the chapter "Precise Credences" (Titelbaum, this volume).[2]

2 BELIEF REVISION THEORIES AS NORMATIVE THEORIES

I mentioned earlier that a belief revision theory is, roughly, a theory saying which ways of belief revision are OK and which are not, which I am going to explain in greater detail in this section.

2.1 *What a Belief Revision Theory Is Like*

Consider the following constraint on an agent at a time.

> PRESERVATION. If the information that agent A receives at time t is compatible with the set of the beliefs that A has right before t, then, right after t, agent A retains all of her beliefs in response to that information.

(By "the" information one receives at t, I mean the conjunction of *all* pieces of information that one receives at t.)[3] This constraint on belief revision is *formal* in the sense that it concerns the logical properties of beliefs rather than their particular contents. Due to its formal nature, Preservation usually receives the following reformulation:

> PRESERVATION. If ϕ is compatible with B, then B is a subset of $B * \phi$, where:

[1] If you wish, you can count one more attitude: disbelieving a proposition. It is debatable whether disbelieving P can be reduced to believing $\neg P$.

[2] This raises an issue: how should the revision of beliefs and the revision of credences be related? For the first few works that address this issue, see Arló-Costa and Pedersen (2012), Lin and Kelly (2012), and Leitgeb (2014). Also see the chapter "Full and Partial Belief" (Genin, this volume).

[3] What if one receives no piece of information at t? What is the conjunction of the empty set of propositions? Answer: it is a tautology. Think of the conjunction of a set S of propositions to be the weakest proposition that entails every proposition in S—or, in terms of algebraic logic, define the conjunction of S as the the greatest lower bound of S in the lattice of propositions under discussion.

- ◇ B is the set of one's beliefs right before the receipt of new information,
- ◇ ϕ is the new information one receives,
- ◇ $B * \phi$ is the set of one's new beliefs in response to new information ϕ.

Preservation offers just one possible constraint on belief revision, and we will discuss more constraints below.

Preservation as just formulated is a mere constraint, a condition that one may turn out to satisfy or violate at a time; there is nothing normative or evaluative in itself. But when a belief revision theory contains Preservation, it is typically understood to make the following normative claim:[4]

> Preservation Thesis (the "Perfect Rationality" Version). One is perfectly rational only if one has never violated, and would never violate, Preservation.

Once a normative thesis is put on the table, a philosopher's first reaction would be to explore potential counterexamples (whether or not she wants to confirm or refute the thesis). Here is one.

> Example (Three Composers).[5] The agent initially believes the following about the three composers Verdi, Bizet, and Satie.

[4] We may want to clearly distinguish what is normative (such as 'ought') from what is evaluative (such as 'good', 'rational', and 'justified'). But this distinction is irrelevant to the purposes of this article. Understand my use of 'normative' to be a shorthand for 'normative or evaluative'.

[5] This scenario is adapted from an example due to Stalnaker (1994). Stalnaker uses it to argue against a different constraint on rational belief revision.

> Rational Monotonicity. If ψ is compatible with $B * \phi$, then $B * \phi \subseteq B * (\phi \wedge \psi)$.

Stalnaker considers two alternative possibilities: the agent could receive E or $E \wedge E'$ as the information at a certain time. And then Stalnaker asks how the agent should set up a belief revision strategy as a contingency plan to deal with these two possibilities. Substituting E and E' for the ϕ and ψ in Rationality Monotonicity, Stalnaker obtains his counterexample to it. That is what Stalnaker does, which appears to be different from what we are doing here about Preservation, for two reasons. First, Preservation is a proper consequence of Rational Monotonicity under the weak assumption that $B * \top = B$, where $\top$ is a tautology. Second, Stalnaker's own example lacks an essential feature of our scenario here: the agent receives two pieces of information, E and E', successively. Indeed, it is the *second* revision, prompted by the later information E', that is alleged to violate Preservation. That is, in terms of the $(*)$-notation, it is the revision of the second belief set $B * E$ into the third belief set $(B * E) * E'$ that is alleged to violate Preservation. That said, it should not be surprising that Stalnaker's case against Rational Monotonicity can be easily modified into a case against Preservation, thanks to the formal resemblance between these two constraints on belief revision. In case you are interested, here is a bit more history about the Composers case: Stalnaker's own example is a variation on an example due to Ginsberg (1986), which is in turn a variation on an example due to Quine (1982). Both Ginsberg and Quine use their examples to talk about counterfactuals rather than belief revision.

(A) Verdi is Italian;

(B) Bizet is French;

(C) Satie is French.

Then the agent receives this information.

(E) Verdi and Bizet are compatriots.

So the agent drops her beliefs in A and in B, and retains the belief in C that Satie is French (after all, information E has nothing to do with Satie). Of course, she comes to believe the new information E that Verdi and Bizet are compatriots, while suspecting that Verdi and Bizet might both be Italian, and that they might both be French. So, at this stage, the agent does not rule out the possibility that Verdi is French (and, hence, a compatriot of Satie). So what she believes at this stage is compatible with the following proposition.

(E') Verdi and Satie are compatriots.

But then she receives a second piece of information, which turns out to be E'. Considering that she started with initial beliefs A, B, and C and received information E and E', which jointly say that the three composers are compatriots, now she drops her belief in C.

Let us focus on this agent's second revision of beliefs, prompted by information E'. Information E' is compatible with what she believes right before receiving this information, and she drops her belief in C nonetheless. So this agent's second revision of beliefs violates Preservation. But there seems nothing in the specification of the scenario that prevents the agent from being perfectly rational. So this seems to be a counterexample to the Preservation Thesis.

This cannot be the end of the dialectic, of course. We want to think about whether one may save the Preservation Thesis by explaining away the alleged counterexample—an issue that we will revisit in Section 5. This is just to give a taste of what it is like to work in belief revision theory.

2.2 *What Normative Interpretations Could Be Intended?*

The Preservation Thesis is only one of the many normative theses that we can formulate in terms of Preservation. Here is a sample:

(T_1) An agent is rational at a time only if she does not violate Preservation at that time.

(T_2) An idealized agent is perfectly rational only if she has never violated and would never violate Preservation.

(T_3) A strategy for belief revision is rational only if every possible revision licensed by it does not violate Preservation.

(T_4) An agent is rational at a time only if, other things being equal, she does not violate Preservation at that time.

(T_5) Other things being equal, an agent should not violate Preservation.

A belief revision theory is meant to affirm or deny some theses like these.

This list is by no means exhaustive. There are at least two dimensions along which we can generate more theses for a belief revision theory to affirm or deny (or be silent about).

As to dimension one: note that Preservation is only one of the many possible constraints on belief revision. So, in theses T_1–T_5, we can easily replace Preservation by a distinct constraint on belief revision.

As to dimension two: note that theses T_1–T_5 are formulated in terms of 'ought' or 'rational'. So, if there are multiple senses of 'ought', then the above ought-thesis will have to be multiplied. Similarly, if epistemic rationality is not identical to, or is only a special kind of, instrumental rationality, then the above rationality-theses will have to be duplicated. One more example: we might be interested in not only whether one's revision is rational, but also whether it is justified. So, for example, we can consider the thesis that an agent is *justified* in revising her beliefs the way she does only if her revision does not violate Preservation.

So, given a constraint on belief revision (such as Preservation), we can formulate various normative theses in terms of that constraint. A belief revision theory is meant to affirm or deny some such theses.

2.3 *Which Normative Interpretation Is to Be Intended?*

Most belief revision theories in the literature are usually understood to make claims only about idealized rationality, e.g. affirming or denying theses of the form T_2. But why?

Here is a potential reason. Many belief revision theories assume that the agent's belief set B is closed under deduction, so those theories can be interpreted as talking about a logically omniscient agent, who believes every logical consequence of what she believes. So those theories *can* be interpreted as talking about a kind of perfect rationality that only a logically omniscient agent can have. But this is not a good reason for *restricting* the interpretation to idealized perfect rationality. For, following Levi (1983), a deductively closed set B of sentences *can also* be used to express the commitments of an ordinary, non-idealized agent's beliefs. Under this alternative interpretation, revision of B is revision of the commitments of one's beliefs.

As it turns out, the decision to focus on certain kinds of normative interpretations rather than some others actually involves a difficult choice among research programs in belief revision theory—or so I shall argue in the following.

As a preliminary step, let me argue that T_1 should not be an intended normative content of a belief revision theory, because T_1 has a quite obvious counterexample.

> Example (One's Embarrassing Past). Suppose that propositions A, B, C are logically independent, in the sense that all the 8 $(= 2^3)$ combinations of their truth values are logically possible. An agent started by believing A without commitment to the truth or falsity of B or C. Then she received information B and, in response, she somehow dropped her old belief in A and came to believe $\neg A \wedge B$, without commitment to the truth or falsity of C. So she violated Preservation at that time. Since then she has retained those beliefs and has not received any new information. Remembering all these in her embarrassing past, now she receives new information C. She is wondering what to believe.

What is she supposed to do in order to be a rational agent *now*? Since the new information C is compatible with what she believed just now, to satisfy Preservation *now* the agent has to continue to believe $\neg A \wedge B$. But, if Preservation really represents such a good standard to abide by, the rational thing for her to do *now* is to retract her belief in $\neg A \wedge B$ and come to believe A, B, and C instead—as if she had never violated Preservation. So T_1 should be rejected *even* by those who are sympathetic to Preservation as a requirement of rationality.

It is not just that T_1 is false. When we replace the Preservation constraint in T_1 by any other formal constraint ever studied in the belief revision literature, the resulting thesis—a *formal variant* of T_1—is also false. The reason is that the constraints studied in belief revision theory are formal, having nothing to do with the contents of one's beliefs and hence making no reference to one's beliefs about one's revision history. So the case of One's Embarrassing Past can be suitably adapted to refute every formal variant of T_1. Lesson: every belief revision theory in the literature, *when interpreted to make claims of the form* T_1, is false.

If we are sympathetic to Preservation as a good standard to abide by, there are two possible ways out.

> Strategy 1 (Get Hands Dirty Today). Fix thesis T_1 by weakening Preservation in such a way that avoids the above counterexample while retaining the spirit of Preservation.

> Strategy 2 (Pay off the Debt in the Future). Deny T_1 but affirm T_2, T_3, T_4, T_5, or their variants. Namely, redirect our attention, at

> least for the moment, to idealized rationality, or the rationality of strategies instead of agents, or *ceteris paribus* norms. But keep in mind that this incurs a debt: we will, at some point, need to say how the truth of theses like T_2–T_5 can be employed to shed light on the rationality of a non-idealized agent's belief revision without a *ceteris paribus* clause.

These two possible ways out correspond to very different projects one may pursue in belief revision theory. Let me illustrate.

Here is what it is like to pursue Strategy 1 (Get Hands Dirty Today). Consider the following weakening of Preservation.

> Preservation*. If (i) the new information one receives at *t* is compatible with the set of beliefs that one has just before *t* and (ii) one does not believe at *t* that one has violated Preservation before, then, right after *t*, one retains all of one's beliefs in response to the new information.

This constraint is non-formal (i.e. referring to contents of one's beliefs), and it weakens Preservation by adding (ii) to the antecedent. Now formulate the following non-formal variant of T_1.

> (T_1^*) An agent is rational at a time only if she does not violate Preservation* at that time.

This thesis is logically weaker than T_1, weak enough to escape the case of One's Embarrassing Past. For the agent violates antecedent (ii) and, hence, satisfies Preservation* vacuously. The problem with this weakened Preservation* is that it is too weak for those who want to save the spirit of Preservation as a constraint on rational belief revision. Do you think that you violated Preservation at least once in the past? I think I did, although I cannot tell when exactly. Most people, if asked, would say that they violated Preservation at least once in the past, too. So most people satisfy Preservation* vacuously by violating antecedent (ii). Lesson: if we think that the spirit of Preservation is on the right track toward a nontrivial constraint on rational belief revision, we need to weaken Preservation by adding an appropriate antecedent that hits the "sweet spot," making the reformulated Preservation weak enough to avoid potential counterexamples and substantial enough to guide our belief revision. Hitting such a sweet spot might require careful addition of complicated clauses into Preservation, making our hands dirty now.

It is possible to keep our hands clean at least for the moment. If Preservation really represents such a good standard to abide by, then it seems pretty safe to affirm thesis T_2. For, in response to One's Embarrassing Past, we can simply judge that the agent in question simply fails to be perfectly rational due to her embarrassing past, no matter how she is going revise her beliefs at the present time. So, to keep our hands clean, we can develop a belief revision theory that only makes claims about idealized, perfect rationality, such as T_2. But this only makes our hands clean *for the time being*, for it actually incurs a debt that we will have to pay off later. There is nothing wrong in developing a theory of perfect rationality for idealized agents. But we want such a theory to shed light on a theory of rational belief revision for ordinary agents like us. What's the light to be shed? To answer this question is to pay off the debt.

Similarly, if Preservation really represents such a good standard to abide by, it seem pretty safe to affirm thesis T_3. For, in response to One's Embarrassing Past, we can say that the revision strategy that the agent has been following through time is simply irrational. But then one day we will have to pay off the debt: we will have to explain how a theory of strategic rationality sheds light on a theory of agential rationality. Similarly, adoption of T_4 or T_5 incurs its own debt: we will have to say how *ceteris paribus* norms would apply to concrete cases, which would require us to develop, for example, a logic for defeasible deontic reasoning.[6] So what confronts us is this problem:

> Choosing Among Research Programs. Should we get our hands dirty today, or should we incur a debt today and promise to pay it off in the future, by directing our attention to perfect rationality, strategic rationality, or *ceteris paribus* rationality?

The literature, as developed today, seems more inclined to opt for the route of perfect rationality.

In the rest of this article, we will follow the literature, talking about theses of the form T_2 most of the time. Just keep in mind that a research program has been chosen (at least tentatively) and it comes with a debt.

3 FORMAL THEORIES OF BELIEF REVISION

A typical belief revision theory has two parts: the *formal* part is meant to formulate certain formal constraints on belief revision, and the *normative* part is meant to make some normative claims in terms of those constraints. It is time to turn to the formal part.

Consider a language $\mathcal{L}$, identified with a set of sentences closed under at least the standard Boolean operations (i.e., 'and', 'or', and 'not'). A finite sequence $(\phi_1, \phi_2, \ldots, \phi_n)$ of sentences in $\mathcal{L}$ can be understood as a *history of inquiry* in which one receives information ϕ_1, then receives information ϕ_2, $\ldots$, and then receives information ϕ_n. A belief revision strategy is meant

[6] See Nute (2012) for a number of approaches to defeasible deontic logic.

to tell one how to change beliefs given any relevant history of inquiry. Accordingly, we make the following definition.

> Definition (Belief Revision Strategy). A *belief revision strategy* over language $\mathcal{L}$ is a function $S : \mathcal{I} \to \wp(\mathcal{L})$, where:
>
> ◇ $\mathcal{I}$ is a nonempty set of finite sequences of sentences in $\mathcal{L}$ that is *closed under subsequences*—that is, whenever $\mathcal{I}$ contains a nonempty sequence $(\ldots, \phi_n)$, it also contains the truncated sequence $(\ldots)$ that results from deleting the last entry. So the empty sequence, denoted by $(\,)$, is guaranteed by definition to be in $\mathcal{I}$. Call $\mathcal{I}$ an information space, meant to contain all the "relevant" histories of inquiry in question.
>
> ◇ $\wp(\mathcal{L})$ is the collection of all subsets of $\mathcal{L}$, i.e. all sets of sentences in $\mathcal{L}$.
>
> ◇ $S(\phi_1, \phi_2, \ldots, \phi_n)$ is understood as the set of beliefs that strategy S would recommend for an agent at the end of inquiry history $(\phi_1, \phi_2, \ldots, \phi_n)$. In the limiting case, the value of function S at the empty sequence $(\,)$, written $S(\,)$, denotes the set of beliefs recommended at the beginning of the inquiry.

I have to confess that the S-notation used here is not quite standard in the literature. But in this article we will encounter three different kinds of belief revision theories, and the S-notation is the simplest one for unifying all the three.

A formal theory of belief revision, no matter how it is presented, works by imposing a constraint on belief revision strategies, allowing for some strategies and ruling out the others. Accordingly, we make the following definition.

> Definition (Formal Theory of Belief Revision). A *formal theory of belief revision* over language $\mathcal{L}$ is (or can be identified with) a set of belief revision strategies over $\mathcal{L}$.

A formal belief revision theory T can be turned into a normative theory once it is given a normative interpretation, such as: "an agent is perfectly rational only if there exists a belief revision strategy in T that she has been following and would continue to follow." (Just a reminder: alternative interpretations have been discussed in Section 2.2.)

3.1 *Simple Belief Revision Theories*

Let $\mathcal{I}_{\leq 1}$ be the set of all sequences of sentences in $\mathcal{L}$ with lengths ≤ 1. So it does not consider successive revisions of belief. A belief revision strategy is *simple* iff it is defined on $\mathcal{I}_{\leq 1}$. A set of such strategies is called a *simple* formal theory of belief revision.

Suppose that we only care about simple belief revision for the moment. Then the S-notation just introduced is an overkill, and it would be more convenient to work with the notation of B and $*$ introduced earlier. Here is the translation between these two notations:

$S(\,) = B$, the initial set of beliefs;

$S(\phi) = B * \phi$, the set of new beliefs in light of new information ϕ.

So Preservation can be reformulated as follows.

> Preservation. For any ϕ compatible with $S(\,)$, $S(\,) \subseteq S(\phi)$. In other words, for any ϕ compatible with B, $B \subseteq B * \phi$.

The set of simple belief revision strategies that satisfy Preservation is a formal theory of belief revision. It corresponds to a strictly weaker constraint than the standard, AGM belief revision theory, as we will see in Section 4.1.

3.2 *Iterated Belief Revision Theories*

The information space $\mathcal{I}_{\leq 1}$ just considered is very small. What about working with a larger information space? Let $\mathcal{I}_{\text{finite}}$ be the set of all finite sequences of sentences in $\mathcal{L}$. A belief revision strategy S defined on $\mathcal{I}_{\text{finite}}$ says a lot. It says how to revise beliefs when one receives information ϕ_{n+1} that follows inquiry history $(\phi_1, \ldots, \phi_n)$: just change the set of beliefs from $S(\phi_1, \ldots, \phi_n)$ to $S(\phi_1, \ldots, \phi_n, \phi_{n+1})$. It even says how to revise beliefs when one receives information ϕ but then, unfortunately, receives information $\neg\phi$: change the set of beliefs from $S(\ldots, \phi)$ to $S(\ldots, \phi, \neg\phi)$. A set of belief revision strategies defined on $\mathcal{I}_{\text{finite}}$ is called an *iterated* belief revision theory.

For example, consider the set of all belief revision strategies $S : \mathcal{I}_{\text{finite}} \to \wp(\mathcal{L})$ that satisfy the following.

> Iterated Preservation. For any finite sequence $(\phi_1, \ldots, \phi_n)$ of sentences and any sentence ϕ_{n+1} in $\mathcal{L}$, if ϕ_{n+1} is compatible with $S(\phi_1, \ldots, \phi_n)$, then $S(\phi_1, \ldots, \phi_n) \subseteq S(\phi_1, \ldots, \phi_n, \phi_{n+1})$.

This constraint is strictly weaker than many iterated belief revision theories in the literature, as we will see in Section 4.5.

3.3 *Belief Revision Theories for Inductive Inferences*

Sometimes we may want to have an information space $\mathcal{I}$ that is just right, not too big and not too small. Consider an empirical problem: *"Are all* *ravens black?"* Call this the *Raven Problem*. Let language $\mathcal{L}$ contain the following sentences:

$$\begin{aligned} h &= \text{the hypothesis "all ravens are black";} \\ b_i &= \text{"the } i\text{-th observed raven is black";} \\ n_i &= \text{"the } i\text{-th observed raven is non-black."} \end{aligned}$$

An inquiry history relevant to the Raven Problem describes the color of every raven observed in that history. For example, (b_1, b_2, b_3, b_4) says that we have observed four ravens and all of them are black; $(b_1, b_2, b_3, b_4, n_5)$ says that we have observed five ravens with the first four being black and the last one being non-black. Let $\mathcal{I}_{\text{raven}}$ be the set of all finite sequences whose i-th entry is either b_i or n_i. $\mathcal{I}_{\text{raven}}$ is meant to exclude any sequence that contains h, because, let us suppose, scientists never receive h as information. In the present case, the point of working with $\mathcal{I}_{\text{raven}}$ (rather than the much larger information space $\mathcal{I}_{\text{finite}}$) is that we want to be clear about which pieces of information can be *available* to a scientist for solving the Raven Problem. Furthermore, reference to $\mathcal{I}_{\text{raven}}$ is essential when we define how well a belief revision strategy performs as a solution to the Raven Problem, as we will see in Section 4.6.

We might come to believe h when they have observed a certain number of black ravens without a single non-black one. But how many black ravens suffice for a rational or justified belief in h? A belief revision strategy defined on $\mathcal{I}_{\text{raven}}$ is meant to give an answer. For example, a strategy S_{skep} that follows *inductive skepticism* would say that no finite amount of black ravens suffices; that is, $h \notin S_{\text{skep}}(b_1, \ldots, b_n)$ for every positive integer n.

4 HOW TO CONSTRUCT FORMAL THEORIES

In this section we will review a number of techniques for constructing formal theories of belief revision. Those techniques can be taken as mere formal tools for constructing formal theories of belief revision. But those formal techniques are usually associated with some motivations or interpretations, which might do some interesting philosophical work. To anticipate, in Section 6 we will examine how interpreted techniques of theory construction could be turned into explicit arguments for normative claims about belief revision.

4.1 *Axiomatization*

Consider the following axiom system, stated in terms of B and $*$, where $B + \phi$ denotes the set of logical consequences of $B \cup \{\phi\}$:

Axiom System AGM.

(Closure) $B * \phi$ is closed under logical consequences.

(Extensionality) If ϕ and ψ are logically equivalent, then $B * \phi = B * \psi$.

(Success) $B * \phi$ contains ϕ.

(Consistency) If ϕ is consistent, then $B * \phi$ is consistent.

(Accretion) If ϕ is compatible with B, then $B * \phi = B + \phi$.

(Super-Accretion) If ψ is compatible with $B * \phi$, then $B * (\phi \wedge \psi) = (B * \phi) + \psi$.

Note that Accretion implies Preservation. These constraints on B and $*$ can be easily translated to constraints on belief revision strategies S—just recall the translation provided earlier: $B = S(\,)$ and $B * \phi = S(\phi)$. So the AGM axiom system defines a formal theory of simple belief revision, i.e. the set of simple belief revision strategies that satisfy those axioms. The ideas of this belief revision theory can be found in Harper (1975), Harper (1976), and Levi (1978). But this theory is usually called AGM because Alchourrón, Gärdenfors, and Makinson (1985) prove a representation theorem for it, to be presented in the next subsection. The axiomatization provided here is equivalent to the standard—but more complicated—axiomatization found in their 1985 paper.

If you think that the AGM axiom system is too strong and would like to work with a weaker one, the following is an option, where the first four axioms are borrowed from AGM:

Axiom System P^{+}

(Closure)

(Extensionality)

(Success)

(Consistency)

(Cautious Monotonicity) If $\psi \in B * \phi$, then $B * \phi \subseteq B * (\phi \wedge \psi)$.

(Or) If $\psi \in B * \phi_1$ and $\psi \in B * \phi_2$, then $\psi \in B * (\phi_1 \vee \phi_2)$.

I call it P^{+} because this axiom system minus Consistency is, in a sense, equivalent to the well-known system P of nonmonotonic logic.[7] Every axiom in P^{+} can be derived from the AGM axiom system, but the converse

[7] This assumes the standard translation between belief revision theory and nonmonotonic logic (Makinson & Gärdenfors, 1991), which I present in the appendix (Section 8.1).

does not hold. In particular, axiom system P^+ does not imply Accretion because it does not even imply a logically weaker constraint: Preservation (and we will be in a position to prove this claim in Section 4.4).

4.2 *Partial Meet Contraction*

Let us turn to a second technique for constructing a simple belief revision theory. This technique works pretty much by telling a story of a rational agent who is deciding which beliefs to retain or to abandon.

Suppose that an agent's new information ϕ is incompatible with her belief set B. Then, before she adds ϕ into her set of beliefs, it seems a good idea for her to drop some old beliefs, i.e. to remove some sentences from B in order to obtain a (smaller) set that does not entail $\neg\phi$, so that the addition of ϕ would not cause any inconsistency. Denote this set by $B \div \neg\phi$, called the *contracted* set of beliefs free from commitment to $\neg\phi$. Once the agent obtains the contracted belief set $B \div \neg\phi$, she can safely add ϕ to it and close it under logical consequences, and thereby obtain $(B \div \neg\phi) + \phi$ as the new belief set. Namely:

Levi Identity. $B * \phi \; = \; (B \div \neg\phi) + \phi.$

At its core, this amounts to constructing a revision procedure as the concatenation of two other procedures: one for removing beliefs ($\div$) and the other for adding beliefs ($+$). The process from B to $B \div \neg\phi$ is call *contraction*, and the problem is how to find the contracted belief set $B \div \neg\phi$. In typical cases there are multiple candidates for $B \div \neg\phi$ (i.e. multiple subsets of B that do not entail $\neg\phi$). Which one would/could serve as the $B \div \neg\phi$ that the agent needs for the sake of rational belief revision?

That problem has a standard, formal solution, called *partial meet contraction*, which is the focus of this subsection. Let $B \perp \neg\phi$ denote the set of all inclusion-maximal subsets of B that do not entail $\neg\phi$. In other words, $B \perp \neg\phi$ contains X iff X is a set obtained by removing no more sentences from B than necessary—retracting no more old beliefs than necessary—in order to achieve compatibility with new information ϕ. Then, to proceed further, a *prima facie* plausible idea is to (i) select "the best" candidate in $B \perp \neg\phi$ and let it be the contracted belief set. What if there is no uniquely best candidate? Then perhaps the agent may try to (ii) arbitrarily select one of the best candidates in $B \perp \neg\phi$, and let it be the contracted belief set. But what if the agent feels unable to make such an arbitrary selection given multiple best candidates? The standard proposal is to (iii) intersect all of those best candidates and obtain an even smaller set of sentences, to be identified with the contracted belief set $B \div \neg\phi$.

This last idea, (iii), is what underlies so-called partial meet contraction, and can be formally presented as follows.

DEFINITION (SELECTION FUNCTION FOR A BELIEF SET). A *selection function* for B is a function γ such that, for every collection M of subsets of B:

(a) $\gamma(M) \subseteq M$ if $M \neq \varnothing$,

(b) $\gamma(M) \neq \varnothing$ if $M \neq \varnothing$,

(c) $\gamma(\varnothing) = \{B\}$.

The idea is that, for any nonempty collection M of candidates, γ is meant to return $\gamma(M)$ as the set of best candidates in M. Then, for each sentence ϕ, let γ generate $B \div \neg\phi$ as follows:

PARTIAL MEET CONTRACTION. $B \div \neg\phi \;=\; \bigcap \gamma(B \bot \neg\phi)$.

(In case you are interested: while the above formalizes idea (iii), it turns out that idea (i) can be modeled by the special case in which γ returns a singleton.)

In general, given a selection function γ for a belief set B, it defines a contraction operator $\div$ by partial meet contraction, which then defines a revision operator $*$ by Levi identity. Initial belief set B and revision operator $*$ then jointly define a simple belief revision strategy. So a set of selection functions generates a set of simple belief revision strategies, i.e. a simple belief revision theory.

We want to sort out selection functions that are "OK" in order to use them to produce belief revision strategies that are "OK." But which selection functions are "OK"? Imagine that there is a binary relation $\geq$ on subsets of B. Understand $X \geq Y$ as saying that X is at least as "good" as Y with respect to $\geq$ (so presumably we want $\geq$ to be at least transitive and reflexive). Then we can require γ to select the "best" items as follows. For any sentence ϕ (which serves as the new information) such that $B \bot \neg\phi \neq \varnothing$:

$$\gamma(B \bot \neg\phi) = \{X \in B \bot \neg\phi : X \geq Y \text{ for all } Y \in B \bot \neg\phi\}.$$

Whereas if $B \bot \neg\phi = \varnothing$, then $\gamma(B \bot \neg\phi) = \{B\}$. Say that a selection function γ for B is *transitively (and reflexively) relational* iff there exists a transitive (and reflexive) relation $\geq$ that generates γ in the way just presented.[8] It seems tempting to think that a selection function is "OK" only if it is transitively and reflexively relational.

It turns out that the transitively relational selection functions generate all and only the simple belief revision strategies that satisfy the AGM axioms—a classic result due to Alchourrón et al. (1985). So we have two

[8] Note that *not* every transitive and reflexive relation $\geq$ generates a selection function for B. This is because a careless design of $\geq$ could easily result in a γ that violates condition (b), which is required by the definition of selection functions.

equivalent presentations of the same set of revision strategies: one is to use the AGM axioms to define a set of revision strategies, and the other is to construct a set of revision strategies from (1) Levi identity, (2) partial meet contraction, and (3) the set of transitively relational selection functions. This is a *representation result*, a result saying that two apparently different constructions or definitions lead to one and the same thing.

If any "at-least-as-good" relation $\geq$ employed to define a selection function should be both transitive and reflexive, then the classic AGM representation result seems to miss something: we see transitivity mentioned, but where is reflexivity? Don't worry. Rott (1993) proves that we can add reflexivity while retaining the representation result; that is, the selection functions that are transitively *and reflexively* relational generate all and only the simple belief revision strategies that satisfy the AGM axioms.

4.3 *Digression: Why Prove Representation Results?*

We have seen a representation result, and will see more. Although representation results are very interesting from a mathematical point of view, it is less clear what their philosophical significance is. So let us step back and think about how a representation result might be put into philosophical service.

Here is the first possible philosophical service. Suppose that we are searching for counterexamples to the belief revision theory based on, say, partial meet contraction. Then, thanks to the above representation theorem, we are *exactly* searching for counterexamples to the belief revision theory based on the AGM axiomatization—with a bonus: it is usually easier to work out putative counterexamples by contemplating on axioms. So a representation result can be instrumental to the search for potential counterexamples.

But we should not overemphasize the importance of this instrumental role in philosophy. A representation result is sometimes overkill for this instrumental role. Without a representation result, it is still possible to find a potential counterexample to the belief revision theory based on partial meet contraction. It is not hard to see that any belief revision strategy, if constructed from partial meet contraction, must satisfy the Preservation constraint.[9] So Preservation provides a sound (albeit incomplete) axiomatization of partial meet contraction. If we can find a counterexample to Preservation interpreted as a normative thesis, then we already have a counterexample to the belief revision theory based on partial meet contraction—all done without applying a representation result.

[9] For, when the new information ϕ is compatible with the initial belief set B, we have that $B \perp \neg\phi = \{B\}$, and hence the contracted set of beliefs $\bigcap \gamma(B \perp \neg\phi)$ must be B itself, to which the agent is going to add ϕ in order to form the new belief set $B * \phi = B + \phi$.

The lesson seems to be the following. A partial, sound axiomatization already starts to facilitate the search for potential counterexamples. It would be great if we also had a representation result. For then we are sure that, if there is any genuine counterexample, it must violate at least one of the axioms mentioned in the representation result—look no further. But it is difficult to decide how much time to invest in proving a representation conjecture, especially if the only payoff is an aid to the search for counterexamples.

A representation result might provide another philosophical service. Consider the belief revision theory T whose formal part is axiomatized by the AGM axioms. Assume the following.

(E) We have tried very hard to work out potential counterexamples to T but in vain.

Then this is good evidence for theory T. Now consider the belief revision theory T' whose formal part is constructed from partial meet contraction with transitively and reflexively relational selection functions. And assume the following.

(E') The construction procedure of T' seems to describe what a rational agent could follow in order to revise beliefs, and this "somehow" lends plausibility to T'.

So now we have evidence for T and distinct, independent evidence for T'. But, given the representation result, T and T' are one and the same belief revision theory. So we have two independent pieces of evidence for a single belief revision theory—this is a case of convergence of evidence. So a representation theorem can play an *argumentative* role in the convergence of evidence for a belief revision theory. But notice that the existence of this argumentative role is contingent on the truth of E and E'. Also notice that what E' means is unclear, depending on what is meant by 'somehow'—this is an issue we will discuss more in Section 6.2.

Enough digressions. Let us return to constructions of formal theories of belief revision.

4.4 *Orderings over Possible Worlds*

If we think that the construction techniques presented above are too restrictive due to their commitment to Preservation, we have to look for more flexible construction techniques, such as the one presented below.

Imagine that we are trying to determine the revised belief set $B * \phi$ in light of new information ϕ. Assume, for the sake of simplicity, that to believe something is to rule out some possibilities (except the limiting case in which one rules out no possibility at all). Which possibilities to rule out? We do not treat all possibilities equally; we treat some as more plausible than some others. We want to rule out the possibilities that are implausible. This inspires the following procedure.

> Step (I). Rule out the possibilities in which new information ϕ is false.
>
> Step (II). Among the possibilities that remain on the table, figure out the worlds that are most plausible, and rule out all the others.
>
> Step (III). Believe that the actual world is one of those that remain on the table—that is, let $B * \phi$ be the set of sentences that are true in every possibility that remains on the table.

So a "more-plausible-than" relation between possibilities can be used to generate a simple belief revision strategy in steps (I)–(III). This idea can be traced at least back to Shoham's (1987) work on so-called "preferential" semantics of nonmonotonic logic,[10] given Makinson and Gärdenfors' (1991) idea that nonmonotonic logic and (simple) belief revision theory are two sides of the same coin.[11]

The informal presentation in the above can be made rigorous as follows. Suppose that we have a set W of possible worlds for interpreting the language $\mathcal{L}$ in use. That is, suppose that every sentence ϕ in $\mathcal{L}$ expresses a proposition $|\phi|$, which is a subset of W and understood to contain all and only the worlds at which ϕ is true. There are metaphysical views about what possible worlds are, and there are many different mathematical models that might or might not reflect what they really are (such as identifying possible worlds with purely set-theoretic entities, or sets of linguistic entities, etc.). For present purposes, we only need to care about how we are going to make use of them, rather than what they really are. Assume that $\mathcal{L}$ is a language for propositional logic. Say that W is a *universe* of possible worlds with assignment function $|\cdot|$ for language $\mathcal{L}$ iff: (1) $|\neg\phi| = W \setminus |\phi|$, (2) $|\phi \wedge \psi| = |\phi| \cap |\psi|$, and (3) W is fine-grained enough so that sentences in $\mathcal{L}$ are assigned the same subset of W iff they are logically equivalent.[12] Here is an example: let $\top$ denote a tautology, so $|\top| = W$. Note that this model of possible worlds is quite flexible: a universe W in use is allowed to be so fine-grained that there are two distinct possible worlds w, w' in W that make exactly the same sentences in $\mathcal{L}$ true. Namely, a W in use is allowed to make distinctions that language

[10] Shoham (1987) talks literally about "more-preferred-to" instead of "more-plausible-than." But his point is to use an ordering over possible worlds, no matter how it is to be interpreted.

[11] See the appendix (Section 8.1) for a presentation of this idea.

[12] This ensures that a set Γ of sentences entails a sentence ϕ iff $\bigcap\{|\psi| : \psi \in \Gamma\} \subseteq |\phi|$, which captures the idea that entailment is truth preservation.

$\mathcal{L}$ does not make (but a richer language possibly does). This flexibility will be crucial later.

Let $\geq$ be a binary relation on a universe W of possible worlds for language $\mathcal{L}$. For any worlds $w, w' \in W$, understand $w \geq w'$ as saying that w is at least as plausible as w' with respect to $\geq$. World w is (strictly) more plausible than w' with respect to $>$ iff $w \geq w' \not\geq w$. Let $\max(U, \geq)$ denote the set of most plausible worlds in U with respect to $\geq$. To be more precise, $\max(U, \geq)$ is defined to be the set of worlds $w \in U$ such that $w < w'$ for no $w' \in U$.[13] Then use $\geq$ to generate a belief revision strategy $S_{\geq}$ as follows: given new information ϕ, let the revised belief set $S_{\geq}(\phi)$ contain a sentence ψ iff ψ is true at every possible world in $\max(|\phi|, \geq)$.

Definition (Order-Generated Revision Strategy).

$S_{\geq}(\phi) \ =_{\text{def}} \ \{\psi \in \mathcal{L} : |\psi| \supseteq \max(|\phi|, \geq)\}$,

which is the revised belief set $B * \phi$;

$S_{\geq}(\,) \ =_{\text{def}} \ S_{\geq}(\top) = \{\psi \in \mathcal{L} : |\psi| \supseteq \max(W, \geq)\}$,

which is the initial belief set B.

So, given an arbitrary binary relation $\geq$ over W, we can use it to generate a simple belief revision strategy $S_{\geq}$. Hence a set R of binary relations can be used to generate a formal theory of simple belief revision, i.e. $\{S_{\geq} : \geq \in R\}$.

But which binary relations $\geq$ are "OK" for generating revision strategies? We may consider requiring, for example, that any relation $\geq$ in use be a *preorder*, i.e. satisfy the following.

Reflexivity. $w \geq w$, for all $w \in W$.

Transitivity. If $w \geq w'$ and $w' \geq w''$, then $w \geq w''$, for all $w, w', w'' \in W$.

And we may consider the stronger requirement that any $\geq$ in use be a *complete order*, i.e. a preorder that also satisfies the following.

Completeness. Either $w \geq w'$ or $w \leq w'$, for all $w, w' \in W$.

Completeness is a substantial constraint.

Observation (I). Whenever we use complete preorders to generate belief revision strategies, Preservation is guaranteed to be satisfied.

Observation (II). Violation of Preservation becomes possible when we no longer require completeness.

[13] Note that this is *not* the condition that $w \geq w'$ for all $w' \in U$.

The second observation can be proved in a quite instructive way. The proof strategy is to construct an incomplete preorder of relative plausibility that captures the Three Composers case (which served as an alleged counterexample to Preservation in Section 2.1). Let I_x mean that x is Italian, F_x mean that x is French. Let Verdi, Bizet, and Satie be denoted by v, b, and s, respectively. Let $I_v F_b F_s$ denote the possible world in which Verdi is Italian, Bizet is French, and Satie is French. In general, a possible world assigns the two nationalities (I and F) to the three composers (v, b, and s). So there are eight possible worlds, shown in Figure 1. The arrows represent the ordering we are going to define: $w \geq w'$ iff either $w = w'$ or there is a chain of arrows linking w' upward to w. (This is called a *Hasse diagram*.) The rationale behind this ordering $\geq$ can be seen from the following, equivalent definition of $\geq$:

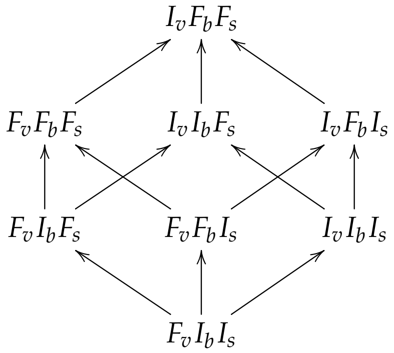

Figure 1: Hasse diagram of the Three Composers problem

- let $I_v F_b F_s$ be the most plausible world, which the agent believes to be the actual world at the initial stage;
- let $\mathsf{diff}(w)$ be the set of composers x such that w differs from the most plausible world $I_v F_b F_s$ in the nationality of composer x.
- $w \geq w'$ iff $\mathsf{diff}(w) \subseteq \mathsf{diff}(w')$; roughly speaking, the less a world differs from the most plausible world, the more plausible it is.

It is not hard to see that this is an incomplete order. Now we are ready to show that the above plausibility order is a countermodel that witnesses Observation (II). At the initial stage, the agent believes that the actual world is the most plausible world: $I_v F_b F_s$. Then the agent receives the first information E, that v and b are compatriots. So the worlds incompatible with that information are ruled out, as shown on the left side of Figure 2. At this stage, the agent believes that the actual world is one of the two most plausible worlds: $F_v F_b F_s$ and $I_v I_b F_s$. Then the agent receives the second

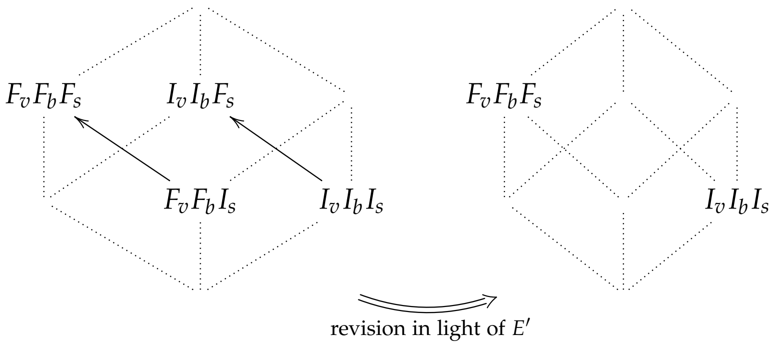

Figure 2: Revising in light of E, and then E'

information E', that v and s are compatriots. So the worlds incompatible with that information are ruled out, as shown on the right side of Figure 2. At this final stage, the agent believes that the actual world is one of the two most plausible worlds: $F_vF_bF_s$ and $I_vI_bI_s$. It is routine to verify that the transition from the left to the right represents the agent's second revision of beliefs in the Three Composers case, which violates Preservation. This establishes Observation (II).

There is one more constraint on orders that we need to consider. The Consistency axiom, which occurs in both axiom systems AGM and P$^+$, seems very plausible. But it might be violated when we use a preorder. To see why, consider a preorder $\geq$ and a consistent piece of new information ϕ such that every world in $|\phi|$ is less plausible than some other world in $|\phi|$. In that case, $\max(|\phi|, \geq) = \varnothing$ and hence:

$$\begin{aligned} B * \phi &= S_{\geq}(\phi) \\ &= \text{the set of sentences in } \mathcal{L} \text{ true at every world in } \max(|\phi|, \geq) \\ &= \text{the set of sentences in } \mathcal{L} \text{ true at every world in } \varnothing \\ &= \text{the set of all sentences in } \mathcal{L}\text{, which is inconsistent.} \end{aligned}$$

And this violates axiom Consistency. To satisfy axiom Consistency, the minimal constraint we need to impose on plausibility orders $\geq$ is this:

> **$\mathcal{L}$-Smoothness.**[14] For every sentence ϕ in $\mathcal{L}$, if $|\phi|$ is nonempty, then there is no infinite sequence $(w_0, w_1, w_2, \dots)$ on $|\phi|$ such that $w_0 < w_1 < w_2 < \dots$.

Now we are in a position to state Grove's (1988) representation result: for any simple belief revision strategy S such that $S() = S(\top)$, S satisfies the AGM axiom system iff S is generated by some $\mathcal{L}$-smooth complete preorder.

[14] This is also called the *limit* assumption in the literature on semantics of conditionals.

Those who would like to relax the Preservation axiom would be more interested in the representation result for axiom system P^+: for any simple belief revision strategy S such that $S() = S(\top)$, S satisfies axiom system P^+ iff S is generated by some $\mathcal{L}$-smooth preorder over some universe W of possible worlds. To ensure that the "only if" side holds, it is crucial to allow W to be sufficiently fine-grained. This result can be obtained by translating a result in nonmonotonic logic into belief revision theory. To be more precise, this result is translated from an immediate corollary of Kraus, Lehmann, and Magidor's (1990) representation theorem for the so-called system P of nonmonotonic logic,[15] where the translation in use is due to Makinson and Gärdenfors (1991).[16]

A technical remark on the use of mathematical tools: Grove (1988) uses the so-called sphere systems, which do the same job as complete preorders in the present context. Kraus et al. (1990) use strict partial orders, which also do the same job as preorders in the present context. It just turns out that, in order to unify these two works in the same setting, it seems most convenient to use preorders.

4.5 *Generalization to Iterated Belief Revision*

The technique we've just discussed—constructing plausibility orderings—can be easily carried over from simple belief revision to iterated belief revision.

Let $\geq$ be an order that represents relative plausibility between worlds. Recall how $\geq$ determines a belief revision procedure—in three steps. First, discard the worlds in which ϕ is false; second, among the worlds that are still on the table, figure out the worlds that are most plausible with respect to $\geq$, and discard all the others; last, let the agent believe that the actual world is one of those that remain on the table. This is a procedure for "one-time" belief revision. Next time we receive new information, how are we to find a plausibility order for our use? It is too bad that the above procedure discards some worlds and thereby destroys the structure of $\geq$. What we need to do, for the sake of iterated belief revisions, is to use the new information to revise the plausibility order $\geq$ we currently have and obtain a new order $\geq_{*\phi}$—a new plausibility order that we can use when we receive the next piece of information.

[15] Kraus et al. (1990) use a setting slightly different from our current setting: (i) instead of preorders they use strict partial orders, (ii) instead of primitive possible worlds they use indexed valuation functions for atomic sentences, and (iii) instead of using $>$ to mean "is more plausible than," they use $\prec$ (but not the other way round!) to mean "is preferred to" or "is more normal than." But these differences between the two mathematical settings do not matter insofar as the underlying idea is concerned.

[16] Their translation is presented in the appendix (Section 8.1).

So let an agent start by having a plausibility order $\geq$ and believing that the actual world is among the most plausible worlds, plausible with respect to $\geq$. When she receives new information ϕ_1, she uses the new information to revise the current order $\geq$ into a new one $\geq_{*(\phi_1)}$, and believes that the actual world is among the most plausible worlds, plausible with respect to the new order $\geq_{*(\phi_1)}$. Then, when she receives another piece of information ϕ_2, let her repeat the above procedure: use the latest information ϕ_2 to revise $\geq_{*(\phi_1)}$ into a new order $\geq_{*(\phi_1,\phi_2)}$, and believe that the actual world is among the most plausible worlds, plausible with respect to the latest order $\geq_{*(\phi_1,\phi_2)}$. In general, after receiving a finite stream of information $\phi_1, \phi_2, \ldots, \phi_n$ and revising her plausibility order successively, she will come to believe that the actual world is among the most plausible worlds, plausible with respect to the latest order $\geq_{*(\phi_1,\phi_2,\ldots,\phi_n)}$. To recap: the idea is to construct iterated revisions of plausibility orders:

$$\begin{tikzcd}
\geq \arrow[r] & \geq_{*(\phi_1)} \arrow[r] & \geq_{*(\phi_1,\phi_2)} \arrow[r] & \geq_{*(\phi_1,\phi_2,\phi_3)} \arrow[r] & \cdots
\end{tikzcd}$$

and let it generate iterated revisions of beliefs (as byproducts or epiphenomena):

$$\begin{tikzcd}
\geq \arrow[r] \arrow[d] & \geq_{*(\phi_1)} \arrow[r] \arrow[d] & \geq_{*(\phi_1,\phi_2)} \arrow[r] \arrow[d] & \geq_{*(\phi_1,\phi_2,\phi_3)} \arrow[r] \arrow[d] & \cdots \\
S() & S(\phi_1) & S(\phi_1,\phi_2) & S(\phi_1,\phi_2,\phi_3) & \cdots
\end{tikzcd}$$

This idea can be formalized as follows. A *strategy for iterated revision of plausibility orders* is a function $\geq_*$ that maps every finite sequence $(\phi_1, \ldots, \phi_n)$ of sentences in language $\mathcal{L}$ to a preorder $\geq_{*(\phi_1,\ldots,\phi_n)}$ over W. Every order revision strategy $\geq_*$ generates a belief revision strategy as follows.

> DEFINITION (ORDER-GENERATED REVISION STRATEGY).
>
> $$S_{\geq_*}(\phi_1, \ldots, \phi_n) =_{\text{def}} \{\psi \in \mathcal{L} : |\psi| \supseteq \max(W, \geq_{*(\phi_1,\ldots,\phi_n)})\},$$
>
> i.e. the set of sentences that are true at every possible world that is most plausible with respect to $\geq_{*(\phi_1,\ldots,\phi_n)}$.

This is how iterations of belief revision can be generated from iterations of plausibility order revision. While it might be difficult to construct the former directly, the latter turns out to be not that difficult to construct. Consider the following construction technique called "cut-and-paste":

> DEFINITION (CUT-AND-PASTE REVISION). Say that $\geq'$ is obtained from $\geq$ by *cut-and-paste revision* on a subset X of W iff:
>
> (1) for all $w, u \in X$, $w \geq' u$ iff $w \geq u$;

(2) for all $w, u \notin X$, $w \geq' u$ iff $w \geq u$;

(3) for all $w \in X$ and $u \notin X$, $w > u$.

Namely, we "grab" the order $\geq$ over the whole W, "cut" the part of $\geq$ over X, and "paste" it on "top" of the other part $W \setminus X$, making any world inside X more plausible than any world outside X (condition 3), without changing the ordering of the worlds inside X (condition 1), nor changing the ordering of the worlds outside X (condition 2). Here are two examples of cut-and-paste revision.

Definition (Conservative and Radical Revisions).

Radical revision of $\geq$ on ϕ is cut-and-paste revision of $\geq$ on $|\phi|$. This is sometimes called *lexicographic revision*.

Conservative revision of $\geq$ on ϕ is cut-and-paste revision of $\geq$ on $\max(|\phi|, \geq)$.

Radical revision changes a lot, while conservative revision just does a little. What if we want to revise not that much nor that little, but something in between? Consider the following, very general kind of order revision:

Definition (Canonical Revision). The revision from $\geq$ to $\geq'$ in light of information ϕ is said to be *canonical* iff:

(1) ϕ is true at all worlds that are most plausible with respect to $\geq'$;

(2) for all $w, u \in |\phi|$, $w \geq' u$ iff $w \geq u$;

(3) for all $w, u \notin |\phi|$, $w \geq' u$ iff $w \geq u$;

(4) for all $w \in |\phi|$ and $u \notin |\phi|$:

- if $w > u$, then $w >' u$,
- if $w \geq u$, then $w \geq' u$,
- if $w \not\leq u$, then $w \not\leq' u$.

Condition (1) ensures that the new information is to be believed. Condition (2) ensures that there is no change to the plausibility relation among the worlds that make ϕ true. Condition (3) does something similar, ensuring that there is no change to the plausibility relation among the worlds that make ϕ false. Condition (4) appears quite complicated, but it is meant to capture this intuitive idea: given any worlds w and u that make the new information true and false, respectively, the plausibility relation of w to u should not be "downgraded." Radical revisions and conservative revisions are both special cases of canonical revisions.

So, to construct a formal theory of iterated belief revision, we can proceed by specifying a set $\mathcal{S}$ of strategies for iterated revision of plausibility orders, and then letting it generate a set of iterated belief revision strategies $\{S_{\geq_*} : \geq_* \in \mathcal{S}\}$.

But which ones to put into $\mathcal{S}$? There are at least two dimensions to consider. First, do we want to allow some strategies in $\mathcal{S}$ to output incomplete orders, or do we want to require every strategy in $\mathcal{S}$ to output only complete preorders? Prefer the former option if you like Preservation; otherwise prefer the latter option. Second, do we want to require that every strategy $\geq_*$ in $\mathcal{S}$ always follow canonical revision, i.e. the revision from $\geq_{*(\ldots)}$ to $\geq_{*(\ldots,\phi)}$ must be a canonical revision on ϕ? If we do, do we want to require something more, such as that every strategy in $\mathcal{S}$ always follow radical revision, or that every strategy in $\mathcal{S}$ always follow conservative revision, or some other constraint?

Darwiche and Pearl (1997), for example, opt for complete preorders together with canonical revision. Some think that the requirement of canonical revision is too weak: Boutilier (1996) adds the requirement of conservative revision; Jin and Thielscher (2007) add the requirement that, for all worlds w, u such that $w \in |\phi|$ and $u \notin |\phi|$, if $w \geq_{*(\ldots)} u$, then $w >_{*(\ldots,\phi)} u$.

This subsection has presented a reductionist approach that tries to reduce iterated belief revision to revision of orders, but there has been the worry that a reductionist approach is too restrictive. See Stalnaker (2009) for an example meant to support this worry (this example will be discussed in Section 5.3). Also see Booth and Chandler (2017) for more examples, meant to argue against any reductionist approach that reduces iterated belief revisions to functions that send a plausibility order and a piece of information to a plausibility order.

4.6 *Learning-Theoretic Analysis*

Perhaps a belief revision strategy is better insofar as it better serves the goal of one's inquiry, e.g. the goal of *learning* whether all ravens are black. In this subsection, we will construct a belief revision theory by addressing the issue of how to choose belief revision strategies that serve the goal of learning well—this is an issue typically addressed in *formal learning theory*. We will be guided by two questions. First, how are we to define when a belief revision strategy performs well with respect to the goal of learning? No matter how we are to define learning performance, the performance of a strategy is typically contingent upon what the world is like, something that we have no control over and lack knowledge about. There might be a strategy that performs well in one case but poorly in another case, and an alternative strategy that performs in the opposite way. This brings out the second question: which strategy is better and which is to be ruled out by our belief revision theory? The following illustrates a learning-theoretic answer with a case study on the Raven Problem Section 3.3: "Are all ravens black?"

To choose among belief revision strategies for tackling the Raven Problem, let us draw a payoff table. Table 1, like any typical decision table, has three kinds of elements: (i) columns, (ii) rows, and (iii) cells. The *columns* correspond to the relevant, mutually exclusive possibilities. Recall that h is the hypothesis that all ravens are black, b_i means that the i-th raven observed is black, and n_i means that it is nonblack. So, for example, the first column "h" corresponds to the possibility in which h is true and, hence, all ravens are black. The column "$\neg h, b_1, \ldots, b_i, n_{i+1}$" corresponds to the possibility in which not all ravens are black and the first nonblack raven to be observed is the $(i+1)$-th one. The *rows* correspond to the options to choose from. In the above table there are only three options—three belief revision strategies—which I will define soon. Each row and each column intersects at a *cell*, in which we will specify the outcome of the corresponding option in the corresponding possibility. Each of those outcomes will concern *how well* a belief revision strategy serves the goal of learning the true answer to the question posed: "Are all ravens black?" When all those outcomes are specified, we will try to figure out which options that are "OK", or at least which are not "OK."

	h	$\neg h, n_1$	$\neg h, b_1, n_2$	$\ldots$	$\neg h, b_1, \ldots, b_{100}, n_{101}$	$\ldots$
S_{ind}						
S_{count}						
S_{skep}						

Table 1: An incomplete payoff table for the Raven Problem

The three strategies listed in the decision table are defined as follows. The *skeptical* strategy S_{skep} always asks one to believe the logical consequences of one's accumulated information, no more and no less. That is:

$$S_{\text{skep}}(\phi_1, \ldots, \phi_i) =_{\text{def}} \mathsf{Cn}\{\phi_1, \ldots, \phi_i\},$$

where $\mathsf{Cn}\, X$ denotes the set of logical consequences of X. So, for example, $S_{\text{skep}}(b_1, b_2, \ldots, b_{i-1}, n_i)$ contains $\neg h$ because n_i entails $\neg h$. But $S_{\text{skep}}(b_1, b_2, \ldots, b_i)$ excludes h no matter how large i is—so this strategy is what the inductive skeptic would recommend.

An *inductive* strategy is a strategy that starts from asking one to believe just the logical consequences of the accumulated information, but after observing a certain amount of black ravens in a row without any counterexample, it asks one to believe h, the inductive hypothesis that all ravens are black. Here is an example:

$$S_{\text{ind}}(\phi_1, \ldots, \phi_i) =_{\text{def}} \begin{cases} \mathsf{Cn}\{\phi_1, \ldots, \phi_i, h\} & \text{if } i \geq 100 \text{ and } \phi_j = b_j \text{ for} \\ & \text{all } j \leq i, \\ \mathsf{Cn}\{\phi_1, \ldots, \phi_i\} & \text{otherwise.} \end{cases}$$

This strategy decides to make the inductive leap at the 100th black raven. We could replace 100 with another positive integer, which would generate another inductive strategy.

A *counter-inductive* strategy works as follows: when seeing more and more black ravens in a row without any nonblack raven, this strategy will start to ask one to believe $\neg h$ at some point—violating Ockham's Razor—and it will ask one to believe h only at a later point. Here is an example:

$$S_{\text{count}}(\phi_1, \ldots, \phi_i) =_{\text{def}} \begin{cases} \mathsf{Cn}\{\phi_1, \ldots, \phi_i, \neg h\} & \text{if } 50 \leq i < 100 \text{ and} \\ & \phi_j = b_j \text{ for all } j \leq i, \\ \mathsf{Cn}\{\phi_1, \ldots, \phi_i, h\} & \text{if } i \geq 100 \text{ and } \phi_j = b_j \\ & \text{for all } j \leq i, \\ \mathsf{Cn}\{\phi_1, \ldots, \phi_i\} & \text{otherwise.} \end{cases}$$

What makes it counter-inductive is the first clause. Replacement of 50 and 100 with other numbers m and n (with $m < n$) would generate other counter-inductive strategies.

For the sake of simplicity, let us compare just the three strategies explicitly defined above, although infinitely many more can be considered if we wish. So we have only three rows in the payoff table to think about.

Next: fill the cells with outcomes. The kind of outcome to be specified should say how well a strategy performs to help one achieve the goal, where the present goal is set to learn whether all ravens are black. The following introduces two performance criteria.

Say that a strategy *will learn* whether h is true given a column C iff, whenever C holds and one obtains more and more information, there will be a "learning moment" at which the strategy asks one to believe, and always continue to believe, the unique answer in $\{h, \neg h\}$ that is true given C. The following definition makes this more precise.

> DEFINITION (LEARNING WITH RESPECT TO THE RAVEN PROBLEM). A strategy S *will learn* whether h is true given column C iff:
>
> for any infinite sequence $(\phi_1, \phi_2, \ldots)$ such that:

- every finite segment of $(\phi_1, \phi_2, \ldots)$ is in the information space $\mathcal{I}_{\text{raven}}$ in use (that is, every entry ϕ_i is either b_i or n_i),
- the conjunction $\bigwedge_{i \geq 1} \phi_i$ is compatible with possibility C,

there exists a natural number n, called a "learning moment," such that:

- for each $i \geq n$, $S(\phi_1, \phi_2, \ldots, \phi_i)$ is consistent and entails the unique sentence in $\{h, \neg h\}$ that is true given C.

Here I only define the concept of learning for solving the Raven Problem, but generalization is straightforward—please see appendix (Section 8.2). An essential feature of this definition is that it refers to the information space $\mathcal{I}_{\text{raven}}$ in use, which is meant to include all and only the pieces of information that can be *available* to the inquirer. In principle we can try to solve the Raven Problem by adopting a strategy for iterated belief revision, which is defined on the much larger information space $\mathcal{I}_{\text{finite}}$ that contains all finite sequences of sentences. But, in that case, we still need to use the smaller information space $\mathcal{I}_{\text{raven}}$ to correctly define (or characterize) when a strategy will learn the true answer given a column.

We are now in a position to fill some cells with (partial) outcomes: see Table 2. An occurrence of "Y" in a cell means: "yes, the strategy will learn whether h is true given this column." Similarly, "N" means: "no, it won't learn." Just to check that we get this part right: given the first column "h" ("all ravens are black"), when more and more black ravens are observed, the skeptic strategy will never ask one to believe the true answer h, and hence, it will not learn whether h is true given column "h." That said, the skeptic strategy will learn whether h is true given any other column "$\neg h, b_1, \ldots, n_i$": the right answer is obtained, and held on to, beginning from the i-th observation, because n_i entails $\neg h$. It is not hard to verify that the cells left blank in the above should all be filled with "Yes."

	h	$\neg h, n_1$	$\neg h, b_1, n_2$	$\ldots$	$\neg h, b_1, \ldots, b_{100}, n_{101}$	$\ldots$
S_{ind}						
S_{count}						
S_{skep}	N	Y	Y	$\ldots$	Y	$\ldots$

Table 2: Payoff table for the Raven Problem continued

We want to think about, not just whether a strategy will learn, but also how well it learns. It would be great to have a strategy that might occasionally points to a falsehood but, once it points to the truth, it will never let it go. If a strategy has that property given a column, say that it is *stable* given that column (which is arguably a virtue that Plato praises towards the end of *Meno*). For example, the counter-inductive strategy is not stable given the column "$\neg h, b_1, \ldots, b_{100}, n_{101}$": it asks one to believe the truth $\neg h$ on the 50th observation, but fails to continue to do so on the 100th, violating stability. Now, for each cell, let us specify (i) whether the strategy will learn and (ii) whether it is stable: see Table 3. The answers to (i) and (ii) for each cell are specified in the first and second components of the ordered pair, respectively.

	h	$\neg h, n_1$	$\neg h, b_1, n_2$	...	$\neg h, b_1, \ldots, b_{100}, n_{101}$	...
S_{ind}	(Y, Y)	(Y, Y)	(Y, Y)	...	(Y, Y)	...
S_{count}	(Y, Y)	(Y, Y)	(Y, Y)	...	(Y, **N**)	...
S_{skep}	(**N**, Y)	(Y, Y)	(Y, Y)	...	(Y, Y)	...

Table 3: Payoff table for the Raven Problem completed. The first component concerns whether it will learn; the second, whether it is stable.

With the payoff table completed, it is time to think about which strategies are "OK" and which are not. Presumably, learning is better than failing to learn; stability is better than instability. With that in mind, a learning theorist would be interested in arguing that both the skeptical strategy S_{skep} and the counter-inductive strategy S_{count} are not "OK". Three styles of arguments are available for consideration.

STYLE 1. We can argue that both the skeptical strategy S_{skep} and the counter-inductive strategy S_{count} are not "OK" if we are happy to apply the Dominance Principle, which says that an option is not "OK" if it is dominated in the sense that some alternative option performs at least as well in all columns and does strictly better in some column.

STYLE 2. To argue for the same conclusion, we do not have to apply the dominance principle. An alternative is to apply the so-called *Maximin* rule. According to Maximin, we are to, first, figure out the worst possible outcome of each option, and then judge that an option is not "OK" if[17] its worst outcome is worse than the worst outcome of some alternative option. Namely, Maximin asks one to *maxi*mize the *min*imal payoff. The worst outcomes are identified in Table 4. So, according to Maximin, both the skeptical strategy S_{skep} and the counter-inductive strategy S_{count} are not "OK".

STYLE 3. Perhaps the minimal argument that suffices to obtain the same conclusion is to rely on a premise of this form:

> TEMPLATE FOR ACHIEVABILIST THESES. If the empirical problem in question is easy enough so that it is possible to achieve epistemic

[17] Note that the Maximin rule is formulated here in terms of 'if' rather than 'if and only if'; this is to ensure that the Maximin rule is in general compatible with the Dominance principle.

	worst outcome
S_{ind}	(Y, Y)
S_{count}	(Y, **N**)
S_{skep}	(**N**, Y)

Table 4: Worst possible outcomes for the Raven Problem

> standard X, then a revision strategy for that empirical problem has to achieve (at least) X in order to be "OK".

Now, let X be "learning the truth with stability in all the columns (all the possibilities under consideration)". The Raven Problem is indeed that easy, as witnessed by the first row of the payoff table. So, if we are happy to accept that premise, we can argue that an "OK" revision strategy for the Raven Problem must (at least) achieve learning and stability in all the columns, ruling out the skeptical and the counter-inductive strategies.

The learning-theoretic analysis presented above is just a "baby version" for the sake of illustration. It is adapted from Genin and Kelly (2018) and Kelly, Genin, and Lin (2016), which build on Schulte (1999) and Kelly (2007). More generally, a belief revision theory can be constructed by considering the learning performances of belief revision strategies in possible scenarios. This idea admits of many possible implementations.

- *We may consider enriching the specifications of outcomes.*

  We have only talked about whether a revision strategy will learn and whether it is stable. But do we also want to consider other kinds of learning performance? Think about these: How many retractions of beliefs will be incurred? How many times will a false answer be believed? How fast will the true answer be learned?

- *We need to find a way to evaluate revision strategies in terms of the payoff table.*

  We may consider using a decision rule such as Dominance and Maximin. But how about other decision rules like Minimax Regret, Maximax, or even Maximization of Expected Utility (if the use of prior probabilities does not beg the inductive skeptic's question)? We may also consider relying on one achievabilist thesis or another, or even a set of such theses. But what achievabilist theses are correct? That is, what epistemic standards have to be achieved when achievable?

All those considerations and their possible variants, in combination, provide what we may call the learning-theoretic toolkit for constructing various formal theories of belief revision. But which specific tools *should* we use in order to construct a belief revision theory that has a plausible normative interpretation? This issue will be revisited in Section 6.3.

Also see Kelly (1999) for an application of learning-theoretic analysis to iterated belief revision, which considers the possibility of receiving mutually contradictory pieces of information.

The learning-theoretic analysis need not be a rival to the aforementioned approaches to belief revision theory. Indeed, as we have seen, the learning-theoretic analysis is able to rule out some notable revision strategies, such as the skeptical and counter-inductive strategies. This ability seems to *complement* the more standard, AGM belief revision theory: the skeptical strategy is not ruled out by the AGM theory because it can be modeled by a complete order of relative plausibility that takes every possible world to be equally plausible; the counter-inductive strategy is not ruled out by the AGM theory, either, because it can also be modeled by an appropriate complete order of relative plausibility. Perhaps the right theory of belief revision should be constrained jointly by the learning-theoretic considerations and the considerations that follow, generalize, or weaken the AGM theory. See Baltag, Gierasimczuk, and Smets (2016) and Genin and Kelly (2018) for more on the possibility of such a joint project.

4.7 *Other Construction Techniques*

There are many other techniques for constructing belief revision theories. Let me mention some of the most influential ones.

- Instead of using plausibility orderings over possible worlds, we may use orderings over sentences, the so-called *epistemic entrenchment* orderings (Gärdenfors & Makinson, 1988). This idea has been applied to both simple belief revision and iterated belief revision (Nayak, 1994).

- On the approach of partial meet contraction, it is standardly assumed that a belief set B be closed under logical consequence, but we may relax that assumption, letting B be a mere set of sentences, called a *belief base*, on which the agent bases other beliefs (Hansson, 1994, 1999).

- If we think that almost all formal theories of simple belief revision in the literature are too strong, we can resort to the standard translation between simple belief revision and nonmonotonic inference (Makinson & Gärdenfors, 1991), which I present in the appendix (Section 8.1), and then translate a sufficiently weak nonmonotonic logic into an equally weak theory of belief revision. The literature of nonmonotonic logic does provide very weak systems, such as Reiter's (1980) default logic.[18] When we translate Reiter's default logic into belief revision theory, the result is even weaker than system P^+, let alone AGM.[19]

- Spohn (1988) proposes an approach to iterated belief revision theory, which considers belief revisions in situations of the following kind: an agent receives new information, but she is not fully certain whether it is true, and somehow has a clear idea of how uncertain she is supposed to be, where the uncertainty in question is measured by ordinal numbers. See the chapter on ranking theory (Huber, this volume).

For an extensive, detailed survey of construction techniques, see Rodrigues et al. (2011).

5 HOW TO ARGUE AGAINST

To argue against a normative theory of belief revision, the paradigmatic way is to provide intuitive counterexamples. But an alleged counterexample usually raises a question: "Is that a genuine counterexample?" Let us think about this issue by discussing concrete examples.

5.1 *Three Composers Revisited*

Recall the case of Three Composers, which we considered in Section 2.1. To facilitate cross reference, let me reproduce it below:

> Example (Three Composers). Consider three composers: Verdi, Bizet, and Satie. The agent initially believes the following.
>
> (A) Verdi is Italian;
>
> (B) Bizet is French;
>
> (C) Satie is French.
>
> Then the agent receives this information.
>
> (E) Verdi and Bizet are compatriots.
>
> So she retains the belief in C that Satie is French (after all, information E has nothing to do with Satie), but drops her beliefs in A and in B. Then the agent receives another piece of information.
>
> (E') Verdi and Satie are compatriots,

[18] Reiter's default logic is only one of the many approaches to nonmonotonic logic; see Brewka, Niemelä, and Truszczyński (2008) for a review.

[19] This observation is due to Makinson (1988).

> which is compatible with what she believes right before this new information arrives. Considering that she started with initial beliefs A, B, and C and has received two pieces of information E and E', which jointly say that the three composers are compatriots, now she drops her belief in C.

Let us recall that the second revision is an alleged counterexample to Preservation as a necessary condition of perfect rationality.

Anyone who wants to defend Preservation as a necessary condition of perfect rationality may try responding in either of the following two ways. First, the defender may try explaining why the agent in the Three Composers case is actually irrational.

The second possible response proceeds as follows. E' seems not the kind of thing that we can actually receive as new information. We would come to believe E' by inferring it from the new information that we can actually receive, such as "my music teacher just told me that Verdi and Satie are compatriots," or "I just saw a chart coloring composers in terms of their nationalities; it assigns the same color, red, to Verdi and Satie but I do not know which nationality corresponds to red." So the scenario *misspecifies* the new information that the agent actually receives. A realistic scenario should be more complicated than the one told above. So the above scenario also *underspecifies* how exactly the agent comes to gain the new belief in E' and drop the old belief in C. The goal of this response is to show that, no matter how we retell the original Three Composers scenario in a way free from misspecification and underspecification, the retold story will not be a counterexample to Preservation.[20]

To see how one may explain an alleged counterexample away by pointing to underspecification or misspecification, let me provide other examples in the following two subsections.

5.2 *Underspecification*

Katsuno and Mendelzon (2003) argue that the AGM theory is not universally applicable. They propose the following counterexample.

> EXAMPLE (BOOK AND MAGAZINE). Suppose that the agent believes that there is either a book on the table (B) or a magazine on the table (M), but not both. Consider two alternative developments of this scenario:
>
> *Case 1:* The agent is told that there is a book on the table. She then concludes B and $\neg M$.

[20] I thank Horacio Arló-Costa for bringing this possible response to my attention.

> *Case 2:* The agent is told that a book has been put on the table. She then concludes B but continues to suspend judgment about M.

So the agent starts by believing $B \vee M$ and $\neg(B \wedge M)$. Katsuno and Mendelzon agree that the AGM theory can easily explain Case 1 as follows: the agent receives information B and, hence, by the Accretion axiom in the AGM theory, she comes to believe $\neg M$. But Katsuno and Mendelzon think that Case 2 is a counterexample to the Accretion axiom in the AGM theory because (i) the new information is compatible with the old beliefs and (ii) the new information plus the old beliefs entails $\neg M$, which the agent does not believe after the revision.

The lesson they want to draw is that we need a theory of belief revision like AGM to deal with Case 1, but we need a distinct theory, what they call a theory of *belief update*, to deal with Case 2.

But the AGM theorist could respond by saying that Katsuno and Mendelzon underspecify Case 2. Here is one possible way to specify Case 2 with sufficient detail.

> *Case 2′:* The agent starts by believing not only that $B \vee M$ and $\neg(B \wedge M)$ are both true at t_0, but also that if a book is put on the table at $t_1 (> t_0)$, then, first, B is true at t_1 and, second, M is true at t_0 iff M is true at t_1. Then the agent is told, at t_1, that a book is indeed put on the table at t_1. In this case she should continue to suspend judgment about M.

Given this more detailed specification of Case 2, the AGM theorist can use the Accretion axiom to explain why the agent should suspend judgment about M at t_1. Note that the new information is consistent with the set of her old beliefs. Furthermore, the new information plus the set of her old beliefs is silent about the truth value of M at t_1 (and this is made clear by explicit references to times t_0 and t_1). Therefore, by Accretion one should suspend judgment about the truth value of M at t_1.

So Katsuno and Mendelzon's alleged counterexample does not really refute the AGM theory. The lesson is that an alleged counterexample may fail to work due to underspecification.

I want to make a second point. Belief revision theory is very interdisciplinary, studied by philosophers, logicians, and computer scientists. There are people belonging to all the three groups, but there are also people belonging to only one or two. So different belief revision theorists might have very different goals in mind when using counterexamples. A sympathetic reading of Katsuno and Mendelzon's paper—a paper in artificial intelligence—suggests that they are interested in situations where the object language is so austere that it contains no tense operators or referential expressions about time. So the conclusion they want to draw can be charitably understood as saying that, given that the object language is so austere (and hence computationally easier to deal with), the AGM theory when restricted to that language cannot accommodate Case 2. This conclusion should be very interesting to computer scientists: it would be interesting to see if Case 2 can be accommodated by an algorithm that manipulates a very simple language and implements a non-AGM belief revision theory. It is just that this conclusion, although interesting in computer science, is not equally interesting in epistemology.

5.3 *Misspecification*

Stalnaker (2009) argues against the following constraint on iterated belief revision.

> AXIOM C2 (DARWICHE AND PEARL, 1997). $S(\phi_1, \ldots, \phi_n, \alpha, \beta) = S(\phi_1, \ldots, \phi_n, \beta)$, whenever the latest information β is incompatible with the preceding information α.

This says, roughly, that when one receives information α and then the next piece of information β contradicts α, one ought to revise beliefs as if one had only received β without receiving α. Darwiche & Pearl's Axiom C2 is among the weakest studied in the belief revision literature. Indeed, it is satisfied by every revision strategy that always follows canonical revision (which is the weakest requirement of iterated belief revision discussed in Section 4.5). But Stalnaker (2009) proposes a counterexample to Axiom C2.

> EXAMPLE (COIN FLIPPING). A fair coin is flipped in each of the two rooms, 1 and 2. Alice and Bert (who I initially take to be reliable) report to me, independently, about the results: Alice tells me that the coin in room 1 came up heads, while Bert tells me the same about the coin in room 2. So I believe what they tell me at *stage one*. But then Carla and Dora, also two independent witnesses whose reliability, in my view, trumps that of Alice and Bert, give me information that conflicts with what I heard from Alice and Bert. Carla tells me that the coin in room 1 came up tails, and Dora tells me the same about the coin in room 2. These two reports are given independently, and simultaneously.[21] This is *stage two*. Finally, *stage three*: Elmer, whose reliability trumps everyone else, tells me that that the coin in room 1 in fact landed heads. (So Alice was right after all.) What should I now believe about the coin in room 2?

[21] This simultaneity assumption is crucial for Stalnaker's purposes. Although this kind of simultaneity (relative to the agent's frame of reference) is extremely rare, it is still possible. So this example is a genuine possibility.

It seems that the agent, at the final stage, should believe that the coin in room 2 came up tails, for Elmer says nothing that contradicts what Dora says. But this result, Stalnaker claims, violates Darwiche & Pearl's Axiom C2. To see why, let:

$\alpha =$ the conjunction of what Carla says and what Dora says;

$\beta =$ what Elmer says.

The latest information β contradicts the information α obtained at the preceding stage, and it does so only because it contradicts the first conjunct of α (i.e. what Carla says). But Axiom C2 asks the agent to act as if information α were not received at all and, hence, as if Dora's testimony were not received. By contrast, we seem to have the intuition that the agent should retain her belief in what Dora says—after all, the latest information β does not undermine what Dora says. The problem with Axiom C2 seems to be this: it requires that Dora's testimony be discredited *only* because it arrived at the same time as someone else's discredited testimony.

Those who want to defend Darwiche & Pearl's Axiom C2 might respond that Stalnaker actually *misspecifies* the information in question. The agent does not really receive any information whose content is that the coin in room Y came up Z. The information received should be of this form: "agent X says that the coin in room Y came up Z." That is, the real information should not be the content of what people say, but should report the fact that those people say such and such things. Then there is no contradiction between the earlier information and the later information in the Coin Flipping case, and hence there is no violation of Axiom C2—or so the response concludes.

So, if the above response is right, Stalnaker's alleged counterexample fails to work due to misspecification.

This hypothetical exchange between Stalnaker and the defender of Axiom C2 raises a deep question. The clash between Stalnaker's counterexample and the defender's response can be taken as a debate over *what counts as information*, assuming that both parties employ the same conception of information. But what if Stalnaker and the defender presuppose distinct conceptions of information? That is, what if they are talking past each other? This question points to a debate concerning the nature or goal of belief revision theory. According to the conception of information used in Stalnaker's specification of the scenario, the information that the agent receives takes the form of E_2 rather than E_1.

(E_1) Agent X says that the coin in room Y came up Z.

(E_2) The coin in room Y came up Z.

But according to another conception of information—the one used in the response—the agent only receives information of the form E_1, while E_2

comes to be believed as a result of revising the agent's old beliefs in light of information E_1. Now, if the two parties do presuppose distinct conceptions of information, the real debate is this:

> Choice among Conceptions of Information. Which conception of information should be the one used in belief revision theory? Or, without presupposing that there is a unique conception of information to be used in belief revision theory, how should those conceptions of information play their respective roles in belief revision theory?

These are difficult questions to answer. If we are going to have two conceptions of information in belief revision theory, then we will have to rewrite the formal theories presented above, for they simply do not distinguish different conceptions of information. If we are to stick with the more permissive conception of information that Stalnaker has in mind, then it seems that we are developing a belief revision theory that does not address an important kind of belief revision, i.e. the cases in which E_2 is believed as a result of belief revision in light of information E_1. But if, instead, we are to stick with the more restrictive conception of information, then we will create a slippery slope. Which of the following is the information that the agent receives?

(E_0) Agent X utters 'the coin in room Y came up Z'.
(E_1) Agent X says that the coin in room Y came up Z.
(E_2) The coin in room Y came up Z.

If we want a restrictive conception that excludes E_2 as information, why not go for the most restrictive conception that allows only E_0 as information, and take the other two to be something that the agent might come to believe by revising old beliefs in light of the sole information E_0? And, if we really adopt such a restrictive conception of information, then it seems pointless to develop a theory of iterated belief revision that aspires to take care of so many cases, including the cases in which one receives information α and later receives information β that contradicts α. These cases would be made impossible or extremely rare by the most restrictive conception of information.

So which conception(s) of information should we use in belief revision theory? That is a tough issue, not usually discussed by belief revision theorists. But Gärdenfors (1988), for example, does elaborate on the conception of information that he intends to work with.

We arrived at a foundational issue from an alleged counterexample to a belief revision theory. Discussions about counterexamples are important because we may use them to refute theories, but also because they sometimes raise deep questions concerning what exactly we want to theorize about.

6 HOW TO ARGUE FOR

Arguments for particular belief revision theories do not usually receive explicit formulations in the literature. But two argumentative approaches are discernible in the literature. On the first approach, one argues for a belief revision theory in terms of how well it survives alleged counterexamples. On the second approach, a formal but motivated construction of a belief revision theory is somehow "transformed" into an argument for the theory. Let me explain these two approaches in turn.

6.1 *Argument from Surviving Alleged Counterexamples*

We use intuitive examples to refute general theories. So a possible argument schema we may use is the following.

(i) We have worked very diligently in search of intuitive counterexamples to this normative theory of belief revision but have not been able to find a genuine counterexample.

(ii) Therefore, this theory is plausible.

This argument is certainly not valid, but perhaps it is harmless to make it valid by adding a premise: if (i) then (ii).

That is the first approach we may adopt in order to argue for a belief revision theory, but hopefully not the only approach. We may have conflicting intuitions about concrete examples. When we do, we will debate over premise (i). So it would be great to explore whether there are more theoretical, general considerations that can help us resolve or mitigate our disagreement. That brings us to the second approach.

6.2 *Argument from Construction: Partial Meet Contraction*

On the second approach, a construction of a formal belief revision theory is to be interpreted and then turned into an argument for a normative theory of belief revision. I will illustrate with two construction techniques: first with partial meet contraction (in this subsection), and then with the learning-theoretic analysis (in the next subsection).

Belief revision theorists working on partial meet contraction seem to have the following line of thought in mind. Recall that this construction technique generates belief revision strategies S as follows:

$$\begin{aligned} S(\phi) &=_{(0)} B * \phi \\ &=_{(1)} (B \div \neg\phi) + \phi \\ &=_{(2)} \bigcap \gamma(B \bot \neg\phi) + \phi \\ &=_{(3)} \bigcap \{X \in B \bot \neg\phi : X \geq Y \text{ for all } Y \in B \bot \neg\phi\} + \phi. \end{aligned}$$

These equations jointly describe a *formal* procedure by which we can use a binary relation $\geq$ over sets of sentences to generate a belief revision strategy S. Under a suitable interpretation, this procedure may tell a *story* about a rational agent who is trying to revise beliefs, about the sensible considerations that she has, and about the rational decisions that she makes. In fact, this story was already sketched in Section 4.2, in which all formal apparatuses—ranging from $\div$, $\bot$, γ, to $\geq$—were introduced with motivations. (Of course, there are details to be filled into the story sketched in that section, and some parts of the story may require fine-tuning to make the whole story plausible.) Some belief revision theorists such as Gärdenfors (1984) do take the story—the interpreted formal procedure—very seriously, and they think that the story somehow lends plausibility to the belief revision theory they construct.

The question I want to discuss here is how the above line of thought can possibly be turned into an explicit argument with a clearly specified normative conclusion. Let us explore some possibilities. Suppose that the procedure (0)–(3) of partial meet contraction has been given an interpretation in line with the motivations provided in Section 4.2. Suppose, further, that the normative thesis to be argued for is the following.

> Putative Conclusion. An agent is perfectly rational only if she has been following, and would continue to follow, a belief revision strategy S that is constructible through procedure (0)–(3).

Note that this putative conclusion does not make the implausibly strong claim that an agent is perfectly rational only if she *actually* follows procedure (0)–(3); there may be distinct procedures leading to the same final product. Now add the following premise.

> Premise (I). Procedure (0)–(3), under such and such an interpretation, describes a possible process for perfectly rational belief revision.

But the above premise *alone* does not suffice, for it only describes procedure (0)–(3) as *one* possible process for perfectly rational belief revision. This leaves us with the following open question.

> OPEN QUESTION. Is there a procedure that describes another possible process for perfectly rational belief revision, but generates a belief revision strategy not constructible through procedure (0)–(3)?

If the answer is "yes," then the putative conclusion is false. So, to make the argument valid, we need to add *at least* the following premise (or something to the same effect).

> PREMISE (II). The answer to the above question is "no."

But this second premise is far from obvious, so an argument for it is required. Indeed, since procedure (0)–(3) is committed to Preservation, the Three Composers case is a potential counterexample to Premise (II). Perhaps one can try to argue that procedure (0)–(3) describes a very "paradigmatic" process for perfectly rational belief revision—so paradigmatic that the answer to the open question is "no," and that the putative conclusion must be true. It remains to explore how one may elaborate on this line of thought.

So, for those who are sympathetic to the philosophical significance of partial meet contraction (0)–(3), a foundational issue in belief revision theory is how we may provide more premises besides (I) and produce a sensible, valid argument for the putative conclusion.

But even if such an argument can be produced, Premise (I) can be challenged. That is, one may challenge the very possibility of a workable interpretation of procedure (0)–(3). Recall the main idea of this procedure. Suppose that one receives information ϕ, and that ϕ is incompatible with the set B of one's old beliefs. Then some old beliefs have to be retracted before ϕ is added to one's stock of beliefs. That is, before one adds ϕ, one needs to find a contracted set $B \div \neg\phi$ of beliefs, a subset of B that is compatible with ϕ. It is hypothesized that one should not retract beyond necessity (but why?).[22] So let the agent consider all elements of the remainder set $B \perp \neg\phi$, i.e. all inclusion-maximal subsets of B that are compatible with ϕ. Then let relation $\geq$ sort out the "best" of those subsets. The intersection of those best subsets, $\bigcap\{X \in B\perp\neg\phi : X \geq Y \text{ for all } Y \in B\perp\neg\phi\}$, is then identified with the contracted set of beliefs, $B \div \neg\phi$. That's the main idea. But that raises an issue concerning the right interpretation of $\geq$. Let us try the following interpretation:

> INTERPRETATION OF $\geq$ (1). $X \geq Y$ means that X is at least as good as Y as a candidate for $B \div \neg\phi$.

Under this interpretation, the intersection of the "best" candidates for $B \div \neg\phi$ ("best" with respect to $\geq$) may not be a "best" candidate for $B \div \neg\phi$ ("best," again, with respect to $\geq$). So a non-optimal candidate may be

[22] For more on this issue, see Rott (2000).

selected! So this particular interpretation of $X \geq Y$ makes the construction process incoherent: one does not choose from the best candidates, but opts for the intersection of the best candidates, which may be sub-optimal.

What else could $X \geq Y$ mean? Let us try Gärdenfors' (1984) suggestion:

> INTERPRETATION OF $\geq$ (2). $X \geq Y$ means that X is epistemically at least as "important" as Y.

Following this interpretation, procedure (0)–(3) assumes that the contracted belief set $B \div \neg\phi$ must be the intersection of the most "epistemically important" elements of $B \perp \neg\phi$. Gärdenfors' interpretation of $\geq$ does not cause any incoherence, but he leaves us with some unanswered questions. First, how should we understand the concept that Gärdenfors refers to as epistemic importance? Second, why *should* the contracted belief set be the intersection of the epistemically most important candidates? That is, why are the concepts of belief contraction and epistemic importance normatively related that way? Plausible answers to these questions are required if we want to use Gärdenfors' interpretation of $\geq$ to defend Premise (I) and, ultimately, to argue for the putative conclusion listed above.

So there are a number of issues to address if we want to take seriously the construction of partial meet contraction and turn it into an explicit argument. For more on how we may take partial meet contraction seriously, see Gärdenfors (1984), Levi (2004), and Arló-Costa and Levi (2006).

6.3 *Argument from Construction: Learning-Theoretic Analysis*

Let us examine another technique for constructing belief revision theories: learning-theoretic analysis. Recall that this construction selects belief revision strategies according to some decision rule or achievabilist thesis (Section 4.6). This suggests the following argument schema, where T is a formal theory of belief revision, i.e. a set of revision strategies.

> PREMISE (I). J (as a decision rule or achievabilist thesis) judges a strategy to be not "OK" if that strategy is not in T.

> PREMISE (II). If J judges a strategy to be not "OK", then that strategy is not rational (or epistemically justified, or the like).

> PUTATIVE CONCLUSION. Therefore, a strategy is rational (or epistemically justified, or the like) only if it is in T.

A candidate for J is the Dominance principle, as mentioned in Section 4.6. When outcomes, or learning performances, are specified in greater detail, it is likely that only very few strategies are dominated. Indeed, a general feature of the Dominance principle is that it becomes weaker when outcomes are specified in greater detail. So, in general, it would be difficult to use the dominance principle to argue for a strong normative thesis.

In that case, one might consider resorting to another candidate for J mentioned in Section 4.6: the Maximin decision rule. But there is a longstanding worry that, in many situations, the Maximin rule is too pessimistic to be the right rule to apply. Indeed, the dominant view in decision theory is that a correct decision rule has to involve one's degrees of belief over the columns in the decision table, rather than (pessimistically) focusing on the worst possible outcomes. There is a possible response in favor of applying Maximin to *some contexts*. The learning-theoretic analysis is actually developed to address the so-called problem of induction. Namely, it is meant to respond to the inductive skeptic's questions: "How can we justify induction?" "How can we justify inductive strategies rather than skeptical strategies?" "How can we justify the use of a particular inductive strategy rather than an alternative inductive strategy?" To properly address these tough questions, we cannot rely on anyone's degrees of belief over the columns in the decision table, for fear of begging the skeptic's question—or so Lange (2002) argues. So, to make a decision without begging the skeptic's question, the right decision rule, if there is one, has to be a qualitative decision rule. And the Maximin rule seems a good candidate—or so this response suggests and promises to elaborate. This idea, which favors the use of Maximin in some contexts, may be traced at least back to the Maximin foundation of statistical inference due to Wald (1950). Note that those sympathetic to the above line of thought do not have to stick with Maximin but can switch to, and argue for, another qualitative decision rule that does not presuppose degrees of belief. Kelly (2007), for example, proposes a kind of dominance principle that applies to the worst-case bounds of "complexity classes"—a decision rule inspired by how computer scientists evaluate the efficiency of problem-solving algorithms.

There is another kind of candidate for J mentioned in Section 4.6, achievabilist theses, which take this form: "If the empirical problem in question is easy enough to allow a revision strategy to achieve epistemic standard X, then a revision strategy for that empirical problem is "OK" only if it achieves (at least) X." Achievabilist theses have seldom been formulated explicitly in learning theory, and they seem to be central to the way Putnam (1965) and Gold (1967) created formal learning theory in the 1960's—or so Kelly (1996) seems to suggest. But which achievabilist theses are correct and how should they be defended or at least motivated? There have been very few systematic discussions on this issue in the literature, but see Kelly et al. (2016) and Genin and Kelly (2018) for examples of how certain epistemic standards may be motivated and put to use in achievabilist theses.

7 CONCLUDING REMARKS

We have discussed a number of foundational issues about belief revision theory. Let us recap what we have covered. Have a look at the italicized terms below.

> A belief revision theory is meant to make *normative or evaluative*$_{\text{(i)}}$ *claims*$_{\text{(ii)}}$ about revision of beliefs in light of new *information*$_{\text{(iii)}}$.

With respect to (i), we have noted that alternative normative interpretations can be given to a formal belief revision theory, and have seen that the choice among those possible interpretations amounts to the choice among very different research programs in belief revision theory (Section 2.3). With respect to (ii), we have examined some methods that we may use to argue for or against the normative claims that a belief revision theory is intended to make (Section 5 and Section 6), including various potential difficulties or issues that we need to address when trying to apply those argumentative methods. With respect to (iii), we have only briefly discussed the issue of what counts as information and the problem of choosing among different conceptions of information (Section 5.3).

For discussions of other philosophical issues, see Levi (1983), Levi (1991), Levi (2004), Gärdenfors (1988), Rott (2000), Rott (2001), Hansson (1999), Hansson (2003), Gillies (2004), and Genin and Kelly (2018).

8 APPENDIX

8.1 *Nonmonotonic Logic and Belief Revision Theory*

A *nonmonotonic consequence relation* is a binary relation $\mid\!\sim$ between sentences. Understand $\phi \mid\!\sim \psi$ as saying of $\mid\!\sim$ that it licenses the inference from ϕ to ψ—a possibly defeasible, inductive, or plausible inference. Nonmonotonic logic, if broadly construed, aims at distinguishing nonmonotonic consequence relations that are good in one sense or another. There are many approaches to nonmonotonic logic; they differ in the procedures that are used to sort out "good" nonmonotonic consequence relations; see Brewka et al. (2008) for a review.

Makinson and Gärdenfors (1991) propose a translation between simple belief revision strategies S and nonmonotonic consequence relations $\mid\!\sim$.

Their translation is based on the following bridge principle (which I state in terms of the S-notation used here):

$$\psi \in S(\phi) \text{ iff } \phi \mathrel{|\!\sim} \psi.$$

To be more precise: given any simple belief revision strategy S, we can use the bridge principle to define a nonmonotonic consequence relation $\mathrel{|\!\sim}_S$ as follows: $\phi \mathrel{|\!\sim}_S \psi$ iff $\psi \in S(\phi)$. Conversely, given any nonmonotonic consequence relation $\mathrel{|\!\sim}$, we can use the bridge principle to define a simple belief revision strategy $S_{\mathrel{|\!\sim}}$ as follows: $S_{\mathrel{|\!\sim}}(\phi) =_{\text{def}} \{\psi : \phi \mathrel{|\!\sim} \psi\}$ and $S_{\mathrel{|\!\sim}}() =_{\text{def}} S_{\mathrel{|\!\sim}}(\top)$, where $\top$ is a tautology.

This establishes a one-to-one correspondence between all nonmonotonic consequence relations and all simple belief revision strategies S such that $S() = S(\top)$.

8.2 *General Definition of Learning*

Let an information space $\mathcal{I}$ be given, which contains some finite sequences of sentences, meant to represent possible *available* pieces of information. Let a question $\mathcal{Q}$ be identified with a set of mutually incompatible sentences, called the *potential answers* to $\mathcal{Q}$. The potential answers to $\mathcal{Q}$ may, or may not, be jointly exhaustive—let the disjunction of the potential answers to $\mathcal{Q}$ be understood as the *presupposition* of question $\mathcal{Q}$. Let a decision table be given, together with a set $\mathcal{C}$ of columns as mutually incompatible possibilities. Those columns/possibilities are assumed to be so specific that each column $C \in \mathcal{C}$ either entails exactly one potential answer to question $\mathcal{Q}$ or it entails the negation of Q's presupposition. With respect to the above setting $(\mathcal{Q}, \mathcal{I}, \mathcal{C})$, define the following concepts.

- An *$\mathcal{I}$-information stream* is an infinite sequence $(\phi_1, \phi_2, \ldots)$ of sentences such that its finite initial segments are all in $\mathcal{I}$.
- Say that an $\mathcal{I}$-information stream $(\phi_1, \phi_2, \ldots)$ is *compatible* with a column $C \in \mathcal{C}$ iff the infinite conjunction $\bigwedge_{i \geq 1} \phi_i$ is compatible with possibility C.
- The *true answer* to question $\mathcal{Q}$ given column $\mathcal{C}$, written $\mathsf{Ans}(\mathcal{Q} \mid \mathcal{C})$, is defined as the unique potential answer to $\mathcal{Q}$ that C entails, if such a unique answer exists; otherwise, $\mathsf{Ans}(\mathcal{Q} \mid \mathcal{C})$ is undefined.

We are finally in a position to define learning with respect to the above setting.

- Say that a strategy S *will learn* the true answer to question Q given column C just in case:

(1) the true answer $\mathsf{Ans}(\mathcal{Q} \mid \mathcal{C})$ exists;

(2) for each $\mathcal{I}$-information stream $(\phi_1, \phi_2, \ldots)$ compatible with C, there exists $n \geq 1$, called a "learning moment," such that for each $i \geq n$, $S(\phi_1, \phi_2, \ldots, \phi_i)$ is consistent and entails $\mathsf{Ans}(\mathcal{Q} \mid \mathcal{C})$.

ACKNOWLEDGEMENTS

I am indebted to Kevin Kelly and Horacio Arló-Costa for teaching me so much about belief revision theory. This article would have been impossible without them. I am indebted to the two editors, Jonathan Weisberg and Richard Pettigrew, for their endless patience and very helpful comments. I am also indebted to Ted Shear for his incredibly detailed comments and useful suggestions, and to Kasey Genin, Jake Chandler, Adam Sennet, and Rachel Boddy for their stimulating questions.

REFERENCES

Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *The journal of symbolic logic*, *50*(2), 510–530.

Arló-Costa, H. & Levi, I. (2006). Contraction: On the decision-theoretical origins of minimal change and entrenchment. *Synthese*, *152*(1), 129–154.

Arló-Costa, H. & Pedersen, A. P. (2012). Belief and probability: A general theory of probability cores. *International Journal of Approximate Reasoning*, *53*(3), 293–315.

Baltag, A., Gierasimczuk, N., & Smets, S. (2016). On the solvability of inductive problems: A study in epistemic topology. *arXiv preprint arXiv:1606.07518, presented at the 2015 TARK.*

Booth, R. & Chandler, J. (2017). The irreducibility of iterated to single revision. *Journal of Philosophical Logic*, *46*(4), 405–418.

Boutilier, C. (1996). Iterated revision and minimal change of conditional beliefs. *Journal of Philosophical Logic*, *25*(3), 263–305.

Brewka, G., Niemelä, I., & Truszczyński, M. (2008). Nonmonotonic reasoning. *Foundations of Artificial Intelligence*, *3*, 239–284.

Darwiche, A. & Pearl, J. (1997). On the logic of iterated belief revision. *Artificial intelligence*, *89*(1-2), 1–29.

Gärdenfors, P. (1984). Epistemic importance and minimal changes of belief. *Australasian Journal of Philosophy*, *62*(2), 136–157.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. The MIT press.

Gärdenfors, P. & Makinson, D. (1988). Revisions of knowledge systems using epistemic entrenchment. In *Proceedings of the 2nd conference on theoretical aspects of reasoning about knowledge* (pp. 83–95). Morgan Kaufmann Publishers Inc.

Genin, K. (2019). Full & partial belief. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*. PhilPapers.

Genin, K. & Kelly, K. T. (2018). Theory choice, theory change, and inductive truth-conduciveness. *Studia Logica*. doi:10.1007/s11225-018-9809-5

Gillies, A. S. (2004). Epistemic conditionals and conditional epistemics. *Noûs*, *38*(4), 585–616.

Ginsberg, M. L. (1986). Counterfactuals. *Artificial intelligence*, *30*(1), 35–79.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, *10*(5), 447–474.

Grove, A. (1988). Two modellings for theory change. *Journal of philosophical logic*, *17*(2), 157–170.

Hansson, S. O. (1994). Taking belief bases seriously. In *Logic and philosophy of science in uppsala* (pp. 13–28). Springer.

Hansson, S. O. (1999). *A textbook of belief dynamics*. Springer Science & Business Media.

Hansson, S. O. (2003). Ten philosophical problems in belief revision. *Journal of logic and computation*, *13*(1), 37–49.

Hansson, S. O. (2017). Logic of belief revision. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2017). Metaphysics Research Lab, Stanford University.

Harper, W. L. (1975). Rational belief change, popper functions and counterfactuals. *Synthese*, *30*(1-2), 221–262.

Harper, W. L. (1976). Rational conceptual change. In *Psa: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1976, *2*, pp. 462–494). Philosophy of Science Association.

Huber, F. (2013a). Belief revision i: The agm theory. *Philosophy Compass*, *8*(7), 604–612.

Huber, F. (2013b). Belief revision ii: Ranking theory. *Philosophy Compass*, *8*(7), 613–621.

Huber, F. (2019). Ranking theory. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*. PhilPapers.

Jin, Y. & Thielscher, M. (2007). Iterated belief revision, revised. *Artificial Intelligence*, *171*(1), 1–18.

Katsuno, H. & Mendelzon, A. O. (2003). On the difference between updating a knowledge base and revising it1. *Belief revision*, *29*, 183.

Kelly, K. T. (1996). *The logic of reliable inquiry*. Oxford University Press.

Kelly, K. T. (1999). Iterated belief revision, reliability, and inductive amnesia. *Erkenntnis*, *50*(1), 7–53.

Kelly, K. T. (2007). How simplicity helps you find the truth without pointing at it. In *Induction, algorithmic learning theory, and philosophy* (pp. 111–143). Springer.

Kelly, K. T., Genin, K., & Lin, H. (2016). Realism, rhetoric, and reliability. *Synthese*, *193*(4), 1191–1223.

Kraus, S., Lehmann, D., & Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial intelligence*, *44*(1-2), 167–207.

Lange, M. (2002). Okasha on inductive scepticism. *The Philosophical Quarterly*, *52*(207), 226–232.

Leitgeb, H. (2014). The stability theory of belief. *The Philosophical Review*, *123*(2), 131–171.

Levi, I. (1978). Subjunctives, dispositions and chances. In *Dispositions* (pp. 303–335). Springer.

Levi, I. (1983). *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT press.

Levi, I. (1991). *The fixation of belief and its undoing: Changing beliefs through inquiry*. Cambridge University Press.

Levi, I. (2004). *Mild contraction: Evaluating loss of information due to loss of belief*. Oxford University Press on Demand.

Lin, H. & Kelly, K. T. (2012). Propositional reasoning that tracks probabilistic reasoning. *Journal of philosophical logic*, *41*(6), 957–981.

Makinson, D. (1988). General theory of cumulative inference. In *International workshop on non-monotonic reasoning* (pp. 1–18). Springer.

Makinson, D. & Gärdenfors, P. (1991). Relations between the logic of theory change and nonmonotonic logic. In *The logic of theory change* (pp. 183–205). Springer.

Nayak, A. C. (1994). Iterated belief change based on epistemic entrenchment. *Erkenntnis*, *41*(3), 353–390.

Nute, D. (2012). *Defeasible deontic logic*. Springer Science & Business Media.

Putnam, H. (1965). Trial and error predicates and the solution to a problem of mostowski. *The Journal of Symbolic Logic*, *30*(1), 49–57.

Quine, W. V. O. (1982). *Methods of logic*. Harvard University Press.

Reiter, R. (1980). A logic for default reasoning. *Artificial intelligence*, *13*(1-2), 81–132.

Rodrigues, O., Gabbay, D., & Russo, A. (2011). Belief revision. In *Handbook of philosophical logic* (pp. 1–114). Springer.

Rott, H. (1993). Belief contraction in the context of the general theory of rational choice. *The Journal of Symbolic Logic*, *58*(4), 1426–1450.

Rott, H. (2000). Two dogmas of belief revision. *The Journal of Philosophy*, *97*(9), 503–522.

Rott, H. (2001). *Change, choice and inference: A study of belief revision and nonmonotonic reasoning*. Clarendon Press.

Schulte, O. (1999). Means-ends epistemology. *The British Journal for the Philosophy of Science*, *50*(1), 1–31.

Shoham, Y. (1987). A semantical approach to nonmonotonic logics. In *Readings in nonmonotonic reasoning* (pp. 227–250). Morgan Kaufmann Publishers Inc.

Spohn, W. (1988). Ordinal conditional functions: A dynamic theory of epistemic states. In W. L. Harper & B. Skyrms (Eds.), *Causation in decision, belief change and statistics: Proceedings of the irvine conference on probability and causation* (pp. 105–134). The University of Western Ontario Series in Philosophy of Science. Kluwer.

Stalnaker, R. (1994). What is a nonmonotonic consequence relation? *Fundamenta Informaticae*, *21*(1, 2), 7–21.

Stalnaker, R. (2009). Iterated belief revision. *Erkenntnis*, *70*(2), 189–209.

Titelbaum, M. G. (2019). Precise credences. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*. PhilPapers.

Wald, A. (1950). Statistical decision functions.