THE OPEN HANDBOOK OF FORMAL EPISTEMOLOGY

RICHARD PETTIGREW & JONATHAN WEISBERG, EDS.

THE OPEN HANDBOOK OF FORMAL EPISTEMOLOGY

Richard Pettigrew & Jonathan Weisberg, Eds.

May 2019

LIST OF CONTRIBUTORS

R. A. Briggs
Stanford University

Michael Caie
University of Toronto

Kenny Easwaran
Texas A&M University

Franz Huber
University of Toronto

Jason Konek
University of Bristol

Hanti Lin
University of California, Davis

Anna Mahtani
London School of Economics

Konstantin Genin
University of Toronto

Johanna Thoma
London School of Economics

Michael G. Titelbaum
University of Wisconsin, Madison

Sylvia Wenmackers
Katholieke Universiteit Leuven

An epigraph: something pithy, and surprisingly apt to the context if you just stop and think about it a moment. — *Someone Famous*

For whosits

ACKNOWLEDGMENTS

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

CONTENTS

1

PRECISE CREDENCES

Michael G. Titelbaum

I am more confident than not that I will go in to my office tomorrow. I'm not *certain* that I will go, and I haven't even hit the point of *believing* that I will: it is the summer, I have no courses to teach or students to meet, I may wake up tomorrow and decide it's not worth the effort. But I'm more confident that I will go than I am that I won't. If I had to place my confidence on a scale of 0 to 100, I'd put it somewhere above 50.

Credences are numerical degrees of confidence. While they could be expressed as percentages—between 0 to 100, inclusive—it has become customary to measure them on a scale from 0 to 1. Credences are also often called "degrees of belief," though that name may hold the connotation that they are a species of ordinary, qualitative belief.

It's better to think of credence not as a kind of qualitative belief, but instead as a member of the same family as qualitative belief. That family—the family of doxastic attitudes—also includes certainty, disbelief, suspension of belief, and probably comparative confidence as well. The members of this family have a variety of commonalities. For example, we tend to think of credences as taking the same sorts of *objects* as outright beliefs. Many authors take these objects to be propositions, and so classify both credences and beliefs as propositional attitudes. I will follow that trend here, but if you think beliefs are adopted towards something other than propositions (sentences, perhaps?), you will be inclined to the same view about credences.

The theory of credences was developed to address a number of philosophical problems. One was the proper interpretation of "probability" locutions. If I say, "The probability that I'll go to the office tomorrow is over 50%," what does this mean, and what are the truth-conditions for my utterance? A number of interpretations of probability have been offered and defended (some of which we will discuss in Section 1.6), and it's not clear that every use of the term "probability" should be interpreted the same way. But one prominent suggestion, the "subjective interpretation of probability," is that probability statements express the speaker's degree of confidence in a proposition. So my utterance expresses a confidence over 0.5 that I shall go to the office.

Yet even if "probability" statements rarely—or never—express an agent's degrees of confidence, such degrees of confidence may still exist, and have philosophical work to do. Degrees of belief play a prominent role in traditional decision theory, the classic formal approach to rational choice

(about which more in Section 2.2). Credences also figure in Bayesian confirmation theory (Section 2.1), an account of evidential support rivaling other statistical approaches such as frequentism and likelihoodism. And they can be applied to such further topics as coherentism, Inference to the Best Explanation, and social epistemology (Section 2.3).

So if we grant that credences exist, what exactly does it take to possess one? In line with contemporary behaviorist approaches in psychology, de Finetti (1937/1964) *defined* the degree of belief assigned to an event by an individual as the rate at which she'd bet that it would occur (more about the details in Section 2.2). But as was typical with operationalism, this definition ran into problems when, say, an agent displayed inconstant betting behaviors over time, and so was difficult to assign a particular credence. Nowadays we may grant than an agent with a particular degree of belief will, *if rational*, display particular betting behavior (Christensen, 2004). But we also tend to think of this normative connection less as a *definition* of credence and more as one aspect of what it is to possess a degree of confidence. Just as our account of qualitative belief has progressed beyond behaviorism to a broader functionalism, we think of credence as a multi-faceted mental state with descriptive and normative connections to a wide variety of behaviors and other attitudes.

Besides their connections to desires, intentions, and decisions contemplated in action theory and decision theory, credences are connected to other varieties of doxastic attitudes (not to mention emotions, sensations, and memories). If comparative confidence is a distinct type of mental state, it clearly is connected to credence: I am more confident of P than Q just in case my credence in P is higher than my credence in Q. As for qualitative attitudes, certainty is often identified with credence 1 in a proposition (though see Section 1.7 below). There must also be links between credence and outright belief: if I believe P, my credence in P should be higher than my credence in $\sim P$.

Can we find a fully general connection between credence and outright belief? Some authors (e.g., Holton, 2014) maintain that to the extent there are any credences, to possess credence x in P is just to hold an outright belief that the probability of P is x. Yet it's difficult to find a single concept of probability that applies to every proposition to which an agent might assign a degree of belief. And it seems agents (such as children) can be more or less confident of propositions without possessing a concept of probability. Moreover, whatever concept of probability we select, it seems conceivable for an agent to adopt a degree of confidence in the proposition that P has probability x. (We'll see a further technical difficulty with the credence-as-outright-belief theory in Section 1.2.) Most theorists now hold that the numerical value of a credence is an attribute of the attitude

adopted *towards* a proposition, not part of the *content* of the proposition towards which that attitude is adopted.[1]

Going in the other direction, the "Lockean Thesis"[2] takes outright belief just to be credence above a particular threshold. The threshold credence is usually lower than 1 (belief need not be certainty) but well above 1/2, and may depend on contextual parameters. The main objection to the Lockean Thesis is that one can describe rationally acceptable credence distributions which, by way of the thesis, generate rationally-unacceptable patterns of belief. In the Lottery Paradox (Kyburg, 1961) an agent assigns to each ticket in a lottery a low credence that it will win, while assigning a high credence (perhaps certainty) that some ticket will win. For any Lockean threshold less than 1, we can arrange the numbers so that the agent winds up believing of each ticket that it will lose, while believing that some ticket will win—a logically inconsistent overall set of beliefs. Similarly, in the Preface Paradox (Makinson, 1965), an author has high confidence in each claim made in her book while also being confident that at least one of those claims is false. Via the Lockean thesis this becomes belief in each conjunct of a conjunction coupled with disbelief in that conjunction.

How, then, to relate credence and outright belief in general? The most radical possibility is to deny either the existence of beliefs or the existence of credences. More conservatively, one could offer a reduction of one category to the other, or at least a principle of descriptive supervenience. Alternatively, one could grant that while beliefs and credences appear in a variety of configurations in *actual* agents, normative principles specify how they'd align in a *rational* agent. The current consensus is that something beyond just the Lockean Thesis would be required to make either of these approaches work; recent attempts to articulate belief-credence principles can be found in Leitgeb (2017), Douven (2012), and Lin and Kelly (2012).

On the other hand, one could concede that beliefs and credences are both genuine kinds of mental states an agent can possess, there are some ways in which they interact (or interact if one is rational), but no systematic general principles are available. While this stance is available to strong realists about beliefs and credences, it is especially attractive to theorists who read belief and credence ascriptions as convenient, simplifying models of a highly complex cognitive system. The belief-model and the credence-model are each effective and efficient in different circumstances, and may be applied toward different ends. In that case, it would be unsurprising if no universal translation from one to the other were available.

1 Moss (2018) takes the numerical value to be part of a credence's content, but takes credal objects to be more complicated than simple propositions.

2 Locke (1689/1975, Bk. IV). See also Foley (1993) for discussion.

1 RATIONAL CONSTRAINTS ON CREDENCE

Once we understand what a credence is, the next question is what it takes for a set of credences to be rational.

1.1 *The Probability Axioms*

The most generally-accepted rational credence norms are Kolmogorov's (1933/1950) axioms. Suppose we have a language $\mathcal{L}$ of propositions, which starts with a finite set of atomic propositions and then closes them under the standard truth-functional connectives. Define a real-valued function c over $\mathcal{L}$ representing the credence values an agent assigns the propositions in $\mathcal{L}$.[3] The precise, real-number values that c assigns each proposition are the "precise credences" of this entry's title; I'll discuss alternative formal approaches in Section 5 below.

Given this setup, Kolmogorov's axioms become the following.

NON-NEGATIVITY. For any $X \in \mathcal{L}$, $c(X) \geq 0$.

NORMALITY. For any tautology $\top \in \mathcal{L}$, $c(\top) = 1$.

FINITE ADDITIVITY. For any mutually exclusive $X, Y \in \mathcal{L}$,
$c(X \vee Y) = c(X) + c(Y)$.

Mathematicians often call these *the probability axioms*, and call any distribution satisfying them a *probability function*. Probabilism is the position that rational credences form a probability function; in other words, rational credences satisfy the Kolmogorov axioms.[4]

The probability axioms set $0 \leq c(X) \leq 1$ for every $X \in \mathcal{L}$. Probabilism also entails a number of intuitive constraints on rational credence. For example:

- for any $X \in \mathcal{L}$, $c(\sim X) = 1 - c(X)$.

Suppose you assign a high confidence that anthropogenic global warming has occurred. This constraint requires you to assign a low confidence that no anthropogenic warming has occurred. And should you become more confident that anthropogenic warming has occurred, this constraint

3 While I will consider languages containing propositions, other authors describe credences as distributed over sentences, or sets of possible worlds, or sets of events, etc.

4 Probabilism is often described as the doctrine that rational *agents* have credences satisfying the probability axioms, or (if that's considered too unrealistic) that *ideally rational agents* have probabilistic credences. Both of these formulations make agents (real or ideal) the targets of evaluation. Strictly speaking, I prefer to evaluate credences (or sets of credences) for rationality, rather than agents. But for ease of locution I will largely treat the two as interchangeable here.

will require your confidence in that proposition's negation to decrease accordingly.

Some other intuitive constraints following from the Kolmogorov axioms.

- For any contradiction $\mathsf{F} \in \mathcal{L}$, $c(\mathsf{F}) = 0$.
- For any $X, Y \in \mathcal{L}$ (mutually exclusive or otherwise),
$$c(X \vee Y) = c(X) + c(Y) - c(X \,\&\, Y).$$
- For any $X, Y \in \mathcal{L}$, if $X \vDash Y$ then $c(Y) \geq c(X)$.
- For any logically equivalent $X, Y \in \mathcal{L}$, $c(X) = c(Y)$.
- For any finite set of mutually exclusive $X_1, \ldots, X_n \in \mathcal{L}$,
$$c(X_1 \vee \ldots \vee X_n) = c(X_1) + \ldots + c(X_n).$$

The last bulleted constraint has an important consequence when an agent considers a *partition*—a set of propositions whose members are mutually exclusive and jointly exhaustive. Because the disjunction of a partition's elements is a tautology, probabilism demands that the credences assigned to elements of a partition sum to 1.

A further important consequence of probabilism is that credences are strongly extensional. If an agent is certain that two propositions X and Y have the same truth-value (that is, if $c(X \equiv Y) = 1$), then for the sake of calculating credences X and Y might as well be logically equivalent. For instance, any credence equation or inequality in which X appears would remain true were any of its Xs replaced with Ys. Any difference in meaning, modal profile, etc. is irrelevant to probability once truth-values are established to be identical.

We can illustrate probabilism with Kyburg's Lottery example from page 3. Given a lottery with, say, 100 tickets, introduce a language whose atomic propositions are W_1 through W_{100} (with W_i indicating that ticket i wins the lottery). If the lottery is fair, an agent might assign $c(W_i) = 1/100$ for each W_i. From our first intuitive consequence of the probability axioms, we then have $c(\sim W_i) = 99/100$; the agent is highly confident of each ticket that it will not win. However, assuming no more than one ticket can win, our final intuitive consequence listed above yields:

$$c(W_1 \vee \ldots \vee W_{100}) = c(W_1) + \ldots + c(W_{100}) = 1. \tag{1}$$

So our agent is certain some ticket will win, as intuitively she ought to be.[5]

5 Notice that none of this solves the Lottery *Paradox*, which brings full beliefs into the lottery picture. My goal is just to illustrate how probabilism is compatible with and supportive of a natural account of rational credences in the lottery case. A similar illustration could be given for Makinson's Preface example.

While proofs in the probability calculus usually proceed from Kolmogorov's axioms, practical-problem solving is often made easier by working with state-descriptions. Define a literal to be an atomic sentence of $\mathcal{L}$ or its negation, then define a state-description in $\mathcal{L}$ to be a maximal consistent conjunction of its literals. Every noncontradictory $X \in \mathcal{L}$ then has a unique disjunctive normal form, a disjunction of state-descriptions logically equivalent to X.[6]

Carnap (1950) makes repeated use of the fact that a distribution c over $\mathcal{L}$ satisfies the probability axioms just in case it assigns: (1) nonnegative values to $\mathcal{L}$'s state-descriptions summing to 1; (2) for every noncontradictory X, a value equal to the sum of the values assigned to the state-descriptions in X's disjunctive normal form; and (3) a value of 0 to every contradictory proposition.[7]

This result is handy in two ways. First, we can completely characterize any probability distribution over $\mathcal{L}$ by specifying the values it assigns to $\mathcal{L}$'s state-descriptions. Second, given partial information about a probability distribution, we can determine what this information says about the values assigned to state-descriptions, then from there work out the values of (or constraints on the values of) other propositions.

For example, suppose I tell you that Bob is certain of $P \supset Q$, and is twice as confident of P as $\sim P$. It immediately follows that Bob's confidence in $\sim Q$ is less than or equal to $1/3$. Why? Well, the disjunctive normal form equivalent of $\sim Q$ is $(P \,\&\, \sim Q) \vee (\sim P \,\&\, \sim Q)$. Since Bob is certain of $P \supset Q$, the first disjunct receives credence 0, so for Bob $c(\sim Q) = c(\sim P \,\&\, \sim Q)$. But since $c(P) + c(\sim P) = 1$, and $c(P) = 2 \cdot c(\sim P)$, we have $c(\sim P) = 1/3$. The disjunctive normal form equivalent of $\sim P$ is $(\sim P \,\&\, Q) \vee (\sim P \,\&\, \sim Q)$. By Non-Negativity Bob's credence in the first disjunct must be greater than or equal to 0, so the second disjunct receives a credence less than or equal to $1/3$.[8]

Finally, with the notion of a probability function in hand we can define the notion of an *expectation*. Suppose we have a numerical quantity for which many values are possible. To calculate an agent's expectation for that quantity, we multiply each value times the agent's credence that the quantity will take that value, then sum over all the values available. For example, if I'm 10% confident that I'll go into my office two days this

6 To make the disjunctive normal form unique, we require literals to appear in a state-description in some canonical order (perhaps alphabetical, if the propositions are designated by letters), and then we require state-descriptions to appear in disjunctive normal forms in a canonical order as well.

7 I have never been able to discover whether this result was original to Carnap or not. I would sincerely welcome any e-mails demonstrating its historical provenance!

8 For more on the mathematical theory underlying this approach, and for a Mathematica routine that will solve many probability problems once they are reduced to algebra using state-descriptions, see Fitelson (2008).

week, 60% confident that I'll go in just one day, and 30% confident that I won't go in at all, then my expectation for the numbers of days I'll go into my office this week is:

$$0.10 \cdot 2 \text{ days} + 0.60 \cdot 1 \text{ day} + 0.30 \cdot 0 \text{ days} = 0.8 \text{ days}. \tag{2}$$

1.2 *The Ratio Formula*

So far we have discussed unconditional credence—an agent's degree of confidence that a particular proposition is true in light of her current understanding of what the world is like. We may also inquire after an agent's *conditional* credence in proposition X given Y; this is the agent's credence in X upon making the additional assumption that Y. Notice that Y may be a proposition in which the agent currently has low unconditional credence. In asking for her credence in X given Y, we ask her to set aside her current actual opinion about Y, temporarily add Y to the stock of propositions she takes to be true, then assess X in light of this enhanced suppositional set.[9]

An agent's conditional credence in X given Y is denoted $c(X \mid Y)$, and is usually taken to be governed by the Ratio Formula.

Ratio Formula. For any $X, Y \in \mathcal{L}$ with $c(Y) > 0$,

$$c(X \mid Y) = \frac{c(X \,\&\, Y)}{c(Y)}.$$

The Ratio Formula can be read as either a descriptive truth or as a normative requirement. On the former approach, an agent's conditional credence X given Y takes a particular value just in case her unconditional credences in $X \,\&\, Y$ and Y stand in that ratio. This reading is most natural if one wants to reduce one type of credence to the other: one could hold that to have a conditional credence just is to have unconditional credences standing in a particular ratio; or one could hold that conditional credences are basic and unconditional credences are a proper subset of those.[10] Alternatively, one could see conditional credence as just another type of doxastic attitude on equal footing with unconditional credences, then read the Ratio Formula

9 Notice that we are discussing indicative, not subjunctive, conditional credences. The supposition Y is to be *added* to the agent's current set of assumptions about the world, with the resulting suppositional set assumed to be consistent. Most discussions of conditional credence concern the indicative form. For a treatment of subjunctive conditional credences, see Joyce (1999).

10 From the Kolmogorov axioms and Ratio Formula, it follows that for any $X \in \mathcal{L}$, $c(X) = c(X \mid \mathsf{T})$. So unconditional credences can be thought of as conditional credences conditional on a tautology. See Easwaran (this volume) for more.

as a rational requirement on how conditional and unconditional credences should align.[11]

Note that as I've defined the Ratio Formula, it remains silent when the agent assigns the condition (proposition Y) a credence of 0. We will return to credences conditional on credence-0 propositions in Section 1.7.

Combining the Ratio Formula and Kolmogorov's Axioms yields the handy Law of Total Probability.

LAW OF TOTAL PROBABILITY. For any $X, Y_1, \ldots, Y_n \in \mathcal{L}$ such that the $Y_1, \ldots, Y_n$ form a finite partition,

$$c(X) = c(X \mid Y_1) \cdot c(Y_1) + \ldots + c(X \mid Y_n) \cdot c(Y_n).$$

The Law of Total Probability calculates the unconditional credence of X as a weighted average of X's credences conditional on members of the Y-partition, weighted by the unconditional credences in the Ys.[12]

To illustrate once more with our lottery scenario, suppose B is the proposition that our agent will benefit from the outcome of the lottery. She holds tickets 1 through 3, so is sure to benefit if they win. Also, her sister holds the very last ticket (ticket 100), and the agent is 1/2 confident that her sister will share the winnings should that ticket come in. Applying the Law of Total Probability (and recalling that W_i is the proposition that ticket i will win), the agent's credence that she will benefit is

$$\begin{aligned} c(B) &= c(B \mid W_1) \cdot c(W_1) + c(B \mid W_2) \cdot c(W_2) + c(B \mid W_3) \cdot c(W_3) \\ &\quad + c(B \mid W_4) \cdot c(W_4) + \ldots + c(B \mid W_{100}) \cdot c(W_{100}) \\ &= 1 \cdot 1/100 + 1 \cdot 1/100 + 1 \cdot 1/100 \\ &\quad + 0 \cdot 1/100 + \ldots + 1/2 \cdot 1/100 \\ &= 0.035. \end{aligned} \tag{3}$$

Conditional credence also plays a crucial role in the notion of credal relevance. When $0 < c(Y) < 1$, all of the following inequalities are equivalent:

$$c(X \mid Y) > c(X), \tag{4}$$
$$c(X) > c(X \mid \sim Y), \tag{5}$$
$$c(Y \mid X) > c(Y), \tag{6}$$
$$c(Y) > c(Y \mid \sim X), \tag{7}$$
$$c(X \,\&\, Y) > c(X) \cdot c(Y). \tag{8}$$

11 For a discussion of how conditional credences interact with an agent's credences in conditionals, see Briggs (this volume).

12 Put another way, the Law of Total Probability requires an agent's unconditional credence in X to equal her expectation of her credence in X conditional on the true element of the Y-partition.

When these inequalities hold, we say that Y is positively relevant to X on the agent's credence function. (Since positive relevance is a symmetric relation, we may also say that X is positively relevant to Y.) Another way to put this is that the agent takes X and Y to be positively correlated. Replacing the greater-thans with less-thans describes when Y is negatively relevant to X (or negatively correlated with X) on an agent's credences. On the other hand, when $c(X \,\&\, Y) = c(X) \cdot c(Y)$ (or any of the other inequalities above becomes equality), we say that X is irrelevant to Y for the agent, or probabilistically independent of Y.

These relevance relations are relative to an agent's credences; they reflect which propositions she assesses as relevant to each other given her current understanding of the world. But we can also temporarily enhance her current set of suppositions about the world, and see whether any relevance relations change. This takes us from a notion of unconditional relevance to conditional relevance. Y is relevant to X conditional on Z just in case

$$c(X \mid Y \,\&\, Z) > c(X \mid Z). \tag{9}$$

For each of the inequalities above, a corresponding characterization of conditional relevance can be given by adding Z as a condition to the expression on each side.

The notion of conditional relevance underlies a crucial notion in the philosophy of science: screening off. We say that Z screens off X from Y when X and Y are unconditionally dependent but the following two equalities hold:

$$c(X \mid Y \,\&\, Z) = c(X \mid Z), \tag{10}$$
$$c(X \mid Y \,\&\, \sim Z) = c(X \mid \sim Z). \tag{11}$$

In other words, X and Y are independent conditional on each of Z and $\sim Z$. In a screening-off situation, supposing either Z or $\sim Z$ makes the correlation between X and Y disappear.[13]

To illustrate one application of this concept, Reichenbach (1956) argues that a common cause screens off its effects from each other. Suppose X is the proposition that my newspaper reports that the Yankees won last night, Y is the proposition that your newspaper reports that the Yankees won last night, and Z is the proposition that the Yankees actually won. On the one hand, while I remain ignorant of Z it would be rational for me to treat X as relevant to Y. X provides information about Z, and therefore also provides information about Y. But once the truth-value of Z is established, X and Y lose the ability to say anything about each other; X and Y become

13 This definition generalizes to the case in which Z is a random variable capable of taking a variety of values z_i. Screening off then occurs when X and Y are unconditionally correlated, but become independent conditional on each proposition of the form $Z = z_i$.

independent conditional on any supposition about Z. Thus Z will screen off X from Y on my credence function.

A proximal cause will also screen off its effect from a distal cause. (Imagine Y is the final score of last night's Yankees game, Z is the proposition that the Yankees won, and X is the proposition that my newspaper reports that they won.) In general, probabilistic correlations (conditional and unconditional) can provide useful evidence about the causal relations among a set of variables. Some philosophers have even *defined* causality in terms of probabilistic relations. For more on all of this, see Hitchcock (2012).

One final point about conditional credences. Earlier (p. 2) I mentioned the theory that a credence of x in P is just the outright belief that the probability of P is x. There I noted a number of problems for that theory; now we can add that the theory seems to lack a good way of understanding conditional credence. A conditional credence in $c(P \mid Q)$ of x cannot be read as a qualitative belief in the proposition "If Q, then the probability of P is x," nor can it be read as the belief that "The probability of 'If Q, then P' is x." This was established by a series of triviality results initiated by Lewis (1976).[14] For instance, Lewis' work shows that if we assume $c(P \mid Q) = x$ just in case $p(Q \rightarrow P) = x$ for some suitable notion of probability p and some indicative conditional $\rightarrow$, then it follows that every proposition is probabilistically independent from every other! This is obviously absurd. A conditional credence just isn't a credence—or a belief—about a conditional.

1.3 *Updating by Conditionalization*

The rational constraints on credence listed to this point have been synchronic—when they relate multiple credences, all the credences related are held at the same time. The degree of belief literature has also proposed a number of diachronic constraints, governing relations among credences assigned at different times.

Suppose we have two times, t_i and t_j, with the latter occurring after the former. Let c_i and c_j be the agent's credences functions at these two times. The most traditional, well-established, and well-known diachronic credal constraint is Conditionalization.

CONDITIONALIZATION. If $E \in \mathcal{L}$ represents everything the agent learns between t_i and t_j, then for any $X \in \mathcal{L}$, $c_j(X) = c_i(X \mid E)$.

The intuitive idea of Conditionalization is simple. Suppose that at t_i you don't know whether E is true. I ask you to hypothetically suppose E (temporarily add it to your stock of assumptions about what the world is like), then ask for your conditional credence in X given this supposition.

14 For the recent state of the art in this area, see Hájek (2011) and Fitelson (2015).

You offer some number. Then, between t_i and t_j, you learn that E is actually true (and learn nothing else besides). If I now ask you at t_j for your unconditional credence in X, it seems you should offer the same number you reported as a conditional credence before. After all, the set of real-world conditions against which you're assessing X is the same at both times; it's just that at t_i you were supposing E as a fact about the world, while at t_j you know E to be true.

Conditionalization integrates nicely with our other credal constraints. For instance, if c_i satisfies the Kolmogorov axioms and $c_i(E) > 0$, then conditionalizing yields a c_j distribution that satisfies the axioms as well. So if an agent begins with a probability distribution and repeatedly updates by conditionalizing, she is guaranteed to respect probabilism on an ongoing basis. The probability axioms and Ratio Formula also make updating by conditionalization cumulative and commutative. If you conditionalize successively on E and then E', this yields the same result as conditionalizing just once on $E \,\&\, E'$, which means it also yields the same result as conditionalizing on E' followed by E.

For a conditionalizing agent, current credences interact in an interesting way with predictions about future credences. Suppose an agent is certain at t_i that her t_j credences will be formed by conditionalizing on a proposition she will learn from some particular finite partition. (Perhaps she will conduct an experiment between t_i and t_j, and the propositions in the partition represent all of its possible outcomes.) Assuming she meets a few other plausible side-conditions, such an agent will satisfy the Reflection Principle.

Reflection Principle. For any $X \in \mathcal{L}$, $c_i(X \mid c_j(X) = r) = r$.

This principle, introduced by van Fraassen (1984), sets the agent's t_i unconditional credence in X equal to her t_i expectation of her unconditional t_j credence in X.[15] Notice that although a c_j appears in the righthand expression, the principle governs *synchronic* credal interactions: it relates the agent's c_i credences in X to her c_i credences about her future credences in X. Given (again) a few side-conditions, Reflection may be derived from the Kolmogorov axioms, the Ratio Formula, and the agent's certainty that she will update by conditionalizing on some member of a particular partition. Van Fraassen, however, argues in the opposite direction: he provides independent motivation for Reflection, then views Conditionalization as a derivable consequence. For more on the arguments in each direction, and the specific side-conditions required, see Weisberg (2007) and Briggs (2009).

15 To see why, return to our formulation of the Law of Total Probability on page 8, and let each Y_i there assert that the agent's unconditional t_j credence in X will take some particular real value r.

When an agent repeatedly updates by Conditionalization, she often finds herself calculating the value of $c(X \mid E)$. This calculation can be streamlined by a famous theorem.

BAYES' THEOREM. For any $X, E \in \mathcal{L}$ with non-zero c-values,

$$c(X \mid E) = \frac{c(E \mid X) \cdot c(X)}{c(E)}.$$

Bayes' Theorem has proved so central to the application of Conditionalization that theorists who work with degrees of belief are often called "Bayesians" (or "subjective Bayesians," or "Bayesian epistemologists"). In a moment I'll describe why Bayes' Theorem is so useful. But first, it's worth noting that Bayes' Theorem is indeed a *theorem*, easily derivable from the Kolmogorov Axioms and Ratio Formula.[16] Bayesianism has generated a great deal of controversy, especially among statisticians. But the controversial claim in Bayesianism isn't that Bayes' Theorem is true. Everyone agrees that the theorem follows from the Kolmogorov Axioms, and that *if* an agent is going to generate new credences over time by conditionalizing, *then* the theorem provides a handy tool for calculating post-update credences from pre-update credences. The controversy is whether agents should really update their credences by conditionalizing, and whether scientific inference is best understood as a series of conditionalizations.

Setting this controversy aside, why is the particular analysis of $c(X \mid E)$ in Bayes' Theorem so useful? Consider a scientific context, in which a theorist has a finite partition of hypotheses $H_1, \ldots, H_n$ about what's going on with some phenomenon. The theorist plans to run an experiment that she hopes will discriminate among the hypotheses. At time t_i, before she has run the experiment, the theorist has a set of unconditional credences c_i, which we call her *priors*. The theorist runs the experiment between t_i and t_j, and let's suppose the observation she makes is represented by proposition E. Given this new evidence, Conditionalization helps her calculate her credences at t_j, which we call her *posteriors*.

Suppose we're interested in the theorist's confidence in some particular hypothesis H_m after the experimental results come in. Applying Conditionalization, Bayes' Theorem, and then the Law of Total Probability to the denominator of Bayes' Theorem, we derive:

$$c_j(H_m) = \frac{c_i(E \mid H_m) \cdot c_i(H_m)}{c_i(E \mid H_1) \cdot c_i(H_1) + \ldots + c_i(E \mid H_n) \cdot c_i(H_n)}. \tag{12}$$

16 The theorem is traditionally attributed to the Reverend Thomas Bayes. Though Bayes never published the theorem, Richard Price found it in his notes and published it after Bayes' death in 1761. Pierre-Simon Laplace rediscovered the theorem independently later on, and was responsible for much of its early popularization.

Consider the components of the right-hand fraction one at a time. First, we have a number of expressions of the form $c_i(H_x)$. These are the theorist's priors in the various hypotheses. Presumably going into the experiment she has some unconditional levels of confidence in the hypotheses she is considering; these supply the priors in question. Then we have expressions of the form $c_i(E \mid H_x)$. An agent's conditional credence in an experimental result E given some hypothesis H_x is called her *likelihood* for that evidence on that hypothesis. A well-defined scientific hypothesis should make a prediction for how the theorist's experiment will come out, or at least should assign probabilities to various possible outcomes. These inform the theorist's likelihoods for various experimental outcomes (such as E) on the various hypotheses she entertains. Thus Bayes' Theorem allows the theorist to form a posterior opinion about each hypothesis H_m that she entertains, based on the evidence she's received, her unconditional priors in the hypotheses, and her t_i likelihoods—elements that are plausibly all easily to hand.

1.4 *Jeffrey Conditionalization*

Statisticians and philosophers of science often worry that Conditionalization allows a scientist's final verdict on a hypothesis to be influenced by her initial credence in that hypothesis—her personal degree of belief in the hypothesis before any evidence came in. Epistemologists worry about Conditionalization's conception of evidence. It seems that for Conditionalization to work, it must be possible to identify some proposition E representing *everything* the agent learns between t_i and t_j. Moreover, the agent must become *certain* of E between t_i and t_j, because updating the agent's credence in E itself using Conditionalization yields $c_j(E) = 1$. Finally, once an agent becomes certain of some proposition, subsequent updates by Conditionalization will retain that certainty forever.[17]

Conditionalization therefore seems to embody a conception of learning on which what is learned is explicitly summarizable in propositional form, becomes certain, and is retained ever after. To epistemologists, this is reminiscent of foundationalist approaches to evidence abandoned decades ago. It also violates the Regularity Principle, which deems it irrational for an agent to assign absolute certainty to an empirical proposition. (After all, what evidence could ever make you *entirely certain* that some empirical claim was true?)

To address these problems, Richard C. Jeffrey offers an updating rule that generalizes Conditionalization to allow for learning experiences in

17 It's easy to show that if an agent conditionalizes on E between t_i and t_j, she will have $c_j(E) = 1$, and then if she conditionalizes on some other evidence between t_j and t_k, she will still have $c_k(E) = 1$ as well.

which no certainties are gained. He introduces his rule using the following example.

> The agent inspects a piece of cloth by candlelight, and gets the impression that it is green, although he concedes that it might be blue or even (but very improbably) violet. If G, B, and V are the propositions that the cloth is green, blue, and violet, respectively, then the outcome of the observation might be that, whereas originally his degrees of belief in G, B, and V were .30, .30, and .40, his degrees of belief in those same propositions after the observation are .70, .25, and .05. (Jeffrey, 1965, p. 154)

Discussing the example, Jeffrey writes:

> If there were a proposition E in [the agent's] preference ranking which described the precise quality of his visual experience in looking at the cloth, one would say that what the agent had learned from the observation was that E is true.... But there need be no such proposition E in his preference ranking; nor need any such proposition be expressible in the English language.... The description 'The cloth looked green or possibly blue or conceivably violet,' would be too vague to convey the precise quality of the experience.... It seems that the best we can do is to describe, not the quality of the visual experience itself, but rather its effects on the observer, by saying, 'After the observation, the agent's degrees of belief in G, B, and V were .70, .25, and .05.' (Jeffrey, 1965, pp. 154–5)

Jeffrey proposes an updating rule he called "probability kinematics"; nowadays everyone calls it "Jeffrey Conditionalization." The rule applies when an agent's experience impinges on her credences by altering her degree of belief distribution across a particular finite partition in $\mathcal{L}$; any other changes in her credences are caused by the changes to this partition. If the originating partition is $B_1, \ldots, B_n$, then Jeffrey's rule is as follows.

Jeffrey Conditionalization. For any $A \in \mathcal{L}$,

$$c_j(A) = c_i(A \mid B_1) \cdot c_j(B_1) + \ldots + c_i(A \mid B_n) \cdot c_j(B_n).$$

Jeffrey did not mean to rule out the possibility that some learning occurs by certainty acquisition. He just wanted to allow for the possibility of other types of learning experiences as well. So in the case where one of the B_m goes to certainty (and therefore every other member of the partition goes to credence-0), Jeffrey Conditionalization reduces to traditional Conditionalization.

Let's see how Jeffrey Conditionalization applies to Jeffrey's cloth by candlelight example. Suppose the agent is interested in the proposition M, that the selected piece of cloth will match her couch. She's certain that anything violet will match, she's certain anything green will not, and she's 50% confident that a blue cloth will match. (The match depends on the specific shade of blue.) Let t_i be the time before she inspects the cloth by candlelight. Using the Law of Total Probability and the initial unconditional credences Jeffrey provides, we have

$$\begin{aligned} c_i(M) &= c_i(M \mid G) \cdot c_i(G) + c_i(M \mid B) \cdot c_i(B) + c_i(M \mid V) \cdot c_i(V) \\ &= 0 \cdot .30 + 0.5 \cdot .30 + 1 \cdot .40 = 0.55. \end{aligned} \tag{13}$$

Jeffrey also provides the agent's unconditional credences in G, B, and V at t_j, after the inspection. With these values, Jeffrey Conditionalization yields

$$\begin{aligned} c_j(M) &= c_i(M \mid G) \cdot c_j(G) + c_i(M \mid B) \cdot c_j(B) + c_i(M \mid V) \cdot c_j(V) \\ &= 0 \cdot .70 + 0.5 \cdot .25 + 1 \cdot .05 = 0.175. \end{aligned} \tag{14}$$

The glimpse by candlelight increases the agent's confidence that the cloth is green and decreases her confidence that the cloth is violet, so the Jeffrey-prescribed posterior that the cloth will match decreases.

Notice how this change in credence is effected. The agent's visual experience changes her credences by directly altering her distribution across the cloth-color partition. Any changes to other propositions in the agent's language (such as M) are downstream effects of this direct alteration. Yet the dependencies between these downstream propositions and the color propositions remain unaltered: changing the agent's opinions about the color of the cloth doesn't change how confident she is that particular colors will match the couch. This is why the same conditional credences appear in both the $c_i(M)$ and the $c_j(M)$ calculations.

Against the background of the Kolmogorov axioms and Ratio Formula, Jeffrey Conditionalization is equivalent to the following condition.

RIGIDITY. For any $A \in \mathcal{L}$ and any B_m, $c_j(A \mid B_m) = c_i(A \mid B_m)$.

In a Jeffrey Conditionalization, experience alters an agent's credences across the B-partition. The agent's credences in other propositions conditional on the B_ms don't change. So the agent sets her posteriors by adopting unconditional credences in the B_ms from experience, copying over her old conditional credences, then applying the Law of Total Probability to calculate her unconditional credences in non-B propositions.

1.5 *Further Rational Requirements*

We have now seen a variety of putative rational constraints on credence: the probability axioms, the Ratio Formula, the Reflection Principle, Regularity,

and the diachronic rules of Conditionalization and Jeffrey Conditionalization. Yet there are infinitely many credence distributions (and sequences of credence distributions over time) compatible with these constraints. Are all of those distributions rationally permissible? Some of them are quite strange, and unintuitive—for instance, some assign very high credence to skeptical scenarios; some will lead agents to reason counter-inductively.

One extreme position about the strength of rational constraints is sometimes called "Objective Bayesianism." This position endorses the Uniqueness Thesis (Feldman, 2007; White, 2005) that given any body of evidence, there is exactly one credence distribution rationally permitted to any agent with that body of total evidence. At the other extreme, what we might call "Extreme Subjective Bayesians" hold that any probabilistic credence distribution is rationally permissible. In between are "Moderate Subjective Bayesians," who hold that there are some rational constraints beyond the ones we've described, but not enough to generate a unique permissible distribution in every case.

What might these further rational constraints be? A constraint that might considerably narrow the field of what's rationally permissible is the

> PRINCIPLE OF INDIFFERENCE. If an agent has no evidence favoring any possibility in a partition over any other, then she should assign equal credence to each element of the partition.[18]

The traditional objection to this principle is that it seems to give conflicting advice when we repartition the same space of possibilities. Following van Fraassen (1989a), suppose I tell you that a cube has been produced from a factory, and its side length is between 0 and 1 meter. Given the paucity of further evidence, if I ask how confident you are that the side length is less than 0.5 meters, the Principle of Indifference seems to require a credence of 1/2. But if I now ask how confident you are that the volume (which must be between 0 and 1 cubic meter) is less than 0.5 cubic meters, the Principle of Indifference also seems to require a credence of 1/2. Since a side length of 0.5 meters corresponds to a volume of 0.125 cubic meters, the only way to assign both these credences consistently with the probability axioms is to be absolutely certain that the volume in cubic meters is not between 0.125 and 0.5![19]

Another family of putative rational constraints has a member we've already seen. The Reflection Principle directs us to set our current uncon-

18 The basic idea here dates back at least to Laplace (1814/1995), who saw it as an application of what Bernoulli (1713) called the "principle of insufficient reason."

19 A more technically-sophisticated cousin of the Principle of Indifference is Jaynes' (1957a, 1957b) Maximum Entropy Principle. This principle applies more naturally over infinite partitions, and adapts well to a variety of forms of evidence. Yet it still succumbs to partition variance problems, and also conflicts with updating by conditionalization in particular cases. See Seidenfeld (1986).

ditional credence in a proposition equal to what we're certain it will be in the future—or if we're not certain of our future credences, equal to our expectation of what they will be. This principle directs us to defer to the opinions of our future self as if she were some sort of expert. But of course there are other experts in the world, such as contemporaries who we think have better judgment or information than ourselves. Following the lead of the Reflection Principle, Elga (2007) suggests that if c_e is the credence distribution of an agent we consider an expert, then for any $X \in \mathcal{L}$ (or at least any X in the expert's area of expertise) we should assign

$$c(X \mid c_e(X) = r) = r. \tag{15}$$

Thinking more metaphorically, an "expert" distribution worthy of our deference need not even be an agent. It may be rational to align our credences with certain objective numerical values in the universe. This brings us to the topic of direct inference principles.

1.6 *Direct Inference Principles*

Page 1 briefly mentioned interpretations of probability—proposals for the meaning of "probability" locutions. For example, the classical interpretation, dating back at least to Laplace (1814/1995), defined probability as the number of favorable outcomes of a process divided by the total number of outcomes possible. Later, the frequency theory of probability (associated most closely with von Mises, 1928/1957), read probability as the frequency with which an outcome would occur were a particular process repeated many times.[20]

My task here is not to assess these notions of probability as proposals in the theory of meaning, or in the theory of probability. Instead, I want to ask what these notions have to do with rational credence. Many Bayesians have endorsed principles of direct inference: principles carrying the agent from information about some notion of probability to specific credences in specific events. For example, it might be that if I'm certain a particular type of experimental setup produces a particular type of outcome with frequency x, then when an experiment of that type is to be run, I should have credence x that it will yield an outcome of that type. This would be a principle of direct inference from frequency facts to credences in outcomes.

Frequency-to-credence principles face notorious difficulties, even when sketched out as roughly as I've just done. For one, a single event (I go

20 The previous section introduced one usage of "Objective/Subjective Bayesian" terminology. That usage should be carefully distinguished from another usage that often comes up in the literature about interpretations of probability. In that literature, "Subjective Bayesianism" describes the position that in everyday talk, "probability" always refers to or expresses subjective credences. "Objective Bayesianism," on the other hand, holds that probability talk refers to something beyond the subject, such as frequencies or chances.

in to my office tomorrow) can be classed as the outcome of a variety of experiment types (choosing whether to go in on a summer day, choosing whether to go in on a Tuesday, etc.), which may yield different frequencies and therefore different credal recommendations. (This is one version of the "reference class problem."[21]) Also, if we tried to use this principle as a general credence-setting strategy, we'd have trouble with experiments that look to be unrepeatable. Before the Large Hadron Collider was switched on, newspapers prominently reported physicists' degrees of belief that doing so would destroy the Earth. It's difficult to align such credences with the frequency with which switching on the collider would cause global destruction; in the event of such destruction, the switching-on only occurs once.

It may therefore be preferable to link rational credence with "objective chance." As a notion of probability, chance is objective, in the sense that its value is determined by the physical makeup of an experimental apparatus. Chance may also be applied to events that occur only once. A frequency-to-credence principle recommends credence $1/6$ that a fair die roll will come up 3 on the grounds that repeating the roll will yield 3 one-sixth of the time. The objective chance theorist recommends $1/6$ on the grounds that a fair die is physically constituted in a particular manner (equally weighted on each side, etc.). This would remain true even if the die had never been rolled before, and was guaranteed to be destroyed after the roll in question.

The most famous direct inference principle linking credence and chance is Lewis' (1980) Principal Principle. *Very* roughly, and skipping over a great many details,[22] the Principal Principle directs an agent to set

$$c(A \mid Ch(A) = x) = x, \tag{16}$$

unless she possesses inadmissible evidence relevant to A. Here $Ch(A) = x$ is the proposition that the objective chance of A is x. So—setting aside the matter of inadmissible evidence for a moment—if the agent is certain that, say, a particular die has a $1/6$ chance of coming up 3, the Principal Principle will set her credence in 3 at $1/6$. If, on the other hand, the agent knows the die is biased, but splits her credence evenly between the number 3s having a $1/10$ chance and a $1/5$ chance of coming up, the Law of Total Probability will combine with the Principal Principle to yield:

$$\begin{aligned} c(3) &= c(Ch(3) = 1/10) \cdot c(3 \mid Ch(3) = 1/10) \\ &\quad + c(Ch(3) = 1/5) \cdot c(3 \mid Ch(3) = 1/5) \\ &= 1/2 \cdot 1/10 + 1/2 \cdot 1/5 \\ &= 0.15. \end{aligned} \tag{17}$$

21 See Hájek (2007) for many more versions.

22 See Meacham (2010) for some of those details.

In other words, her credence that the die will come up 3 is her expectation of the objective chance of getting a 3. We can therefore think of the Principal Principle as an expert deference principle in which the expert is objective chance.

The key innovation of Lewis' Principal Principle is its treatment of evidence the agent takes to be relevant to the outcome of a chance event. Lewis divides such evidence into two sorts: admissible evidence is evidence that the agent takes to be relevant to the outcome because it affects her opinion of the objective chance of the event. For example, information about the weighting of the die is admissible with respect to the outcome of the roll—it affects how the agent thinks the roll will come out *by way of* affecting what the agents thinks are the chances of a 3. Inadmissible evidence affects the agent's opinion in some other way. For instance, if a confederate tells her how the roll came out, this affects the agent's opinion of whether it came out 3, but not by making her think the chances of a 3 were any different going in. Lewis' insight was that chance facts about an outcome screen off admissible information relevant to that outcome. So if E is admissible, the Principal Principle also gives us:

$$c(A \mid Ch(A) = x \,\&\, E) = c(A \mid Ch(A) = x) = x. \tag{18}$$

Admissible evidence relates to chances much the way a distal cause relates to the proximal cause of an event.

1.7 *Countable Additivity*

Up to this point the examples we've considered have typically involved only finitely many possibilities. But what if an agent considers a partition of infinitely many possible outcomes, and distributes her credence equally among all of them? How can this be modeled in our Bayesian epistemology?

To have a concrete example, let's suppose that a positive integer has been selected by some process, and our agent wants to assign equal credence to each integer's having been selected. Presumably that should be possible. But what numerical value might that credence take? It's easy to show that the probability axioms prevent its being a positive real. For suppose the agent assigns

$$r = c(1) = c(2) = c(3) = \ldots. \tag{19}$$

(Where $c(1)$ is her credence that 1 was selected.[23]) For any positive real r, there will exist a positive integer n such that $r > 1/n$. Now consider the

23 Notice we are now dealing with a language containing infinitely many atomic propositions. While this is a change from our earlier setup, it's not too difficult to manage, and is fairly common in formal models.

agent's credence that the selected integer is between 1 and n (inclusive). If you look back at the list of intuitive constraints following from the Kolmogorov axioms (Section 1.1), the last principle on the bulleted list will give us

$$c(1 \vee 2 \vee \ldots \vee n) = c(1) + c(2) + \ldots + c(n) = r \cdot n > 1, \tag{20}$$

which violates the axioms.

What other options are available? One popular suggestion is that when an agent assigns equal confidence to infinitely many possibilities, we represent that level of confidence as a credence of 0. So we would say that $c(1) = c(2) = \ldots = 0$.

Using credence 0 in this way introduces a few problems. First, up until this point we've conceived credence 1 as representing certainty in a proposition, and credence 0 as certainty that the proposition is false. Now we'll have to allow an agent to assign $c(P) = 0$ even if the agent admits P might be true, and $c(\sim P) = 1$ even if the agent isn't certain P is false. And we'll have to phrase the Regularity principle carefully: we may still prohibit agents from assigning certainty to empirical propositions, but no longer ban credences of 1 and 0 in such propositions.

Second, the Ratio Formula we've provided only relates the conditional credence $c(X \mid Y)$ to unconditional credences when $c(Y) > 0$. We'll need to expand this principle to handle cases in which $c(Y) = 0$ yet the agent doesn't rule Y out. For instance, our agent assigning equal credence to the selection of each positive integer might assign $c(2 \mid 2 \vee 4) = 1/2$, even though $c(2 \vee 4) = c(2) + c(4) = 0$.[24]

Third and most importantly, we'll want a way to sum credences over infinite disjunctions. Finite Additivity only covers disjunctions with finitely many disjuncts—what if we want to calculate our agent's credence that the selected integer is even? A natural extension of Finite Additivity is the following.

Countable Additivity. For any countable partition $\{Q_1, Q_2, Q_3, \ldots\} \subseteq \mathcal{L}$,

$$c(Q_1 \vee Q_2 \vee Q_3 \vee \ldots) = c(Q_1) + c(Q_2) + c(Q_3) + \ldots.$$

Countable Additivity is not only natural; it also allows us to establish a very important constraint on credences.

Conglomerability. For any proposition $P \in \mathcal{L}$ and partition $\{Q_1, Q_2, Q_3, \ldots\} \subseteq \mathcal{L}$, $c(P)$ is no greater than the largest $c(P \mid Q_i)$ and no less than the least $c(P \mid Q_i)$.

24 One way to manage this situation is to take conditional credences as basic. See footnote 10 for more information.

Given Conglomerability, the $c(P \mid Q_i)$ establish upper and lower bounds on the value of $c(P)$. This makes sense if you think of $c(P)$ as a weighted average of the credences the agent would assign to P conditional on all the different possible Q_i. And it's especially important when the agent has a partition $\{E_1, E_2, E_3, \ldots\}$ of possible new pieces of evidence she might receive before her next update. Assuming she plans to update by Conditionalization, she knows that her future credence in P will be one of her current $c(P \mid E_i)$; Reflection then demands she satisfy Conglomerability.[25]

The Conglomerability/Countable Additivity package is attractive. But it's inconsistent with assigning a credence of 0 to each positive integer in our example. The reason is simple: given Countable Additivity, the agent's credence that any positive integer will be selected at all is the sum of her credences in each individual integer. But the former value should be 1, while the latter individual values are each 0. So advocates of Countable Additivity have suggested instead that in this situation the agent assign an infinitesimal value to each integer's being selected. The infinitesimals are an extension of the set of real numbers, defined to be greater than 0 but less than any given real number. Thus they don't fall prey to the problem of our Equation 20. At the same time, adding up infinitely many infinitesimals can yield a real number, so we can maintain both Countable Additivity and a credence of 1 that any integer will be selected at all.

Yet infinitesimals introduce difficulties of their own; for some of the difficulties, and many of the mathematical details, see Williamson (2007), Easwaran (2014), Hájek (2003, Section 5), and Wenmackers (this volume).

2 APPLICATIONS OF CREDENCE

I've presented the Bayesian study of credence as the study of a doxastic attitude type, and what it takes to make such attitudes rational. This study is valuable in its own right, as a contribution to epistemology and the philosophy of mind. But historically it's also been pursued to enhance our understanding of other topics, some of which we'll discuss in this section.

2.1 *Confirmation Theory*

A Bayesian epistemologist or philosopher of science studies justification and evidential support by thinking about "confirmation." The type of con-

25 Notice that my statement of Conglomerability doesn't specify the cardinality of the Q_i partition. For finite partitions, Conglomerability can be proven from the standard probability axioms. Adopting Countable Additivity extends Conglomerability to countable partitions. For an agent who entertains larger disjunctions than that, Seidenfeld, Schervish, and Kadane (manuscript) show that at each cardinality we need the relevant Additivity principle to secure conglomerability for partitions of that size.

firmation studied is usually incremental, rather than all-things-considered; when we say that "evidence E confirms hypothesis H," we mean that E provides at least some positive evidential support for H, not that it settles the matter of H or even pushes H past some crucial threshold.[26] For a Bayesian, confirmation is also always relative to a probability distribution, and to a background corpus of propositions. Most commonly, the probability distribution will be some agent's credence function, and the background corpus will be the total evidence informing that credence function. (On a Conditionalization regime, the corpus is represented formally by the set of all propositions X such that $c(X) = 1$.[27]) So we take a given agent at a given time, and ask whether E confirms H for her, relative to her credences and background corpus at that time.

Letting K represent a background corpus, and c_k represent a probability distribution informed by that corpus, Bayesian confirmation theory posits that

$$E \text{ confirms } H \text{ relative to } c_k \text{ just in case } c_k(H \mid E) > c_k(H).$$

Bayesian confirmation is just positive probabilistic relevance relative to c_k. (Similarly, disconfirmation is usually defined as negative relevance relative to c_k.)

Though fairly simple, this theory of confirmation turns out to be surprisingly subtle, powerful, and convincing. To illustrate—and fix the intended notion of evidential support in the reader's mind—suppose a fair die has just been tossed, and you know nothing of the outcome. Perhaps in accordance with the Principal Principle, some frequency principle, or even the Principle of Indifference, you assign equal credence to each of the six possible outcomes. Relative to your credence distribution and background corpus, if you received evidence that the toss came up with a prime number, this would confirm for you that the toss came up odd. Why? Because if you satisfy the Kolmogorov axioms and Ratio Formula, then you assign

$$2/3 = c(\text{odd} \mid \text{prime}) > c(\text{odd}) = 1/2. \tag{21}$$

This doesn't mean that prime evidence should make you *certain* the toss came up odd, or even that it would justify you in *believing* the toss came up odd. But if you update by Conditionalization, learning that the toss came up prime would make you at least somewhat more confident that the toss came up odd. Again, the confirmation here is incremental.

26 This contrasts with the way "confirms" is sometimes used in English, as when we speak of a nominee's being confirmed, or even a dinner reservation.

27 Notice that despite our suggestion in Section 1.7 that it might sometimes be interpreted otherwise, I have gone back to treating credence 1 as representing certainty. To simplify discussion, I will continue to do this going forward.

This Bayesian theory of confirmation gives the confirmation relation some interesting and intuitive formal properties.[28]

- If $E \dashv\models E'$ and $H \dashv\models H'$, then E confirms H just in case E' confirms H'.
- E confirms H just in case E disconfirms $\sim H$.
- If $E \,\&\, K \models H$ but $K \not\models H$, then E confirms H.
- If $H \,\&\, K \models E$ but $K \not\models H$, then E confirms H.

The first of these properties ensures that logical equivalents behave the same within the confirmation relation. The second relates confirmation to disconfirmation. The third and fourth properties[29] specify how confirmation relates to entailment. The third property tells us that entailment is a form of confirmation; if E entails H jointly with K while K didn't entail H on its own, then E confirms H. As for the fourth property, it captures the idea[30] that a hypothesis which predicts an evidential observation (in concert with one's background corpus) is confirmed by that observation.

On the other hand, the Bayesian theory withholds from the confirmation relation certain properties that are sometimes mistakenly ascribed to it. Here are two examples.

- If E confirms both H and H', then the set H, H', K is logically consistent.
- If X confirms Y and Y confirms Z, then X confirms Z.

The first of these properties is important to reject because we're talking about incremental confirmation. For example, in Jeffrey's example in which an agent inspects a piece of cloth by candlelight, his brief glimpse may confirm that the cloth is green, while also confirming that it's blue or even that it's violet. (Perhaps the glimpse disconfirms that the cloth is red and disconfirms that it's orange.) This is perfectly reasonable, despite the fact that green, blue, and velvet are inconsistent hypotheses about the color of the cloth. Similarly, in scientific settings the same observation may confirm mutually exclusive theories from a partition, while at the same time (perhaps) ruling others out.

The latter property is the supposed property of confirmation transitivity. This is one of the most common mistakes made about confirmation, support, justification, and other related notions.[31] Just because X confirms Y

28 In every one of these properties, the expressions "E confirms H" and "E disconfirms H" should be followed by the phrase "relative to c_k." Going forward I'll simplify locutions by leaving the relativization to c_k implicit whenever possible.

29 Both of which require a side-condition that the set E, K, H is logically consistent.

30 Familiar from hypothetico-deductivism (Crupi, 2016, Section 2).

31 Correcting this mistake has been a theme of the epistemology literature about epistemic and justificatory closure. See, e.g., Dretske (1970), Davies (1998) and Wright (2003).

and Y confirms Z does not mean that X confirms Z—even in the special case when Y *entails* Z! To see why, imagine a card has been drawn at random from a standard playing card deck. Information that the card is a spade confirms (incrementally!) that the card is the Jack of Spades. But information that the card is a spade does not even incrementally confirm that the card is a jack.

Another common mistake is to conflate what Carnap (1962) called "firmness" and "increase in firmness" accounts of confirmation.[32] The Bayesian account we've been discussing is an increase in firmness account. A firmness account, on the other hand, says that E confirms H relative to c_k just in case $c_k(H \mid E)$ is high (where the necessary level of height may be influenced by, say, contextual parameters). Among many other problems, the firmness account errs by maintaining that E confirms H in cases when $c_k(H \mid E)$ is high simply because the prior $c_k(H)$ is high. In fact, a firmness account may say that E confirms H relative to c_k even though $c_k(H \mid E)$ is lower than $c_k(H)$ (as long as $c_k(H \mid E)$ is nevertheless high)! The Bayesian account focuses on the relation between E and H—how E would *alter* the agent's opinion of H—rather than just on where that opinion would land were E taken into account.

We can provide more information about E's effect on the agent's opinion of H by measuring the degree of incremental confirmation. The simplest way to measure confirmation is to calculate $c_k(H \mid E) - c_k(H)$; this measure simply asks how much conditionalizing on E would increase the agent's confidence in H. Yet as a measure of E's bearing on H, this simple difference has some drawbacks. For example, the degree to which E can confirm H will be limited by the value of $c_k(H)$. If, say, $c_k(H) = 0.99$, then even if E entails H, the maximal degree to which it can confirm H will be 0.01. Bayesian confirmation theory thus has a considerable literature proposing and assessing alternative measures of confirmational strength; see Crupi (2016, Section 3.4) for a recent summary and references.

One upshot of the literature on measuring confirmation is a new approach to "solving" traditional paradoxes of confirmation. For example, we usually think that universal generalizations are confirmed by their positive instances. The hypothesis that all ravens are black is typically confirmed by the evidence that a particular raven is black.[33] In symbols, $(\forall x)(Rx \supset Bx)$ is confirmed by $Ra \,\&\, Ba$. But now suppose we discover an item that is a non-black non-raven. The evidence $\sim Ba \,\&\, \sim Ra$ is a positive

32 Carnap was well-acquainted with this mistake, having made it himself in the first 1950 edition of his *Logical Foundations of Probability*.

33 I say "typically" because it is possible to generate a deviant background corpus against which it would be reasonable for the observation of a black raven to *dis*confirm that all ravens are black. (For examples, see Swinburne, 1971, and Rosenkrantz, 1977, Chapter 2.) The generation of the paradox doesn't rely on such deviant corpora, so we will set them aside for the rest of the discussion.

instance of the generalization $(\forall x)(\sim Bx \supset \sim Rx)$, so it should confirm that generalization. Yet the latter generalization is (by contraposition) logically equivalent to our former one. So by the first property of confirmation I endorsed above, $\sim Ba \,\&\, \sim Ra$ should confirm that all ravens are black. This is Hempel's (1945) famous "Paradox of the Ravens," which seems to generate the absurd conclusion that a hypothesis about the color of ravens may be confirmed by the observation of a white shoe.

Recently, a number of Bayesian confirmation theorists have conceded that perhaps a white shoe does confirm that all ravens are black—it's just that observing a white shoe confirms this hypothesis much *less* than observing a black raven would.[34] Fitelson and Hawthorne (2010), for instance, specify conditions on c_k such that as long as these conditions are met, evidence of a black raven will confirm the ravens hypothesis much more strongly than evidence of a non-black non-raven, on virtually every proposed measure of confirmation in the literature. It's highly plausible that most of us in the real world have credence distributions satisfying Fitelson and Hawthorne's conditions, accounting for our intuitions about the asymmetry of favoring in this case. Similar approaches have been taken to the problem of irrelevant conjunction (Hawthorne & Fitelson, 2004) and Goodman's (1955) grue paradox (Chihara, 1981; Eells, 1982).

2.2 *Decision Theory*

Since this handbook contains an extensive article on decision theory (Thoma, this volume), I will give only a brief sketch here. In formal decision theory, an agent is confronted with a decision problem, represented by a partition of acts she may perform. Once she performs an act, some outcome will occur, and the agent values different outcomes to different degrees. These valuations are represented by a utility function, which assigns real-number utilities to each possible outcome. (The key assumption about utilities is that they measure value *uniformly*—the agent takes each added unit of utility to be as valuable as the next. The same is not true of money; your first dollar may be much more valuable to you than your billionth.)

So what's difficult about that—shouldn't the agent just choose the act leading to the most valuable outcome? The trouble is that the agent may be uncertain which acts will lead to which outcomes. Put another way, the agent may be unsure what state the world is in, and the outcome that follows her decision may depend both on the act she chooses and on the remaining state of the world. For example, suppose I'm trying to decide whether to go into my office tomorrow. I know that if I go, it *may* be quiet

34 Though the idea dates all the way back to Hosiasson-Lindenbaum (1940).

and peaceful there, in which case I'll get a great deal of writing done, which is an outcome I highly value. On the other hand, there may be loud construction happening outside my office window, in which case I'll dally on the internet and get no writing done, an outcome to which I assign little utility. Since I don't know the state of construction around my building tomorrow, it's unclear to me which available act (go into the office, stay home) correlates with which outcomes, complicating my decision.

The standard solution to this problem is to have the agent assign an expected value to each available act. An agent's expected value for an act is her expectation for the amount of utility that will accrue if she performs the act—calculated using her credences that various states of the world obtain. Given a decision between two acts, a rational agent prefers the act to which she assigns the higher expected value (and is indifferent in case of ties). We can thus use her credence and utility assignments to develop a preference ordering over the acts available to her in any decision problem.

For example, suppose I assign a utility of 100 to a day of peaceful writing at my office, but a utility of 0 to spending the day there with construction going on. If I'm 40% confident there'll be no construction tomorrow, my expected utility of going into the office is

$$\begin{aligned} EU(\text{go to office}) &= c(\text{no construction}) \cdot u(\text{peaceful writing}) \\ &\quad + c(\text{construction}) \cdot u(\text{wasted day}) \\ &= 0.40 \cdot 100 + 0.60 \cdot 0 \\ &= 40, \end{aligned} \tag{22}$$

where the function u designates the amount of utility I assign to a given outcome. Given this expected utility for going to the office, I should prefer to stay home only if I expect doing so to yield me a utility greater than 40.

We can prove that if an agent sets her preferences by maximizing expected utility, her preference ordering over acts will satisfy various intuitive conditions, commonly known as the "preference axioms." For example, her preferences will be asymmetric (she never prefers both A to B and B to A) and transitive (if she prefers A to B and B to C, then she prefers A to C).

As I said, I'm going to avoid the many subtleties of developing a full-blown decision theory. One crucial concern is cases in which the agent's act may be correlated with the state of the world. Evidential decision theorists (Jeffrey, 1965) respond by working with the agent's credence in a state *conditional* on her performing a particular act, while causal decision theorists (Gibbard & Harper, 1978; Lewis, 1981; Joyce, 1999; Weirich, 2012) consider the agent's credence that her act will *cause* a particular state to obtain. Another concern is modeling risk-averse agents—such as an agent who prefers a guaranteed payout with utility 1 to a fair coin flip on which heads yields a prize with utility 3 (Allais, 1953; Buchak, 2013).

There is, however, one more notion from decision theory that we'll need in what follows: fair betting prices. Consider a proposition P and a betting slip that guarantees its possessor \$1 if P turns out to be true. How much is that betting slip worth to you? That depends how confident you are that P obtains. If you're certain of P, that slip is worth \$1 to you. If you're certain P is false, the slip is worth nothing. In between, the more confident you are of P, the more value you assign to the betting slip.

To be more precise, your expected value in dollars of the fair betting slip is $c(P) \cdot \$1$. We call this your *fair betting price* for this gamble on P. In general, if a bet pays out $\$X$ dollars when P is true, your fair betting price for the bet is

$$c(P) \cdot \$X. \tag{23}$$

What does it mean to say this is your fair betting price? Suppose someone offers to sell you a betting slip that pays off on P. Your fair betting price is the price at which you'd expect to break even on such an investment. Assuming you value money linearly (so that each additional cent confers the same amount of additional utility on you), decision theory says that you should be willing to purchase the betting slip for any amount lower than your fair betting price, and indifferent about buying it at exactly your fair betting price. Conversely, if you possess such a slip, you should be willing to *sell* it for any amount *above* your fair betting price.

2.3 *Other Applications*

Historically, confirmation and decision theory have been major drivers of Bayesianism's development and the two most common applications to which the approach has been put. But the Bayesian theory of credences has been applied to many other philosophically significant topics as well. Here are a few examples.

- Probabilities have been used to measure when the propositions in a set cohere. Coherentism about justification has then been evaluated by asking whether coherence among propositions makes it rational to invest a higher credence in each of them. See Shogenji (1999), Bovens and Hartmann (2003), Huemer (2011), and Olsson (2017).

- It's been debated whether an agent who updates by conditionalization will thereby increase her credence in the hypothesis that best explains evidence observed. Van Fraassen (1989b) argues that Bayesianism is incompatible with Inference to the Best Explanation. Replies have been offered by, *inter alia*, Okasha (2000), Lipton (2004), Weisberg (2009), and Henderson (2013).

- Elga (2007) argues that when an agent discovers that an epistemic peer has assigned different credences than her based on the same evidence, that agent should move her credences closer to her peer's. A great deal of debate has ensued about whether such conciliationism is the rational response to peer disagreement. Christensen (2009) presents a useful survey that is unfortunately now outdated; Christensen and Lackey (2013) is a more recent collection. (Though plenty has been published on the subject since then!)

- The peer disagreement controversy intersects with broader questions about the rational response to higher-order evidence—evidence concerning whether one has responded rationally to one's evidence. New essays on higher-order evidence and its connection to disagreement may be found in Rasmussen and Steglich-Petersen (forthcoming).

- Peer disagreement is also an aspect of social epistemology, which has considered for decades how groups and individuals should combine the opinions of multiple experts to form a coherent single view. The literature on probabilistic opinion pooling dates back at least to Boole (1952). More recent discussions, with copious additional references, include Bradley (2007), Russell, Hawthorne, and Buchak (2015), and Easwaran, Fenton-Glynn, Hitchcock, and Velasco (2016).

3 ARGUMENTS FOR CREDAL CONSTRAINTS

Many of the constraints on credences presented in Section 1 have an intuitive claim on being rationally required. It's just plausible that the more confident you are it will rain tomorrow, the less confident you should be that it won't rain. But can we provide *arguments* for the various rational constraints? Here I'll survey three historically-significant approaches to arguing for rational constraints on credence.

3.1 *Representation Theorem Arguments*

In Section 2.2 I suggested that if an agent has credence and utility functions, decision theory can combine these to determine her rational preferences among acts. But decision theory can also work in the opposite direction. Suppose I observe an agent make a number of decisions over her lifetime. Assuming these choices express her preferences among acts, I can construct credence and utility functions for her that would rationalize such preferences if she is an expected utility maximizer. I might then use

these credence and utility functions to predict choices she'll make in the future.[35]

We can prove that as long as an agent's preferences are rational, she can be represented in this way as maximizing expected utility by combining credence and utility functions. More precisely, a representation theorem shows that given a preference ordering over acts satisfying certain preference axioms, there exists a utility function and a probabilistic credence function on which those preferences maximize expected utility. Since there are many different versions of decision theory, there are many sets of preference axioms, and so many different representation theorems.[36] But typically the preference axioms can be divided up into two sorts: substantive constraints such as the asymmetry and transitivity requirements I mentioned earlier; and what Suppes (1974) calls "structure axioms" specifying that the preference ordering is complete, has acts available at a variety of levels of preference, etc. (Structure axioms are usually considered a convenience to make the theorems cleaner and the proofs easier.)

Representation theorems can be highly useful. For instance, economists engaged in rational choice theory often model market participants as maximizing expected utility based on a utility function and a probabilistic credence function. A representation theorem assures us that as long as an agent remains rational—in the sense of making choices that satisfy the preference axioms—her behavior will continue to conform to such a model.

Yet there's a big step from arguing that rational agents can be *modeled* as employing a probabilistic credence function to arguing that rational agents actually *possess* probabilistic credence functions (Hájek, 2009; Meacham & Weisberg, 2011). We can begin to see the problem by noting that an agent's preferences will often underdetermine her utility and credence distributions. That is, if all we know is an agent's preferences, there are (infinitely) many different pairs of utility and credence functions that will generate that preference ordering by maximizing expected utility. Moreover, many of those pairs feature credence functions that don't satisfy the probability axioms. Standard representation theorems prove only that if an agent's preferences satisfy the axioms, *there exists* a corresponding credence/utility

35 We can think of this as a formalization of the folk deployment of a "theory of mind." I watch what you do, I surmise what you want and what you believe, then I let that information guide my interactions with you going forward.

36 Representation theorems were inspired by early, suggestive results in Ramsey (1931). The first rigorous representation theorem of the type we're discussing is in Savage (1954). (Though see also von Neumann and Morgenstern, 1947.) A representation theorem for evidential decision theory appears in Jeffrey (1965), while Joyce (1999) proves one for causal decision theory.

pair in which the credence function satisfies the probability rules. This hardly shows that rationality requires probabilistic credences.

Matters can be improved with a representation theorem proven by Lara Buchak and myself. (A sketch of the proof appears in the Appendix.) This theorem shows that if an agent's preferences satisfy various preference axioms, and she maximizes expected utility, then her credence function must be a *positive scalar transformation* of a probability distribution. In other words, her credences will be non-negative, they will be finitely additive, they will assign the same value to every tautology, and that value will be greater than the value assigned to contradictions. A credence function like this will have all the same properties as a probabilistic function, except that the maximal value it assigns to tautologies may be some positive number other than 1. Yet nothing substantive hangs on whether we measure credence on a 0 to 1 scale or instead, say, a percentage scale from 0 to 100.

Still, even the improved theorem assumes that the agent's credences and utilities interact with preferences through the maximization of expected utility. Zynda (2000) notes that there are many other mathematical quantities combining credence and utility that an agent could choose to maximize. So to argue for probabilism (or something close to it) using one of these representation theorems, we need to assume not only that rationality requires satisfying the preference axioms, but also that it requires maximizing expected utility.

3.2 *Dutch Book Arguments*

Like representation theorems, an inspiration for Dutch Book arguments can be found in Ramsey's (1931), in which he commented,

> These are the laws of probability, which we have proved to be necessarily true of any consistent set of degrees of belief.... If anyone's mental condition violated these laws, his choice would depend on the precise form in which the options were offered him, which would be absurd. He could have a book made against him by a cunning better and would then stand to lose in any event. (p. 84)

Suppose, for instance, that I am both 0.7 confident that I will go to my office tomorrow and 0.7 confident that I will not. Now consider two betting slips—one that pays a dollar if I go to the office, and another that pays a dollar if I don't go to the office. Given my credences, my fair betting price for each of these slips is \$0.70. That means I'm willing to pay up to \$0.70 for each of them. So suppose I buy both, at a price of \$0.70 each. I've now spent a total of \$1.40, and no matter what happens tomorrow, I will only make \$1. My non-probabilistic credence distribution has made me

susceptible to a combination of bets on which I will lose $0.40, come what may!

De Finetti (1937/1964) proved that if an agent's credences violate the probability axioms, a set of bets exists such that if the agent purchases each of them at her fair betting price, she will lose money in every possible world. For unknown reasons, such a set of bets is called a "Dutch Book." The proof works by going through each of the axioms one at a time, and showing how to construct a Dutch Book against an agent who violates the relevant axiom. Moreover, we can establish what Hájek (2009) calls a "Converse Dutch Book Theorem," showing that if an agent *satisfies* the probability axioms, no Dutch Book of the types described in de Finetti's proof can be constructed against that agent.

Other proofs show how to construct Dutch Books against agents who violate the Reflection Principle (van Fraassen, 1984), the Principal Principle (Howson, 1992), Regularity (Kemeny, 1955; Shimony, 1955), and Countable Additivity (Adams, 1962). We can also construct what is known as a "Dutch Strategy" against any agent who violates Conditionalization (Teller, 1973, reporting a result of David Lewis') or Jeffrey Conditionalization (Armendt, 1980; Skyrms, 1987b). A Dutch Strategy is not strictly speaking a particular set of bets guaranteed to give the agent a sure loss; instead, it's a strategy for placing bets with the agent in which certain bets are placed at an initial time, then future bets are placed depending on what the agent learns after that time. Still, the idea of a Dutch Strategy is that no matter what happens (and no matter what the agent learns), if she purchases the bets at her fair betting prices when they're offered, she'll face a net loss come what may.

Avoiding Dutch Books and Dutch Strategies seems an important advantage for the probabilistic agent. Still, can we *argue* that rationality forbids susceptibility to Dutch Strategies and Books? One problem is that the negative effects of violating probabilism highlighted by Dutch Books seem oddly practical. We might have thought that the Kolmogorov axioms provided constraints of *theoretical* (rather than *practical*) rationality on agents' credences. Yet here we're arguing for those axioms by pointing to financial consequences of violating them. Moreover, it's unclear how seriously we should take those potential consequences. Are non-probabilistic agents ever really going to face the precise set of bets that would expose them to a Dutch Book? And what if the non-probabilistic agent has read about Dutch Books, and decides that instead of changing her credences, she'll just be more cautious in her betting behavior? In the example above concerning my going to the office, I might pay $0.70 for the bet that pays off if I go into the office, but then refuse to buy the second bet because I see a Dutch Book coming. In that case I'll still have non-probabilistic credences, but manage by practical strategizing to avoid the prospect of a sure loss.

Taking a cue from the second sentence of the Ramsey quote above, a number of authors have tried to "depragmatize" Dutch Book arguments. Skyrms writes that "For Ramsey, the cunning bettor is a *dramatic* device and the possibility of a dutch book a striking *symptom* of a deeper incoherence" (Skyrms, 1987a, p. 227, emphases mine). For these authors,[37] susceptibility to Dutch Book merely brings out an underlying inconsistency in the agent's credences—the inconsistency of evaluating the same thing different ways depending on how it's presented.

Return to my bets on whether I'll go into the office tomorrow. Given my 0.7 confidence that I'll go, my fair betting price for a bet that pays \$1 if I go and nothing otherwise is \$0.70. So I value that bet at \$0.70; if I'm offered the opportunity to purchase that bet at any lower amount—say, \$0.50—I'd consider that a favorable deal. On the other hand, my 0.7 confidence that I won't go gives me a fair betting price of \$0.70 for a bet that pays \$1 if I *don't* go and nothing if I do. So I would consider it unfavorable to sell that bet at any price less than \$0.70—for instance, \$0.50. Yet buying the first bet at \$0.50 and selling the second bet at \$0.50 *are the exact same transaction*; each one would net me \$0.50 if I go to the office and lose me \$0.50 if I don't. So do I view that transaction favorably or not? One of my credences suggests I view it favorably, while the other demands I don't. How those credences evaluate those bets reveals the conflict between them.[38]

Still, even depragmatized Dutch Book arguments make potentially controversial assumptions. First, we're assuming that a rational agent's fair betting prices equal her expected payouts—an assumption that might fail for risk-averse agents. And second, to construct a Dutch Book against some violations of Finite Additivity, we need to assume a "package principle"—that a rational agent's fair betting price for a combination of two bets equals the sum of her betting prices for each bet considered singly. Each of these assumptions would follow easily if we assumed that rational agents always choose to maximize expected utility. But if we could assume *that*, we'd already have a representation-theorem argument for something very close to probabilism (Section 3.1).[39] So it's unclear why the detour through cunning bettors would be required.

37 See also Armendt (1992), Christensen (2004), and Howson and Urbach (2006).

38 Notice that I wouldn't have this problem if I satisfied the probability calculus by, say, assigning credence 0.7 that I'll go and credence 0.3 that I won't. In that case I'll look favorably on buying the first bet at \$0.50 and also look favorably on selling the second one at \$0.50, so my evaluations will harmonize.

39 In fact, the representation theorem proof in the appendix closely mirrors the structure of traditional Dutch Book theorems for probabilism.

3.3 *Accuracy Arguments*

In his 1998, James M. Joyce sets out to provide a "nonpragmatic vindication of probabilism" that would explicitly avoid invoking practical consequences in its defense of the probability axioms as rational constraints on credence. His work builds on mathematical results from de Finetti (1974) and Rosenkrantz (1981), but uses those results to construct a new kind of argument.

Joyce's key idea is that from a point of view of pure theoretical rationality, agents should aim to make their credences as accurate as possible. How might we measure the accuracy of a credence function? Historically, one option had been to focus on calibration. Function c is perfectly calibrated if, for every $0 \leq x \leq 1$, when we look at all the propositions in $\mathcal{L}$ to which c assigns credence x, the fraction of those propositions that are true is exactly x. If I'm perfectly calibrated, exactly half of the propositions to which I assign credence $1/2$ are true, exactly a third of the propositions to which I assign credence $1/3$ are true, etc.

Van Fraassen (1983) and Shimony (1988) argue for probabilism by showing that in order for a credence distribution to be embeddable in larger and larger systems approaching perfect calibration, that credence distribution must satisfy the probability axioms. This might stand as a good argument for probabilism, except that calibration has some intuitively undesirable features as a measure of accuracy. For example, consider two agents who assign credences to four propositions as in Table 1. I hope you'll agree

	A	*B*	*C*	*D*
Agent 1	0.5	0.5	0.5	0.5
Agent 2	1	1	0.01	0
Truth-values	T	T	F	F

Table 1: Two credence assignments

that intuitively, Agent 2's credences are much more accurate (close to the truth) than Agent 1's. Yet Agent 1 is perfectly calibrated—exactly half the propositions to which she assigns credence $1/2$ are true—while Agent 2 is not.

Our intuitions about accuracy work by looking at each credence assignment one at a time, assessing how accurate that credence is given the truth-value of the proposition, and then aggregating those local accuracy assessments across all the propositions. Yet calibration works with *global* features of a probability distribution, which (as we've just seen) can lead to distorting effects.

So Joyce uses a gradational accuracy approach instead. On this approach, we select a scoring rule to measure how far each individual credence assignment to a proposition is from the truth about that proposition. Intuitively, when proposition P is true, higher credences in P are more accurate; when P is false, lower credences are better. We can formalize this by having a function I that assigns 1 to P if it's true and 0 if it's false, then measuring how far $c(P)$ is from $I(P)$. Historically, it's been popular to measure this distance as

$$(I(P) - c(P))^2. \tag{24}$$

Notice that this measurement *increases* the *farther* you are from the truth; so it's a measure of credal *in*accuracy. A rational agent aiming to be as accurate as possible should look to minimize this quantity for each proposition. Globally, she should look to minimize the sum of this quantity across all the propositions she entertains. (This sum is commonly known as the *Brier score*, named for meteorologist George Brier's discussion of it in his 1950.)

Joyce shows that if we use the Brier score to measure accuracy, then any non-probabilistic credence distribution would be accuracy-dominated by another, probabilistic distribution over the same set of propositions. That is, if you take an agent whose credences over some language violates the probability axioms, there will be another, probabilistic credence distribution over the same language that has a more accurate Brier score than hers *in every possible world*. When the nonprobabilistic agent considers that alternative distribution, she will know that it's more accurate than hers, even without knowing anything about which possible world is actual. Joyce argued that for an agent to maintain her nonprobabilistic distribution, despite this information that another distribution was certainly more accurate, would be irrational. And since the same situation will confront any agent whose credences violate the probability axioms, this constitutes an argument for probabilism.[40]

Related accuracy arguments have been offered for a variety of other Bayesian norms: Conditionalization (Greaves & Wallace, 2006; Briggs & Pettigrew, forthcoming), the Principal Principle (Pettigrew, 2013), the Principle of Indifference (Pettigrew, 2014), Reflection (Easwaran, 2013), and Conglomerability (Easwaran, 2013).

There are two main concerns in the literature about these accuracy arguments. First, there's a general concern about assessing the rationality of credences by measuring their distance to the truth. The gradational accuracy approach evinces a sort of epistemic consequentialism, in which

40 Importantly, the same kind of argument cannot be run *against* probabilism. A credence function that satisfies the probability axioms will not be accuracy-dominated in the manner Joyce describes by any other function (probabilistic or otherwise).

attitudes aim for some outcome (in this case, truth), and are evaluated by how well they approximate that goal. Just as teleological approaches to normativity have aroused suspicion in ethics and other areas of philosophy, the gradational accuracy program has been criticized by such authors as Greaves (2013), Berker (2013), and Carr (2017).

Second, among those who accept the gradational accuracy program, there's a concern about how to select an appropriate scoring rule for measuring accuracy. Maher (2002) suggests that instead of using the Brier score, we might gauge the distance between an individual credence $c(P)$ and a truth-value $I(P)$ by calculating

$$|I(P) - c(P)|. \tag{25}$$

Historically, the Brier score was favored over this absolute-value score because the former is a "proper" scoring rule while the latter is not. To understand the difference, suppose a six-sided die has just been rolled, and we have two characters who do not yet know the outcome. Our first character, Chancey, assigns credence $1/6$ to each of the possible outcomes. Our second character, Pessimist, assigns credence 0 to each outcome. Chancey's credence function satisfies the probability axioms, while Pessimist's does not.

Now suppose each of our characters calculates an expected inaccuracy value for herself and for the other person. To give an example of how this works, suppose Chancey calculates an expected inaccuracy value for her own distribution using the Brier score. To do so, Chancey considers each of the six possible worlds available (that is, each of the six possible outcomes of the die roll), evaluates what her Brier score would be in that possible world, multiplies by her credence that that possible world is actual, then sums across all the possibilities. If, for instance, the die roll comes up 3, Chancey's Brier score will be

$$\begin{aligned}
&(I(1) - c(1))^2 + (I(2) - c(2))^2 + (I(3) - c(3))^2 \\
&\qquad + (I(4) - c(4))^2 + (I(5) - c(5))^2 + (I(6) - c(6))^2 \\
=\ &(0 - 1/6)^2 + (0 - 1/6)^2 + (1 - 1/6)^2 \\
&\qquad + (0 - 1/6)^2 + (0 - 1/6)^2 + (0 - 1/6)^2 \\
=\ &1/36 + 1/36 + 25/36 + 1/36 + 1/36 + 1/36 \\
=\ &30/36 \\
=\ &5/6.
\end{aligned} \tag{26}$$

A bit of reflection will show that this is Chancey's Brier score in each of the six possible worlds. So her expected Brier score across all those worlds is also $5/6$. In the meantime, I'll leave it to the reader to calculate that Pessimist's expected Brier score is 1. Since higher scores mean more

inaccuracy—and less accuracy—Chancey expects her credences to be more accurate than Pessimist's when the Brier score is used to calculate accuracy.

Exactly the opposite happens if we use the absolute-value measure. Again, I'll leave it to the reader to calculate that Chancey's expected absolute-value score is 5/3, while Pessimist's is again 1. So by the lights of the absolute-value score, the nonprobabilistic Pessimist is expected to be more accurate than the probabilistic Chancey.

Proper scoring rules are rules on which a probabilistic agent will never expect some other agent to be more accurate than herself. The Brier score is one of many proper scoring rules, while the absolute-value score is improper. In general, it seems irrational for an agent to hold onto a credence distribution when she expects some other agent's credences to be more accurate than her own (Lewis, 1971). So a theorist who has already accepted that probabilistic distributions are rational has good reason to work with proper scoring rules rather than improper ones. The accuracy-based arguments for Conditionalization, the Principal Principle, the Indifference Principle, etc. mentioned above all confine themselves to working with proper scoring rules.

Predd et al. (2009) show that Joyce's accuracy-dominance argument for probabilism could be run using any proper scoring rule. Yet in the context of an argument for probabilism, favoring proper scoring rules over improper ones seems question-begging. Proper scoring rules are defined as those on which probabilistic distributions are rated more expectedly accurate than the alternatives. Unless you have an antecedent reason to think probabilistic distributions *should* come out looking better than the alternatives, this is no reason to prefer a proper score.[41]

4 ARGUMENTS AGAINST CREDAL CONSTRAINTS

Having surveyed some arguments in *favor* of various rational constraints on credences, what are the arguments *against* these constraints? Of course there are many, and they multiply over time. Here I will focus on a handful that have generated insightful discussion and interesting positive responses.

4.1 *The Problem of Logical Omniscience*

Savage (1967) famously considered the plight of "a person required to risk money on a remote digit of π." His concern was that according to the Normality axiom, an agent is required to assign certainty to every tautology in her language $\mathcal{L}$. Arguably, the fact that a given digit of π

41 Though there may be other reasons. See, e.g., Joyce (2009) and Pettigrew (2016).

takes a particular value is a tautology.[42] So according to probabilism, a rational agent should be certain of all the digits of π. Yet this seems too much for rationality to demand of any real agent.

Savage's discussion initiated a literature on what is known as "the problem of logical omniscience." I actually think there are multiple, related problems here, which we might label as follows.[43]

CREDAL COMPLETENESS. Probabilism requires an agent to assign a credence to each proposition in her language.

LOGICAL DISCERNMENT. Probabilism forbids an agent from assigning a credence other than 1 to any tautology.

LOGICAL LEARNING. A probabilistic agent will never pass from a lower credence in a tautology to a higher credence.

The problem of Credal Completeness is that the probability axioms require an agent to assign a credence to every proposition in her language. For instance, Non-Negativity says that every $X \in \mathcal{L}$ receives some non-negative credence value. Even in a language with finitely many atomic propositions, closure under truth-functional connectives will generate a language of infinite size. Yet it seems not only impossible for a finite agent to assign that many credences, but also inadvisable under Harman's (1986) principle of Clutter Avoidance.

CLUTTER AVOIDANCE. One should not clutter one's mind with trivialities.

Yet we can slightly alter our formalism so that it no longer demands credal completeness and evades clutter avoidance concerns. The idea is to require not that an agent's credence distribution actually satisfy the probability axioms, but only that it be *extendable* to a distribution that does. In other words, we permit an agent to adopt a partial credence distribution that assigns numerical values to only some of the propositions in $\mathcal{L}$, but we require that there be some possible way of assigning values to the rest of $\mathcal{L}$ so that the resulting full distribution satisfies the axioms. This approach recovers intuitive results such as the stricture that if an agent assigns credences to both P and $\sim P$, those credences must sum to 1. But it will not fault an agent if she fails to adopt attitudes towards P, $\sim P$, or both.

Moving to partial distributions avoids the problem of Credal Completeness, but leaves the problem of Logical Discernment intact. It seems

42 If your views about logicism in the philosophy of mathematics entail that facts about digits of π are not tautologies, we can always substitute in a conditional whose antecedent is various arithmetic axioms and whose consequent reports a digit of π. Or we can work instead with some highly complex logical truths.

43 The "Logical Learning" label is common in the literature; I invented the other two labels for our discussion here.

perfectly rational for me to assign credence 1/10 that the trillionth digit of π is a 2. Yet any credence distribution—partial or complete—containing that assignment is not extendable to a probabilistic distribution. It's either a tautology that the billionth digit is a 2, or it's a tautology that the billionth digit isn't, so probabilism either demands that I assign that proposition a credence of 1 or demands that I assign it a credence of 0. Whichever is the true demand, it seems a bit too demanding, since I don't have any good way to figure out which demand it is.

Before considering responses to this problem of Logical Discernment, let's quickly consider Logical Learning. The following credal sequence seems quite reasonable: I assign credence 1/10 that the trillionth digit of π is a 2, Talbott (1991) tells me that it is indeed a 2, so my credence that it is dramatically increases (perhaps all the way to 1). It seems in this case that I have learned a logical truth, and my credal increase is a rational response to that learning episode. Yet a traditional Bayesian system will not approve of this response, or be able to usefully model it, since a probabilistic system countenances only credence distributions (at any time) that assign that proposition a value of 1.

If we solved the Logical Discernment problem by building a Bayesian theory that allowed rational credences in tautologies other than 1, presumably that theory would also allow increases and decreases in such credences. So there's hope that a solution to Logical Discernment would open up a solution to Logical Learning.

How, then, might we model a Bayesian agent without perfect logical discernment? Responding to Savage, Hacking (1967) suggests we identify a proposition as "personally possible" for an agent if the agent doesn't know it's false. We then adjust Normality to demand certainty only in propositions whose negations are personally impossible, and Finite Additivity to apply only when $P \,\&\, Q$ is personally impossible. This allows an agent to be ignorant of arbitrarily many logical truths, and therefore less-than-certain of those truths.

Yet this approach creates three problems. The first is formal. Hacking works with credence distributions over sentences, and he's free to treat whatever sentences he wants as personally possible or impossible. But if we think of those sentences as representing underlying propositions, and those propositions in turn as representing underlying sets of possibilities, it seems natural to ask what possibilities an agent entertains when she entertains as personally possible that which is logically impossible. To address this sort of gap, Hintikka (1975) constructs a semantics admitting of logically impossible worlds, which can enter into the content of propositions in just the manner of classical possible worlds.

A second, intuitive problem is that Hacking's approach allows for arbitrarily large amounts of logical non-omniscience—nothing in Hacking's

formalism indicts an agent who assigns less-than-certainty to $P \vee \sim P$, as long as that agent doesn't know the proposition is true. Bjerring and Skipper (manuscript) complain Hacking's formalism is so permissive that in sacrificing logical omniscience, it fails to capture any rational requirement of basic logical competence. They make similar complaints about a framework from Garber (1983), and various formalisms developed using Hintikka's semantics.

Finally, it's important to see what a Bayesian system loses when it's redefined in terms of personal rather than logical possibility. If an agent fails to know that $P \,\&\, \sim P$ is impossible, then by Hacking's lights she need not apply Finite Additivity to P and $\sim P$. As a result, such an agent may assign P and $\sim P$ credences summing to more than 1. She may increase her credence in P without decreasing her credence in $\sim P$. In our relevance-based theory of confirmation, she may not see P as disconfirming $\sim P$. And when she selects actions by maximizing expected epistemic utility, she may violate the preference axioms in a variety of ways. In other words, the very features and applications that make Bayesianism a plausible picture of rationality begin to dissolve once logical discernment requirements are loosened.

So perhaps we should go in the other direction? A number of theorists have begun to wonder if logical omniscience requirements are not an annoying side-effect of our epistemic formalisms, but instead a hint from those formalisms about the underlying normative domain. Smithies (2015) argues that certainty in logical truths is in fact a requirement of rationality; Titelbaum (2015) and Littlejohn (2018) advocate related positions.

4.2 *The Problem of Old Evidence*

Clark Glymour initiated the Old Evidence debate with a famous example.

> Scientists commonly argue for their theories from evidence known long before the theories were introduced. . . . The argument that Einstein gave in 1915 for his gravitational field equations was that they explained the anomalous advance of the perihelion of Mercury, established more than half a century earlier. Other physicists found the argument enormously forceful, and it is a fair conjecture that without it the British would not have mounted the famous eclipse expedition of 1919. Old evidence can in fact confirm new theory, but according to Bayesian kinematics, it cannot. (Glymour, 1980, pp. 306–7)

We've already seen (Section 1.3 and 1.4) that a traditional Bayesian models evidence acquisition as the gaining of certainties, which are then retained. At the same time (Section 2.1), confirmation is understood as positive

relevance. Combining these two approaches, we have a problem: once an evidential proposition has been learned, it receives credence 1. When $c(E) = 1$, $c(H \mid E) = c(H)$ for any $H \in \mathcal{L}$. So once an agent learns something, that piece of information is confirmationally inert ever after.

Given these basic facts about Bayesianism, we can identify two challenges in Glymour's story about Einstein. Christensen (1999) calls them the "synchronic" and "diachronic" problems of old evidence.[44] The diachronic problem is about changes in credence. Over the course of 1915, Einstein increased his confidence in the General Theory of Relativity (GTR), and we think this had something to do with the perihelion of Mercury. Yet it can't be that Einstein increased his confidence because he learned of the anomalous advance—he already knew about that well before 1915. So what changed his opinion, and how can we reflect it in a Bayesian system?

The synchronic problem of old evidence comes up after 1915, when the perihelion of Mercury has already had its effect on Einstein's attitudes toward GTR. Presumably even after 1915, Einstein would have cited the perihelion advance of Mercury as a crucial piece of evidence supporting GTR. Yet relative to Einstein's credence function at that time—which assigns 1 to the perihelion facts—those facts are not positively relevant to GTR. So how can a Bayesian about confirmation interpret that evidential support?

Proposals to solve the synchronic problem usually work by relativizing confirmation to some probability function other than the agent's current credence distribution. Since the agent currently assigns $c(E) = 1$, E can't confirm anything relative to that current distribution. So we look for some other relevant distribution that doesn't assign 1 to E. For instance, we might adopt a "historical backtracking" approach on which we look back to some time when the agent wasn't yet certain of E, and ask whether E was positively relevant to H in her credence distribution at that time. But this approach is limited for a number of reasons. For instance, Einstein probably knew about the perihelion of Mercury long before he ever considered GTR. So if we backtrack to a time well before 1915 when he wasn't yet certain of E, we won't be able to find any conditional or unconditional credences he assigned to the relevant H at that time. And so we won't be able to say that E confirms H for Einstein now because at some time in the past he assigned $c(H \mid E) > c(H)$.

In light of this and other difficulties, Howson and Urbach (2006) advocate a "counterfactual backtracking" approach. Instead of looking to a time in the past when the agent didn't know E, we look to a close possible world in which the agent knows everything she knows now except E. Well,

44 I'm using Christensen's terminology because I find it the most helpful. But earlier, related disambiguations of the Problem of Old Evidence can be found in Garber (1983), Eells (1985), and Zynda (1995).

not quite *everything*—we will probably also want a world in which she doesn't know logical equivalents to E, immediate entailments of E, etc. But Howson and Urbach (p. 300) have a technical proposal for identifying the propositions that should be subtracted out. Setting the technical details aside, Earman (1992, p. 123) worries this counterfactual approach will suffer from similar defects to other counterfactual analyses; moving to a non-actual world may have side-effects that spoil the analysis. For example, the historical record suggests that Einstein was motivated to formulate GTR in part to explain Mercury's anomalous advance. So the closest possible world in which Einstein doesn't know E yet still assigns credences to H may be very far—and very different from our own—indeed.

Perhaps the best approach is to say that when an agent explains the evidence supporting some hypothesis, the support she's describing may be relative not to her own personal credences but to some other probabilistic distribution. That distribution may be one assumed pertinent by her audience, or by a particular scientific community. Or if we are Objective Bayesians (Section 1.5), it may be the objective distribution that determines how all rational agents should set their credences. Maher (1996), for instance, develops a proposal of the latter sort. Yet many details remain to be resolved. For example, how does either a scientific community or an objective rational distribution assign a prior probability to the proposition that GTR expresses the physical laws of our universe?[45]

As for the diachronic problem of old evidence, the typical response is to identify something *other* than learning of Mercury's perihelion advance that gave Einstein new confidence in GTR over the course of 1915. For one, Einstein might have discovered sometime in 1915 not that Mercury's perihelion advances anomalously, but that GTR predicts such an anomalous advance. Since it's a logical fact that GTR (along with other empirical information of which Einstein was already aware) entails the details of the advance, this would be an instance of logical learning. So a Bayesian implementation of this explanation will depend on the logical omniscience issues discussed in Section 4.1.

Another possibility is that Einstein's high confidence in GTR at the end of 1915 was new because he hadn't had *any* attitude towards GTR at the beginning of 1915. Perhaps Einstein hadn't yet conceived of GTR at the beginning of 1915, so the language over which he assigned credences at that time didn't contain a proposition expressing GTR's truth. This approach would certainly explain why Einstein had a new, high credence at the end of the year that he didn't have at the beginning. But it probably doesn't generalize to all cases of confirmation by old evidence (and may not

45 Even if we wanted to use an Indifference Principle (Section 1.5) here, we'd need a partition to divide our credence evenly across, and it's difficult to determine what alternative sets of physical laws should go into such a partition.

even be historically accurate in Einstein's case). Moreover, cases in which agents add new propositions to their cognitive language pose another challenge for Bayesianism. All of the updating norms we have considered (Conditionalization, Jeffrey Conditionalization) worked over a language that remains fixed over time. The so-called "problem of new theories" challenges us to build a formalism that allows an agent's language to change over time, and that places reasonable constraints on how the agent's credences should evolve across such changes.

Finally, we might focus on the fact that both versions of the problem of old evidence seem to arise because Conditionalization treats acquiring evidence as gaining certainties. If newly-acquired evidence didn't go to (and remain at) a credence of 1, then we wouldn't have the problem that old evidence always has credence 1 and therefore can't be positively relevant to anything. Suppose we adopt the Regularity principle (forbidding certainty in empirical propositions), and mandate Jeffrey Conditionalization as the rational updating scheme. Then evidence acquisition will increase credence in particular propositions, but never send it to 1, and the problem of old evidence will never arise.

Christensen (1999) pursues this approach and finds much to recommend it, but eventually encounters a new difficulty. The problem of old evidence is that acquiring a piece of evidence shouldn't rob it of its ability to confirm hypotheses. Generalizing this idea, we should agree that becoming more *confident* in a piece of evidence shouldn't affect the *degree* to which it confirms a hypothesis. So Christensen seeks a confirmation measure (Section 2.1) on which Jeffrey Conditionalizations that change $c(E)$ don't affect E's level of confirmation of H. He is unable to find a measure that satisfies this constraint, meets other plausible formal conditions, and works intuitively in examples.

4.3 *Memory Loss and Context-sensitivity*

Certainty acquisition and retention also pose other problems for a Conditionalization-based updating framework. For instance, many of us have the experience of gaining a piece of evidence one day and then forgetting it a short time later. Yet if we are constant conditionalizers, a proposition that achieves credence 1 at some time may never sink to a lower credence later. So Conditionalization deems memory loss irrational.[46]

46 Or at least, the version of Conditionalization we've been discussing deems memory loss irrational, because it governs an agent's updating across any arbitrary interval of times t_i to t_j. One might embrace a more limited version of Conditionalization (compare Titelbaum, 2013a, Chapter 6) that applies only across intervals during which the agent's information

While this problem was recognized at least as far back as Levi (1987), Talbott (1991) puts it particularly forcefully. He considers the response that Bayesian rules are meant to model ideally rational agents—not everyday agents—"and an ideally rational agent would not be subject to the fallibility of human memory." (p. 141) For what it's worth, I don't see why elephantine recall should make one agent more *rational* than another (though see Carr, 2015), but the whole question may be sidestepped by an ingenious example due to Arntzenius (2003). While I won't work through the details here, the upshot of Arntzenius's example is that Conditionalization indicts not only agents who actually forget evidence, but also agents who suspect they might have forgotten evidence (even if they actually haven't). Surely we can't require of ideally rational agents certainty of the empirical proposition that they have never forgotten anything in their lives!

Can we alter Conditionalization to allow for certainty loss? One popular approach is to take advantage of a feature traditional Conditionalization already displays. Suppose we have an agent who conditionalizes throughout her entire life. As she gains evidence, she will accumulate certainties; the total set of certainties she possesses at any time will represent her total evidence at that time. Let's refer to the proposition expressing the conjunction of all the agent's evidence/certainties at time t_i as E_i. If the agent is a faithful conditionalizer, there will exist at least one regular[47] probability distribution p_h such that for any time t_i at which that agent assigns credences, and any proposition X in her language $\mathcal{L}$, $c_i(X) = p_h(X \mid E_i)$. In other words, there exists a single function p_h relating to every moment in the agent's life, such that her credence distribution at any moment can be recovered by conditionalizing p_h on her total evidence at that moment.

I'll refer to this distribution p_h as the agent's *hypothetical prior*; it is sometimes also called an "*ur*-prior" or an "initial credence distribution." This last moniker comes from thinking of p_h as representing the agent's credences at some earliest moment in her life when she lacked any empirical certainties. Because conditionalization is cumulative and commutative, if an agent *did* have such an initial moment in her life—before her first update by Conditionalization—the credences she assigned at that time *would* relate to her later opinions in the way that p_h relates to c_i. Yet it's difficult to imagine that any actual agent has ever had a moment when she entirely lacked empirical information.

So I prefer to think of an agent's hypothetical prior as a convenient tool for separating out two influences on her credences. On the one hand,

strictly increases. In that case the problem would be that rather than deeming memory loss irrational, the limited updating rule fails to give any guidance in memory loss cases at all.

47 By saying the distribution is "regular," we mean that it assigns credence 1 only to logical truths.

there's her evidence; on the other, there are her epistemic standards, which encapsulate her principles and tendencies for interpreting evidence. The agent's total evidence changes over time, and is represented at time t_i by E_i. Yet as her evidence changes, she may retain a constant set of standards for interpreting evidence, represented by her hypothetical prior p_h. Applying these standards to the agent's total evidence at t_i—by conditionalizing p_h on E_i—yields her credence distribution c_i.[48]

This generally attractive picture is entailed by Conditionalization: *if* an agent conditionalizes at every update, *then* her credences throughout her life will be representable as faithful to a constant hypothetical prior. Yet interestingly, the entailment does not run in the opposite direction. That is, an agent may maintain fealty to a constant hypothetical prior even if her updates do not always satisfy Conditionalization. For instance, it's possible that an agent could both gain and lose certainties between two times t_i and t_j, and yet there still exists a single hypothetical prior p_h such that for every $X \in \mathcal{L}$, $c_i(X) = p_h(X \mid E_i)$ and $c_j(X) = c_j(X \mid E_j)$.

We can therefore achieve a plausible diachronic model of agents who both gain and lose certainties by generalizing Conditionalization not to demand that an agent conditionalize between each earlier time t_i and later time t_j, but instead to demand (whatever happens to her certainties) that she set her credences in line with a constant hypothetical prior throughout her life. This new diachronic norm generates plausible results for a number of forgetting stories, such as those featured by Talbott. In cases where an agent does strictly gain certainties between two times, it mimics the effects of traditional Conditionalization. And in cases where an agent strictly *loses* certainties between times, it gives us reverse-temporal Conditionalization. That is, the agent's earlier unconditional credences will equal her later credences conditional on the information she lost. Thus forgetting becomes like learning backwards in time.

Unfortunately, shifting to this new diachronic norm does not suffice alone to address another problem with Conditionalization: the way it treats context-sensitive information. Here I refer to "self-locating" claims that change their truth-values across times, persons, and locations—such as "Today is Tuesday," "I am a sailor," and "We are in Detroit." For one thing, to model these sorts of claims in our formalism we will need to add to our language $\mathcal{L}$ something like what Lewis (1979) called "centered propositions." But even then, Conditionalization will face challenges. It may be rational right now to be certain that it's Tuesday, but that certainty will not remain rational into perpetuity.

The context-sensitivity challenge is sometimes described as yet another problem with Conditionalization's certainty-retention. But even when we shift to a diachronic norm that requires fealty only to a constant

48 Compare Schoenfield (2014) and Meacham (2016).

hypothetical prior (and therefore allows for certainty loss), problems still remain. This is because the Bayesian system was designed to model agents whose evidence changed over time, but who used that evidence to evaluate hypotheses with truth-values that were fixed targets.[49] Adding in another level of shiftiness generates complications for Conditionalization, Jeffrey Conditionalization, or hypothetical priors.

A number of formal frameworks have been proposed to model credence updates in context-sensitive propositions. Some retain Conditionalization, some make use of hypothetical priors, but in every case new, additional norms are required to capture the full range of phenomena. There isn't space to survey the various approaches here.[50] But I will note that solving the problem of updating self-locating beliefs may have important consequences beyond fun philosophical thought-experiments like the Sleeping Beauty Problem (Elga, 2000). For instance, fine-tuning arguments for the existence of the multiverse, and debates about the proper interpretation of quantum mechanics, may both turn on how agents should manage credences in context-sensitive propositions.[51]

5 OTHER CONFIDENCE FORMALISMS

In closing, I should note that there are a number of alternative formalisms for modeling agents' varying levels of confidence in claims. First, we can think simply about whether an agent is more confident in one proposition than another. Composing these comparisons together yields a confidence ordering that may float free of any numerical assignments (see Konek, this volume). A second approach, called "ranking theory" (Spohn, 2012; Huber, this volume), attaches numbers to the confidence ranking but works only with the structure of non-negative integers. Third, we can employ a formal structure even richer than the reals. For instance, instead of representing an agent's levels of confidence at a given time with a single probability distribution, we may represent them with a *set* of such distributions (Mahtani, this volume). Or we may have one real-valued function to track the agent's attitudes and a separate (though related) one to track her evidence. This yields a fourth approach, commonly called "Dempster-Shafer Theory" (Dempster, 1966; Shafer, 1976).

Each of these approaches may be supported by some of the argument-types described above, and each is plagued by some of the problems above as well. Some allow formal structures more flexible and expressive than

49 In philosophy of science applications, for instance, scientific hypotheses about the physical laws of the universe or the evolutionary origins of hominids do not typically change their truth-values over time.

50 Titelbaum (2016) provides a big-picture summary with copious references.

51 For these applications and others, see Titelbaum (2013b).

Bayesianism, while some trade expressive power for added psychological plausibility. I will not attempt to choose a favorite here. But it's worth noting that among all the formalisms for representing disparate confidence levels, none is currently more studied or more often applied than the real-valued credal approach.[52]

6 APPENDIX

Here's a proof sketch for the representation theorem mentioned in Section 3.1. We will assume that in the decision theory of interest, the following hold.

- Structural axioms ensuring that betting acts with various structures (as described in the proof below) are always available to the agent.
- Equivalence principle: when acts are independent of states, if two acts yield the same utility as each other in every possible world, the agent assigns them the same expected utility.
- Dominance principle: when acts are independent of states, if act A yields at least as great a utility as act B in some worlds, and a greater utility than B in at least one world, then $EU(A) > EU(B)$.

The equivalence and dominance principles each employ a notion of act-state independence, and the relevant notion will vary depending on which decision theory (evidential, causal, etc.) is in play. So fleshing out the proof below for a given decision theory will require showing that the acts and states appearing in each step of the proof are independent in the relevant sense. Given the types of acts involved, that should be fairly straightforward.

To show that the credence function c appearing in such a decision theory is a positive scalar transform of a probabilistic function, we need to prove that it satisfies four conditions.

1. Every tautology in $\mathcal{L}$ receives the same c-value.

   Proof. Suppose for reductio we have two tautologies $\mathcal{T}_1, \mathcal{T}_2 \in \mathcal{L}$ such that the agent assigns a credence of x to the first and y to the second. Consider an act that pays 1 util on $\mathcal{T}_1$ and 0 utils otherwise, and an act that pays 1 util on $\mathcal{T}_2$ and 0 utils otherwise. The agent will assign the first act an expected utility of x, the second act an expected utility of y. Yet the two acts each yield the same payout (1 util) in every possible world. So we've violated the equivalence principle.

52 Thanks to the editors, Richard Pettigrew and Jonathan Weisberg, and especially to the latter for detailed comments and many citation suggestions. Much of the material in this piece has been adapted from my forthcoming book (Titelbaum, forthcoming), which covers almost all of the topics here in much greater depth.

2. For any tautology and contradiction $\mathsf{T}, \mathsf{F} \in \mathcal{L}$, $c(\mathsf{T}) > c(\mathsf{F})$.

   Proof. Suppose for reductio we have a T and F such that $c(\mathsf{T}) \leq c(\mathsf{F})$. Now consider an act that pays 1 util on T and 0 utils otherwise, and another act that pays 1 util on F and 0 utils otherwise. The agent will assign the first act an expected utility no greater than the second, yet the first act dominates the second. So we've violated the dominance principle.

3. For any mutually exclusive $X, Y \in \mathcal{L}$, $c(X \vee Y) = c(X) + c(Y)$.

   Proof. First consider the act of purchasing a bet that pays 1 util on X, 1 util on Y, and 0 utils otherwise. Since X and Y are mutually exclusive, we may partition the possible states of the world into X, Y, and $\sim X \,\&\, \sim Y$. Using this partition, the expected utility of this act is:

   $$\begin{aligned} & c(X) \cdot u(X) + c(Y) \cdot u(Y) + c(\sim X \,\&\, \sim Y) \cdot u(\sim X \,\&\, \sim Y) \\ = \;& c(X) \cdot 1 + c(Y) \cdot 1 + c(\sim X \,\&\, \sim Y) \cdot 0 \\ = \;& c(X) + c(Y). \end{aligned} \tag{27}$$

   Now consider the act of purchasing a bet that pays 1 util on $X \vee Y$, and 0 utils otherwise. Partitioning the states into $X \vee Y$ and $\sim(X \vee Y)$, the expected utility of this act is:

   $$\begin{aligned} & c(X \vee Y) \cdot u(X \vee Y) + c(\sim[X \vee Y]) \cdot u(\sim[X \vee Y]) \\ = \;& c(X \vee Y) \cdot 1 + c(\sim[X \vee Y]) \cdot 0 \\ = \;& c(X \vee Y). \end{aligned} \tag{28}$$

   These two acts have the same payout in every possible world, so to satisfy the equivalence principle their expected utilities must be equal. Thus $c(X \vee Y) = c(X) + c(Y)$.

Notice it's a corollary of results (1) and (3) that for every contradiction $\mathsf{F} \in \mathcal{L}$, $c(\mathsf{F}) = 0$. The reason is that for any tautology $\mathsf{T} \in \mathcal{L}$, T and F are mutually exclusive, and $\mathsf{T} \vee \mathsf{F}$ is also a tautology. So by (1), $c(\mathsf{T}) = c(\mathsf{T} \vee \mathsf{F})$. But by (3), $c(\mathsf{T}) + c(\mathsf{F}) = c(\mathsf{T} \vee \mathsf{F})$, so $c(\mathsf{F}) = 0$.

Finally, we have our fourth result.

4. For any $X \in \mathcal{L}$, $c(X) \geq 0$.

   Proof. Given the corollary just proven, we only have to prove this for non-contradictory X. So suppose for reductio that for some non-contradictory $X \in \mathcal{L}$, the agent assigns $c(X) = x < 0$. Now consider a bet that pays 1 util if X, 0 utils otherwise. The agent's expected utility for this bet is x, which is negative. Compare with a constant act that yields 0 utils in every possible world. The agent's expected utility for the constant act is 0, which is greater than x. Yet the bet on X dominates the constant act, so we've violated the dominance principle.

REFERENCES

Adams, E. (1962). On rational betting systems. *Archiv für mathematische Logik und Grundlagenforschung*, *6*, 7–29.

Allais, M. (1953). Le Comportement de l'homme rationnel devant le risque: Critique des postulates et axiomes de l'ecole Américaine. *Econometrica*, *21*, 503–46.

Armendt, B. (1980). Is there a Dutch Book argument for probability kinematics? *Philosophy of Science*, *47*, 583–588.

Armendt, B. (1992). Dutch strategies for diachronic rules: When believers see the sure loss coming. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1*, 217–229.

Arntzenius, F. (2003). Some problems for conditionalization and reflection. *The Journal of Philosophy*, *100*, 356–370.

Berker, S. (2013). Epistemic teleology and the separateness of propositions. *Philosophical Review*, *122*, 337–93.

Bernoulli, J. (1713). *Ars conjectandi*. Basiliae.

Bjerring, J. C. & Skipper, M. (manuscript). *Bayesianism for average Joe*. Unpublished manuscript.

Boole, G. (1952). On the application of the theory of probabilities to the question of the combination of testimonies or judgments. *Studies in Logic and Probability*, 308–85.

Bovens, L. & Hartmann, S. (2003). *Bayesian epistemology*. Oxford: Oxford University Press.

Bradley, R. (2007). Reaching a consensus. *Social Choice and Welfare*, *29*, 609–32.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Briggs, R. (2009). Distorted reflection. *The Philosophical Review*, *118*, 59–85.

Briggs, R. & Pettigrew, R. (forthcoming). An accuracy-dominance argument for Conditionalization. *Noûs*.

Buchak, L. (2013). *Risk and rationality*. Oxford: Oxford University Press.

Carnap, R. (1950). *Logical foundations of probability*. Chicago: University of Chicago Press.

Carnap, R. (1962). *Logical foundations of probability* (2nd). Chicago: University of Chicago Press.

Carr, J. R. (2015). Don't stop believing. *Canadian Journal of Philosophy*, *45*.

Carr, J. R. (2017). Epistemic utility theory and the aim of belief. *Philosophy and Phenomenological Research*, *95*, 511–34.

Chihara, C. (1981). Quine and the confirmational paradoxes. In P. French, H. Wettstein, & T. Uehling (Eds.), *Midwest studies in philosophy 6: Foundations of analytic philosophy* (pp. 425–52). University of Minnesota Press.

Christensen, D. (1999). Measuring confirmation. *The Journal of Philosophy*, *96*, 437–61.

Christensen, D. (2004). *Putting logic in its place*. Oxford: Oxford University Press.

Christensen, D. (2009). Disagreement as evidence: The epistemology of controversy. *Philosophy Compass*, *4*, 756–67.

Christensen, D. & Lackey, J. (Eds.). (2013). *The epistemology of disagreement: New essays*. Oxford: Oxford University Press.

Crupi, V. (2016). Confirmation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016). Metaphysics Research Lab, Stanford University.

Davies, M. (1998). Exernalism, architecturalism, and epistemic warrant. In C. Wright, B. Smith, & C. Macdonald (Eds.), *Knowing our own minds* (pp. 321–61). Oxford: Oxford University Press.

de Finetti, B. (1937/1964). Foresight: Its logical laws, its subjective sources. In H. E. Kyburg Jr & H. Smokler (Eds.), *Studies in subjective probability* (pp. 94–158). Originally published as "La prévision; ses lois logiques, ses sources subjectives" in *Annales de l'Institut Henri Poincaré*, Volume 7, 1–68. New York: Wiley.

de Finetti, B. (1974). *Theory of probability*. New York: Wiley.

Dempster, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, *37*, 355–74.

Douven, I. (2012). The Lottery Paradox and the pragmatics of belief. *Dialectica*, *66*, 351–73.

Dretske, F. I. (1970). Epistemic operators. *The Journal of Philosophy*, *67*, 1007–1023.

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory*. Cambridge, MA: The MIT Press.

Easwaran, K. (2013). Expected accuracy supports conditionalization—and conglomerability and reflection. *Philosophy of Science*, *80*, 119–142.

Easwaran, K. (2014). Regularity and hyperreal credences. *Philosophical Review*, *123*, 1–41.

Easwaran, K., Fenton-Glynn, L., Hitchcock, C., & Velasco, J. D. (2016). Updating on the credences of others: Disagreement, agreement, and synergy. *Philosophers' Imprint*, *16*, 1–39.

Eells, E. (1982). *Rational decision and causality*. Cambridge Studies in Philosophy. Cambridge: Cambridge University Press.

Eells, E. (1985). Problems of old evidence. *Pacific Philosophical Quarterly*, *66*, 283–302.

Elga, A. (2000). Self-locating belief and the Sleeping Beauty problem. *Analysis*, *60*, 143–7.

Elga, A. (2007). Reflection and disagreement. *Noûs*, *41*, 478–502.

Feldman, R. (2007). Reasonable religious disagreements. In L. M. Antony (Ed.), *Philosophers without gods: Meditations on atheism and the secular life*. Oxford: Oxford University Press.

Fitelson, B. (2008). A decision procedure for probability calculus with applications. *The Review of Symbolic Logic*, *1*, 111–125.

Fitelson, B. (2015). The strongest possible Lewisian triviality result. *Thought*, *4*, 69–74.

Fitelson, B. & Hawthorne, J. (2010). How Bayesian confirmation theory handles the Paradox of the Ravens. *Boston Studies in the Philosophy of Science*, *284*.

Foley, R. (1993). *Working without a net*. Oxford: Oxford University Press.

Garber, D. (1983). Old evidence and logical omniscience in Bayesian confirmation theory. In J. Earman (Ed.), *Testing scientific theories* (Vol. 10, pp. 99–132). Minnesota Studies in the Philosophy of Science. Minneapolis: University of Minnesota Press.

Gibbard, A. & Harper, W. (1978). Counterfactuals and two kinds of expected utility. In C. A. Hooker, J. L. Leach, & E. F. McClennan (Eds.), *Foundations and applications of decision theory* (Vol. 13a, pp. 125–62). University of Western Ontario Series in Philosophy of Science. Dordrecht: D. Reidel Publishing Company.

Glymour, C. (1980). *Theory and evidence*. Princeton, NJ: Princeton University Press.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Greaves, H. (2013). Epistemic decision theory. *Mind*, *122*, 915–52.

Greaves, H. & Wallace, D. (2006). Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind*, *115*, 607–632.

Hacking, I. (1967). Slightly more realistic personal probability. *Philosophy of Science*, *34*, 311–325.

Hájek, A. (2003). What conditional probability could not be. *Synthese*, *137*, 273–323.

Hájek, A. (2007). The reference class problem is your problem too. *Synthese*, *156*, 563–85.

Hájek, A. (2009). Arguments for—or against—probabilism? In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (Vol. 342, pp. 229–251). Synthese Library. Springer.

Hájek, A. (2011). Triviality pursuit. *Topoi*, *30*, 3–15.

Harman, G. (1986). *Change in view*. Boston: The MIT Press.

Hawthorne, J. & Fitelson, B. (2004). Re-solving irrelevant conjunction with probabilistic independence. *Philosophy of Science*, *71*, 505–514.

Hempel, C. G. (1945). Studies in the logic of confirmation (I). *Mind*, *54*, 1–26.

Henderson, L. (2013). Bayesianism and inference to the best explanation. *British Journal for the Philosophy of Science*, *65*, 687–715.

Hintikka, J. (1975). Impossible possible worlds vindicated. *Journal of Philosophical Logic*, *4*, 475–84.

Hitchcock, C. R. (2012). Probabilistic causation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2012).

Holton, R. (2014). Intention as a model for belief. In M. Vargas & G. Yaffe (Eds.), *Rational and social agency: The philosophy of Michael Bratman* (pp. 12–37). Oxford: Oxford University Press.

Hosiasson-Lindenbaum, J. (1940). On confirmation. *The Journal of Symbolic Logic*, *5*, 133–48.

Howson, C. (1992). Dutch Book arguments and consistency. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *2*, 161–8.

Howson, C. & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd). Chicago: Open Court.

Huemer, M. (2011). Does probability theory refute coherentism? *Journal of Philosophy*, *108*, 463–72.

Jaynes, E. T. (1957a). Information theory and statistical mechanics i. *Physical Review*, *106*, 620–30.

Jaynes, E. T. (1957b). Information theory and statistical mechanics ii. *Physical Review*, *108*, 171–90.

Jeffrey, R. C. (1965). *The logic of decision* (1st). McGraw-Hill series in probability and statistics. New York: McGraw-Hill.

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, *65*, 575–603.

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge: Cambridge University Press.

Joyce, J. M. (2009). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (Vol. 342, pp. 263–297). Synthese Library. Springer.

Kemeny, J. G. (1955). Fair bets and inductive probabilities. *The Journal of Symbolic Logic*, *20*, 263–273.

Kolmogorov, A. N. (1933/1950). *Foundations of the theory of probability*. Translation edited by Nathan Morrison. New York: Chelsea Publishing Company.

Kyburg, H. E., Jr. (1961). *Probability and the logic of rational belief*. Middletown: Wesleyan University Press.

Laplace, P.-S. (1814/1995). *Philosophical essay on probabilities*. Translated from the French by Andrew Dale. New York: Springer.

Leitgeb, H. (2017). *The stability of belief: How rational belief coheres with probability*. Oxford: Oxford University Press.

Levi, I. (1987). The demons of decision. *The Monist*, *70*, 193–211.

Lewis, D. (1971). Immodest inductive methods. *Philosophy of Science*, *38*, 54–63.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, *85*, 297–315.

Lewis, D. (1979). Atittudes *de dicto* and *de se*. *The Philosophical Review*, *88*, 513–543.

Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. 2, pp. 263–294). Berkeley: University of California Press.

Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, *59*, 5–30.

Lin, H. & Kelly, K. T. (2012). A geo-logical solution to the Lottery Paradox. *Synthese*, *186*, 531–75.

Lipton, P. (2004). *Inference to the best explanation* (2nd). London: Routledge.

Littlejohn, C. (2018). Stop making sense? On a puzzle about rationality. *Philosophy and Phenomenological Research*, *96*(2), 257–272.

Locke, J. (1689/1975). *An essay concerning human understanding* (P. H. Nidditch, Ed.). Oxford: Oxford University Press.

Maher, P. (1996). Subjective and objective confirmation. *Philosophy of Science*, *63*, 149–74.

Maher, P. (2002). Joyce's argument for probabilism. *Philosophy of Science*, *96*, 73–81.

Makinson, D. C. (1965). The paradox of the preface. *Analysis*, *25*, 205–7.

Meacham, C. J. (2010). Two mistakes regarding the Principal Principle. *British Journal for the Philosophy of Science*, *61*, 407–31.

Meacham, C. J. (2016). Ur-priors, conditionalization, and ur-prior conditionalization. *Ergo*, *3*(17), 444–492.

Meacham, C. J. & Weisberg, J. (2011). Representation theorems and the foundations of decision theory. *Australasian Journal of Philosophy*, *89*, 641–663.

Moss, S. (2018). *Probabilistic knowledge*. Oxford: Oxford University Press.

Okasha, S. (2000). Van Fraassen's critique of inference to the best explanation. *Studies in History and Philosophy of Science*, 691–710.

Olsson, E. (2017). Coherentist theories of epistemic justification. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2017). Metaphysics Research Lab, Stanford University.

Pettigrew, R. (2013). A new epistemic utility argument for the Principal Principle. *Episteme*, *10*, 19–35.

Pettigrew, R. (2014). Accuracy, risk, and the Principle of Indifference. *Philosophy and Phenomenological Research*, *90*.

Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford: Oxford University Press.

Predd, J., Seiringer, R., Lieb, E., Osherson, D., Poor, V., & Kulkarni, S. (2009). Probabilistic coherence and proper scoring rules. *IEEE Transactions on Information Theory*, *55*, 4786–4792.

Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logic essays* (pp. 156–198). New York: Harcourt, Brace and Company.

Rasmussen, M. & Steglich-Petersen, A. (Eds.). (forthcoming). *Higher-order evidence: New essays*. Oxford: Oxford University Press.

Reichenbach, H. (1956). The principle of common cause. In *The direction of time* (pp. 157–160). University of California Press.

Rosenkrantz, R. (1977). *The paradoxes of confirmation*. Synthese Library. Dordrecht: D. Reidel Publishing Company.

Rosenkrantz, R. (1981). *Foundations and applications of inductive probability*. Atascadero, CA: Ridgeview Press.

Russell, J., Hawthorne, J., & Buchak, L. (2015). Groupthink. *Philosophical Studies*, *172*, 1287–1309.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Savage, L. J. (1967). Difficulties in the theory of personal probability. *Philosophy of Science*, *34*, 305–310.

Schoenfield, M. (2014). Permission to believe: Why permissivism is true and what it tells us about irrelevant influences on belief. *Noûs*, *48*, 193–218.

Seidenfeld, T. (1986). Entropy and uncertainty. *Philosophy of Science*, *53*, 467–491.

Seidenfeld, T., Schervish, M. J., & Kadane, J. (manuscript). Non-conglomerability for countably additive measures that are not κ-additive.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.

Shimony, A. (1955). Coherence and the axioms of confirmation. *Journal of Symbolic Logic*, *20*, 1–28.

Shimony, A. (1988). An Adamite derivation of the calculus of probability. In J. Fetzer (Ed.), *Probability and causality* (pp. 151–161). Dordrecht: Reidel.

Shogenji, T. (1999). Is coherence truth-conducive? *Analysis*, *59*, 338–45.

Skyrms, B. (1987a). Coherence. In N. Rescher (Ed.), *Scientific inquiry in philosophical perspective* (pp. 225–42). Pittsburgh: University of Pittsburgh Press.

Skyrms, B. (1987b). Dynamic coherence and probability kinematics. *Philosophy of Science*, *54*, 1–20.

Smithies, D. (2015). Ideal rationality and logical omniscience. *Synthese*, *192*, 2769–93.

Spohn, W. (2012). *The laws of belief: Ranking Theory & its philosophical applications*. Oxford: Oxford University Press.

Suppes, P. (1974). *Probabilistic metaphysics*. Uppsala: University of Uppsala Press.

Swinburne, R. (1971). The paradoxes of confirmation: A survey. *American Philosophical Quarterly*, *8*, 318–30.

Talbott, W. J. (1991). Two principles of Bayesian epistemology. *Philosophical Studies*, *62*, 135–150.

Teller, P. (1973). Conditionalization and observation. *Synthese*, *26*, 218–258.

Titelbaum, M. G. (forthcoming). *Fundamentals of Bayesian epistemology*. Oxford: Oxford University Press.

Titelbaum, M. G. (2013a). *Quitting certainties: A Bayesian framework modeling degrees of belief*. Oxford: Oxford University Press.

Titelbaum, M. G. (2013b). Ten reasons to care about the Sleeping Beauty Problem. *Philosophy Compass*, *8*, 1003–17.

Titelbaum, M. G. (2015). Rationality's fixed point (or: In defense of right reason). In T. S. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (Vol. 5, pp. 253–94). Oxford University Press.

Titelbaum, M. G. (2016). Self-locating credences. In A. Hájek & C. R. Hitchcock (Eds.), *The Oxford handbook of probability and philosophy* (pp. 666–680). Oxford: Oxford University Press.

van Fraassen, B. C. (1983). Calibration: A frequency justification for personal probability. In R. Cohen & L. Laudan (Eds.), *Physics philosophy and psychoanalysis* (pp. 295–319). Dordrecht: Reidel.

van Fraassen, B. C. (1984). Belief and the will. *The Journal of Philosophy*, *81*, 235–256.

van Fraassen, B. C. (1989a). *Laws and symmetry*. Oxford: Oxford University Press.

van Fraassen, B. C. (1989b). *Laws and symmetry*. Oxford: Clarendon Press.

von Mises, R. (1928/1957). *Probability, statistics and truth*. (English edition of the original German *Wahrscheinlichkeit, Statistik und Wahrheit*.) New York: Dover.

von Neumann, J. & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd). Princeton, NJ: Princeton University Press.

Weirich, P. (2012). Causal decision theory. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2012).

Weisberg, J. (2007). Conditionalization, reflection, and self-knowledge. *Philosophical Studies*, *135*, 179–197.

Weisberg, J. (2009). Locating IBE in the Bayesian framework. *Synthese*, *167*, 125–44.

White, R. (2005). Epistemic permissiveness. *Philosophical Perspectives*, *19*, 445–459.

Williamson, T. (2007). How probable is an infinite sequence of heads? *Analysis*, *67*, 173–80.

Wright, C. (2003). Some reflections on the acquisition of warrant by inference. In S. Nuccetelli (Ed.), *New essays on semantic externalism*. Cambridge, MA: MIT Press.

Zynda, L. (1995). Old evidence and new theories. *Philosophical Studies*, *77*, 67–95.

Zynda, L. (2000). Representation theorems and realism about degrees of belief. *Philosophy of Science*, *67*, 45–69.

2 DECISION THEORY

Johanna Thoma

Suppose I am deliberating whether I should live on a boat and sail the Caribbean for a year. This is a decision not to be taken lightly. Many factors will matter for my decision. Several of these depend on uncertain states of the world. Will I be able to make a living? Is my boat really seaworthy? Will I miss my friends? How bad will the next winter be in my home town?

1 DECISION PROBLEMS AND THE USES OF DECISION THEORY

Giving a decision problem like this some formal structure may be helpful for a number of interrelated purposes. As an agent, it might help me come to a better decision. But giving formal structure to a decision problem may also help a third party: prior to an action, it may help them predict my behaviour. And after the action, it may help them both understand my action, and judge whether I was rational. Moreover, giving formal structure to a decision problem is a pre-requisite for applying formal decision theories. And formal decision theories are used for all the aforementioned purposes.

In the case of the decision whether to live on a boat, we could perhaps represent the decision problem as shown in Table 1. In this matrix, the rows represent the actions I might take. In our case, these are to either live on a boat, or not to live on a boat. The columns represent the relevant states of the world. These are conditions that are out of my control, but matter for what I should do. Suppose these involve my boat either being seaworthy, or not being seaworthy. I am uncertain which of these states of affairs will come about. Finally, the entries in the matrix describe the possible outcomes I care about that would result from my action combined with a state of the world.

	Boat seaworthy	Boat not seaworthy
Live on a boat	Life on a boat, no storm damage	Life on a boat, storm damage
Stay in home town	Life as usual	Life as usual

Table 1: Should I live on a boat?

Since Savage's (1954) decision theory, it has become standard to characterise decision problems with state-outcome matrices like the one I just introduced. More generally, let $A_1 \dots A_n$ be a set of n actions that are open to the agent, and let $S_1 \dots S_m$ be m mutually exclusive and exhaustive states of the world. These actions and states of the world combine to yield a set of $n \cdot m$ outcomes $O_{11} \dots O_{nm}$. Table 2 shows this more general state-outcome matrix.

	S_1	$\dots$	S_m
A_1	O_{11}	$\dots$	O_{1m}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_n	O_{n1}	$\dots$	O_{nm}

Table 2: State-outcome matrix

Given such a representation of a decision problem, formal decision theories assume that agents have various attitudes to the elements of the state-outcome matrix. Agents are assumed to have preferences over the outcomes their actions might lead to. Depending on our interpretation of decision theory, we may also assume that agents can assign a utility value to the outcomes, and a probability value to the states of the world. Decision theories then require the preferences the agent has over actions, which are assumed to guide her choice behaviour, to relate to those other attitudes in a particular way.

1.1 *Expected Utility Maximisation*

Traditionally, the requirement that decision theories place on agents under conditions of uncertainty has been that agents should maximise their expected utility, or act as if they did. Decision theories which incorporate this requirement are known under the heading of 'expected utility theory'. In the special case where an agent is certain about the consequences of each of her actions, this requirement reduces to the requirement to maximise utility. Since we are always to some extent uncertain about the consequences of our actions, I will focus on the uncertain case here.[1] However, much of the following discussion will also apply to decision-

1 I understand decision-making under 'uncertainty' here to refer to any case where an agent is not certain what the consequences of her actions will be, or what state will come about. A distinction is sometimes made between risk, uncertainty, ignorance and ambiguity, where 'risk' refers to the case where objective probabilities are known, 'uncertainty' refers to the case where an agent can make a subjective judgement about probabilities, an agent is in a state of 'ignorance' if she cannot make such probability assignments, and 'ambiguity' occurs when an agent can make probability assignments for some states, but not others.

making under certainty. Moreover, most of this entry will focus on expected utility theory. Some alternative decision theories are discussed in Section 6.

As we will see, the requirement to maximise expected utility takes different forms under different interpretations of expected utility theory. For now, let us assume that agents can assign utility values $u(O)$ to outcomes, and probability values $p(S)$ to states of the world. The expected utility is then calculated by weighting the utility of each possible outcome of the action by the probability that it occurs, and summing them together. Expected utility theory instructs us to prefer acts with higher expected utility to acts with lower expected utility, and to choose one of the acts with the highest expected utility.

In our example, suppose that I think that the chances that my boat is seaworthy are 50%, and that the relevant utilities are the ones given in Table 3. In that case, the expected utility of living on a boat will be $0.5 \cdot 200 + 0.5 \cdot 20 = 110$, while the expected utility of staying in my home town is 100. I conclude I should live on a boat.

	Boat seaworthy	Boat not seaworthy	**EU**
Live on a boat	200	20	**110**
Stay in home town	100	100	**100**

Table 3: Decision problem with utilities

Formally, the expected utility $EU(A)$ of an action can be expressed as follows:

$$EU(A_i) = \sum_{j=1}^{m} p(S_j) \cdot u(O_{ij}).$$

Expected utility theory requires agents to prefer acts for which this weighted sum is higher to acts for which this weighted sum is lower, and to choose an action for which this weighted sum is maximised.

1.2 *The Uses of Decision Theory*

Now we can see how expected utility theory could be put to each of the different uses mentioned above. The requirement to maximise expected utility (or to act as if one did), however it is understood, is considered as a requirement of practical rationality by proponents of expected utility theory. In particular, the requirements of expected utility theory are often interpreted to capture what it means to be instrumentally rational, that

While these differences will play a role later in this entry, it is not helpful to make these distinctions at this point.

is, what it means to take the appropriate means to one's ends, whatever those ends may be. We will see how this may be cashed out in more detail in Section 3, when we discuss different interpretations of expected utility theory. For now, note that if we take the utility function to express the agent's ends, then the requirement to maximise the expectation of utility sounds like a natural requirement of instrumental rationality.

Sometimes, the requirements of expected utility theory are also understood as expressing what it means to have coherent ends in the first place. Constructivists about utility (see Section 3.1) often understand expected utility theory as expressing requirements on the coherence of preferences. But on that understanding, too, expected utility theory does not make any prescriptions on the specific content of an agent's ends. It merely rules out certain combinations of preferences. And so for those who think that some ends are irrational in themselves, expected utility theory will at best be an incomplete theory of practical rationality.

If we understand the requirements of expected utility theory as requirements of practical rationality, it seems like expected utility theory could help me as an agent make better decisions. After I have formally represented my decision problem, expected utility theory could be understood as telling me to maximise my expected utility (or to act as if I did). In the above example, we employed expected utility theory in this way. Expected utility theory helped me decide that I should live on a boat. In this guise, expected utility theory is an *action-guiding* theory.

From a third party perspective, expected utility theory could also be used to judge whether an agent's action was rational. Having represented the agent's decision problem formally, we judge an action to be rational if it was an act with maximum expected utility. This understands expected utility theory as a *normative* theory: a theory about what makes it the case that somebody acted rationally.

It is important to note the difference between the action-guiding and the normative uses of expected utility theory.[2] An action can be rational according to normative expected utility theory even if the agent did not use expected utility theory as an action-guiding theory. One could even hold that expected utility theory is a good normative theory while being a bad action-guiding theory. This would be the case if most agents are bad at determining their expected utility, and do better by using simpler heuristics.[3]

2 Herbert Simon famously drew attention to this difference when he distinguished between *procedural* and *substantive* rationality, drawing on a similar distinction made by Max Weber (1922/2005). See Simon (1976).

3 Starting with Tversky and Kahneman (1974), there has been a wealth of empirical literature studying what kind of heuristics decision-makers use when making decisions under uncertainty, and how well they perform. See, for instance, Payne, Bettman, and Johnson (1993) and Gigerenzer, Todd, and Group (2000).

Expected utility theory is also often put to an explanatory or predictive use, especially within economics or psychology. If we assume that agents follow the requirements of expected utility theory, and we know enough of their preferences or utility and probability assignments, we can use the theory to predict their behaviour. In this context, philosophers have been interested more in whether decision theory can help us *understand* an agent's actions. Interpreting an agent as maximising her expected utility in a formal decision problem may reveal her motives in action, and thus explain her action.

In fact, there is a tradition in the philosophy of action that claims that explaining another's behaviour always involves rationalising her behaviour to some extent. Davidson (1973) introduced the label 'radical interpretation' for the attempt to infer an agent's attitudes, such as her beliefs and desires, from her actions. He believed that this was only possible if we assume certain rationality constraints on how these attitudes relate. Ramsey (1926/2010) had already used expected utility theory to infer an agent's probabilities, and thus, he argued, her beliefs from her behaviour. Lewis (1974) showed that expected utility theory captures Davidson's constraints on the relationship between beliefs and desires, and thus can be used to elicit beliefs and desires. Davidson himself later argued, in Davidson (1985), that expected utility theory can be extended to further elicit an agent's *meanings*, that is, her interpretation of sentences. This is sometimes known as the *interpretive* use of decision theory.

And so in the philosophical literature, expected utility theory has been used as an action-guiding theory, a normative theory, and an interpretive theory.[4] Other decision theories have been put to the same uses. As we will see in Section 6, there are alternatives to expected utility theory that offer rival prescriptions of practical rationality. However, most alternatives to expected utility theory have been introduced as primarily descriptive theories, that are used to predict and explain behaviour that need not be rational.

Now that we have seen what kinds of uses expected utility theory can be put to, the next section will look at some influential applications of expected utility theory.

1.3 *Some Applications*

Expected utility theory has proven to be an enormously fruitful theory, that has been applied in various different fields and disciplines. Originally, it found application mostly in the theory of consumer choice. This field

4 Bermudez (2009) draws a similar tri-partite distinction between the normative, action-guiding and explanatory/predictive dimensions of decision theory. Similarly, Buchak (2016) distinguishes between the normative and interpretive uses of decision theory.

of economics studies why consumers choose some goods rather than others, and helps to predict market outcomes. Expected utility theory has been used to explain the shape of demand curves for goods. The demand for insurance, in particular, is difficult to understand without a formal theory of choice under uncertainty. Expected utility theory has also helped to explain some phenomena that had previously seemed surprising. A classic example here is adverse selection, which occurs when there is an information asymmetry between buyers and sellers in the market. In these kinds of situations, sellers of high quality goods may be driven out of the market. Akerlof (1970) first explained this phenomenon, and a rich literature has developed since. Einav and Finkelstein (2011) provide a helpful overview of work on adverse selection in insurance markets.

Decision theory has also found application in many fields outside of economics. For instance, in politics, it has been used to study voting and voter turn-out,[5] in law it has been used to study judicial decisions,[6] and in sociology it has been used to explain class and gender differences in levels of education.[7]

Expected utility theory has also been influential in philosophy. Apart from it being an important contender as a theory of practical rationality, expected utility theory plays an important role in ethics, in particular in consequentialist ethics. Along with Jackson (1991), many consequentialists believe that agents ought to maximise expected moral goodness. Moreover, expected utility theory has been applied to the question of what agents ought to do in the face of moral uncertainty—uncertainty about what one ought to do, or even about which moral theory is the right one.[8]

Recently, expected utility theory has found application in epistemology in the form of *epistemic decision theory*. Here, agents are modeled as receiving epistemic utility from being in various epistemic states, such as being certain of the proposition that my boat is sea-worthy. I will receive a high epistemic utility from being in that state in the case where my boat in fact turns out to be seaworthy, and low epistemic utility when my boat turns out not to be seaworthy. Agents are then modeled as maximising their expected epistemic utility. Epistemic utility theory has been used to justify various epistemic norms, such as probabilism (the norm that an agent's credences should obey the probability calculus), and conditionalisation (the norm that agents should update their credences by conditionalizing their old credence on the new evidence they received). For an overview of these arguments, see Pettigrew (2011).

5 Downs (1957) counts as the first systematic application of decision theoretic models from economics to politics. For recent work on voting specifically, see Feddersen (2004).

6 See, for instance, Epstein, Landes, and Posner (2013).

7 See, for instance, Breen and Goldthorpe (1997).

8 See, for instance, Lockhart (2000), and Sepielli (2013) for a criticism of Lockhart's approach.

1.4 *Formulating Decision Problems*

How should the decision problems that formal decision theories deal with be formulated in the first place? In order to apply a formal decision theory, the choices an agent faces need to already be represented as a formal decision problem. Table 1 offered one representation of my choice of whether to live on a boat. But how can we be sure it was the right one?

For his decision theory, Savage (1954) assumed that states are descriptions of the world that include everything that might be relevant to the agent. Similarly, he thought that descriptions of outcomes are descriptions of "everything that might happen to the person" (p. 13). Joyce (1999, p. 52) cashes out a rule for specifying outcomes that also appeals to relevance. He claims that a description of an outcome should include everything that might be relevant to the agent, in the following sense: whenever there is some circumstance such that an agent would strictly prefer an outcome in the presence of that circumstance to the same outcome in the absence of that circumstance, the outcome has been underspecified. Importantly, this implies that an agent's evaluation of an outcome should be independent of the state it occurs in, and the act that brought it about. All of this means that the sets of states and outcomes will end up being very fine-grained. Moreover, Savage also thinks of actions as functions from states to outcomes. This means that in each state, each action leads to a unique outcome. To ensure this, the set of actions, too, will have to be very fine-grained.

Note that this means that the decision problem I presented in Table 1 was hopelessly underspecified. When it comes to the decision of whether to live on a boat for a year or not, I do not only care about whether my boat will have storm damage or not. I also care, for instance, about whether I will have enough money for the year. I will evaluate the outcome "Life on a boat, no storm damage" differently depending on whether I will have enough money for the year or not. In fact, the exact amount of money I will have is going to matter for my decision. And so my decision problem should really distinguish between many different states of affairs involving me having more or less money, and the many different outcomes that occur in these states of affairs.

Jeffrey (1965/1983), who offered a famous alternative to Savage's decision theory (see Section 2.4), and treated states, acts, and outcomes all as propositions, went so far as to define outcomes such that they entail an act and a state. An act and a state are also supposed to entail the outcome, and so we can simply replace outcomes with the conjunction of an act and a state in the decision matrix.

These ways of individuating outcomes will obviously lead to very large decision matrices for any real life decision. There are two reasons why we

might find this problematic. The first reason has to do with the efficiency of the decision-making process. If we want our decision theory to be an action-guiding theory, then decision problems can't be so complex that ordinary agents cannot solve them. An action-guiding theory should be efficient in its application. Efficiency may also be a concern for the interpretive project. After all, this project wants to enable us to interpret each other's actions. And so doing so should not be overly complicated.

Savage called decision problems that specify every eventuality that might be relevant to an agent's choice "grand world" decision problems. Joyce (1999) holds that we should really be trying to solve such a grand-world problem, but acknowledges that real agents will always fall short of this. Instead, he claims, they solve "small world" decision problems, which are coarsenings of grand-world decision problems. If we treat acts, states and outcomes as propositions, this means that the acts, states and outcomes of the small world decision problems are disjunctions of the acts, states, and outcomes of the grand-world decision problem. The decision problem described in Table 1 is such a small-world decision problem.

Joyce (1999, p. 74) holds that an agent is rational in using such small-world decision problems to the extent that she is justified in believing that her solution to the small-world decision problem will be the same as her solution to the grand-world decision problem would be. This permits the use of small world decision problems both for the action-guiding and normative purposes of decision theory whenever the agent is justified in believing that they are good enough models of the grand-world decision problem.

Joyce argues that this condition is met in Jeffrey's decision theory *if* an agent correctly evaluates all coarse outcomes and actions, while it is not generally met in Savage's decision theory. As will be explained in Section 2.4, this is due to the feature of *partition invariance*, which Jeffrey's theory has and Savage's theory does not. Despite these arguments, if efficiency in decision-making is an important concern, as it is for an action-guiding theory, one might think that an agent should sometimes base her decision on a small-world decision problem even if she is fairly certain that her decision based on the grand-world decision problem will be different. She might think that her solution to a small-world decision problem will be close enough to that of the grand-world decision problem, while solving the small-world decision problem will save her costs of deliberation.

The second argument against having too fine-grained a decision problem is that this makes expected utility theory not restrictive enough. As will be explained in more detail in Section 2, the axioms used in the representation theorems of expected utility theory concern what combination of preferences are permissible. If preferences attach to outcomes, and outcomes can be individuated as finely as we like, then the danger is that

the norm to abide by the axioms of decision theory does not constrain our actions much.

For instance, consider the following preference cycle, where a, b and c are outcomes, and $\prec$ expresses strict preference:

$$a \prec b \prec c \prec a.$$

Preference cycles such as this are ruled out by the transitivity axiom, which all representation theorems we shall look at in Section 2 share. When outcomes can be individuated very finely, the following two problems may arise. Firstly, a number of authors have worried that any potential circularity in an agent's preferences can be removed by individuating outcomes more finely, such that there is no circularity anymore. Secondly, and relatedly, fine individuation may mean that no outcome can ever be repeated. In that case, an agent cannot reveal a preference cycle in her actions, and so we cannot interpret her as being irrational.

To see this, note that if we treat the first and the second occurrence of outcome a above as two different outcomes, say a_1 and a_2, the circularity is removed:

$$a_1 \prec b \prec c \prec a_2.$$

The worry is that this can always be done, for instance by distinguishing "option a if it is compared to b" from "option a if it is compared to c". If this strategy is always available, in what sense is the transitivity axiom a true restriction of the agent's preferences and actions? If we can't show that decision theory puts real restrictions on an agent's choices, then this is a problem especially for the action-guiding and normative projects.

A number of authors[9] have held that this problem shows that the axioms of decision theory on their own cannot serve as a theory of practical rationality (even a partial one), but have to be supplemented with a further principle in order to serve their function. Broome (1991, chapter 5) notes that the problem can be dealt with by introducing rational requirements of indifference. Rational requirements of indifference hold between outcomes that are modeled as different, but that it would be irrational for the agent to have a strict preference between. If there was a rational requirement of indifference between a_1 and a_2, for instance, the preference cycle would be preserved.

However, we may also restrict how finely outcomes can be individuated to solve the problem, by not allowing a distinction between a_1 and a_2. Broome (1991, chapter 5) advocates a rule of individuation by justifiers that serves the same role as the rational requirements of indifference. According to this rule, two outcomes can only be modeled as distinct if it is not irrational to have a strict preference between them.

9 See, especially, Broome (1991), Pettit (1991) and Dreier (1996).

Pettit (1991) proposes an alternative rule for individuation: two outcomes should be modeled as distinct just in case they differ in some quality the agent cares about, where caring about a quality cannot itself be cashed out in terms of preferences over outcomes. And Dreier (1996) argues that two outcomes should be distinguished just in case there are circumstances where an agent has an actual strict preference between them. Note that this rule for individuation is equivalent to the one proposed by Joyce, but Pettit's and Broome's rules may lead to coarser grained individuations of decision problems. The coarser grained the individuations, the more restrictive the axioms of expected utility theory end up being.

2 REPRESENTATION THEOREMS

2.1 *The Preference Relation*

In decision theory, representation theorems are proofs that an agent's preferences are representable by a function that is maximised by the agent. In the case of expected utility theory, they are proofs that an agent's preferences are such that we can represent her as maximising an expected utility function. As we will see in Section 3, many decision theorists believe that utility is nothing more than a convenient way to represent preferences. Representation theorems are crucial for this interpretation of utility. The significance of the representation theorems will be further discussed in Section 3.2.

A weak preference relation is a binary relation $\succcurlyeq$, which is usually interpreted either as an agent's disposition to choose, or her judgements of greater choiceworthiness.[10] An agent weakly prefers x to y if she finds x at least as choiceworthy as y, or if she is disposed to choose x when x and y are available.

We can also define an indifference relation $\sim$ and a strict preference relation $\succ$ in terms of the weak preference relation $\succcurlyeq$:

1. $x \sim y$ if and only if $x \succcurlyeq y$ and $y \succcurlyeq x$.
2. $x \succ y$ if and only if $x \succcurlyeq y$ and not $y \succcurlyeq x$.

Representation theorems take such preference relations as their starting point. They then proceed by formulating various axioms that pose restrictions on the preference relation, some of which are interpreted as

10 Many economists interpret preference as 'revealed preference', and claim that an agent counts as preferring x to y just in case she actually chose x when y was also available. Such pure behaviourism is usually rejected in the philosophical literature because it takes away from the explanatory power of preferences, and does not allow for counter-preferential choice. For a critique of the notion of revealed preference, see Hausman (2000).

conditions of rationality. Let X be the domain of the preference relation. What representation theorems prove is the following. If an agent's preferences conform to the axioms, there will be a probability function and a utility function such that:

$$\text{For all } x \text{ and } y \in X,\ EU(x) \geq EU(y) \text{ if and only if } x \succcurlyeq y.$$

All the representation theorems described in the following assume that the preference relation is a *weak ordering* of the elements in its domain. That means that the preference relation is transitive and complete:

TRANSITIVITY: For all x, y and $z \in X, x \succcurlyeq y$ and $y \succcurlyeq z$ implies that $x \succcurlyeq z$.

COMPLETENESS: For all x and $y \in X, x \succcurlyeq y$ or $y \succcurlyeq x$.

Section 4 will discuss potential problems with both completeness and transitivity.

Different representation theorems differ both in terms of the domain over which the preference relation is defined, and in terms of the other axioms needed for the representation theorem. They also differ in how many of the agent's attitudes other than preferences they take for granted. Consequently, they result in representation theorems of different strength.

2.2 *Von Neumann and Morgenstern*

One of the first representation theorems for expected utility is due to von Neumann and Morgenstern (1944) and takes probabilities for granted.[11] In this representation theorem, the objects of preference are *lotteries*, which are either probability distributions $L = (p_1, ..., p_m)$ over the m outcomes, or probability distributions over these 'simple' lotteries. Probabilities are thus already part of the agent's object of preference.

While it helps to think of lotteries in the ordinary sense of monetary gambles where there is a known probability of winning some prize, von Neumann and Morgenstern intended for their representation theorem to have wider application. In our original example, if there is a 50% chance that my boat is seaworthy, then I face a 50/50 lottery over the outcomes described in Table 1. Note furthermore that, since we are dealing directly with probability distributions over outcomes, there is no need to speak of states of the world.

While von Neumann and Morgenstern's representation theorem is perhaps most naturally understood given an objective interpretation of probability, their representation theorem is in fact compatible with any interpretation of probability. All we need is to already have access to the relevant

11 An earlier representation theorem is due to Ramsey (1926/2010) and derives probabilities as well as utilities. It is often considered as a precursor to Savage's and Bolker's representation theorems, discussed below. See R. Bradley (2004).

(precise) probabilities when applying the representation theorems. If we think of probability as the agent's subjective degrees of belief, we already need to know what those subjective degrees of belief are. If we think of it as objective chance, we need to already know what those objective chances are.

What von Neumann and Morgenstern go on to prove in their representation theorem is that, provided an agent's preferences over lotteries abide by certain axioms, there is a utility function over outcomes such that an agent prefers one lottery over another just in case its expected utility is higher. One crucial axiom needed for this representation theorem is the independence axiom, discussed in Section 5.1.

Note that the result is not that there is one unique utility function which represents the agent's preferences. In fact, there is a family of utility functions which describe the agent's preferences. According to von Neumann and Morgenstern's representation theorem, any utility function which forms part of an expected utility representation of an agent's preferences will only be unique up to positive, linear transformations. The different utility functions that represent an agent's preferences will thus not all share the same zero point. What outcome will yield twice as much utility will then also differ between different utility functions. It is therefore often claimed that these properties of utility functions represent nothing "real". What is invariant between all the different utility functions that represent the agent's preferences, however, are the ratios of utility differences, which can capture the curvature of the utility function. Such ratios are often used to measure an agent's level of risk aversion.[12]

2.3 *Savage*

While von Neumann and Morgenstern's representation theorem provides a representation of an agent's preferences where probabilities are already given, Savage (1954) infers both a utility function and probabilities from an agent's preferences.[13] As we have already seen, the standard tripartite distinction of actions, outcomes and states of the world goes back to Savage. Instead of assuming, like von Neumann and Morgenstern did, that we can assign probabilities to outcomes directly, we introduce a set

12 Risk aversion is further discussed in Section 5.3. Also see Mas-Colell, Whinston, and Green (1995), chapter 6 for more detail on expected utility theory's treatment of risk aversion.

13 This is why von Neumann and Morgenstern's theory is sometimes referred to as a theory of decision-making under risk, and Savage's is referred to as a theory of decision-making under uncertainty. In the former, probabilities are already known, in the latter, subjective probabilities can be assigned by the agent. However, note that, as we pointed out above, von Neumann and Morgenstern's theory can also be applied when probabilities are subjective.

of states of the world, which determine what outcome an act will lead to. The agent does not know which of the states of the world will come about.

Savage takes the agent's preferences over acts as input, and introduces a number of axioms on these preferences. He derives both a probability function over states, which abides by the standard axioms of probability, and a utility function over outcomes which, like the one von Neumann and Morgenstern derived, is unique up to positive linear transformations. Together, they describe an expected utility function such that an act is preferred to another just in case it has a higher expected utility. Importantly, the agents in Savage's decision theory abide by the sure-thing principle, which serves a role similar to the independence axiom in von Neumann and Morgenstern's representation theorem, and will also be discussed in Section 5.1.

Acts, states and outcomes are all treated as theoretical primitives in Savage's framework. But Savage's representation theorem relies on a number of controversial assumptions about the act, state and outcome spaces and their relation. For one, probabilities apply only to states of the world, and utilities apply only to outcomes. Preferences range over both acts and outcomes. Savage assumed that an act and a state together determine an outcome. Most controversially, Savage assumes that there are what he calls *constant acts* for each possible outcome, that is, acts which bring about that outcome in any state of the world. For instance, there must be an act which causes me great happiness even in the event that the apocalypse happens tomorrow. What makes things worse, by completeness, agents are required to have preferences over all these acts. Luce and Suppes (1965) take issue with Savage's theory for this reason.

While the results of Savage's representation theorem are strong, they rely on these strong assumptions about the structure of the act space. This is one reason why many decision theorists prefer Jeffrey's decision theory and Joyce's modification thereof.

2.4 *Jeffrey, Bolker and Joyce*

Jeffrey's decision theory, developed in Jeffrey (1965/1983), uses an axiomatisation by Bolker (1966). While he does not rely on an act space as rich as Savage's, Jeffrey preserves the tripartite distinction of acts, states and outcomes. However, for him, all of these are propositions, which means he can employ the tools of propositional logic. Moreover, preferences, utility and probability all range over all three. Agents end up assigning

probabilities to their own acts,[14] and assigning utilities to states of the world.

Jeffrey's theory is sometimes known as conditional expected utility theory, because agents who follow the axioms of his decision theory are represented as maximisers of a conditional expected utility. In Savage's decision theory, the utilities of outcomes are weighted by the unconditional probability of the states in which they occur. This is also the formulation we presented in Section 1.1. In the example there, we weighted the possible outcomes by the probability of the state they occur in. For instance, we weighted the outcome of enjoying a year on a boat without damages by the probability of my boat being seaworthy.

Jeffrey noted that the unconditional nature of Savage's decision theory may produce the wrong results in cases where states are made more or less likely by performing an action. In our example, suppose that, for whatever reason, my choosing to live on a boat for a year makes it more likely that my boat is seaworthy. The unconditional probability of the boat being seaworthy is lower than the probability of it being seaworthy given I decide to live on the boat. And thus using the unconditional probability may lead to the judgement that I shouldn't spend the year on the boat, because the probability of it not being seaworthy is too high—even if the boat will be very likely to be seaworthy if I choose to do so. To avoid this problem, Jeffrey argued, it is better to use probabilities that are in some sense conditional on the action whose expected utility we are evaluating. We should weight the outcome of spending a year on a boat without damage by the probability of the boat being seaworthy given that I choose to live on the boat for a year.[15]

Let the probability of a state given an act be $p_A(S)$. There is much disagreement on how this probability is to be interpreted. The main disagreement is whether it should be given a causal or an evidential interpretation. I postpone this discussion to Section 3.3. But let me note here that Jeffrey himself falls on the evidential side. Conditional expected utility theory advises us to maximise the following:

$$EU(A_i) = \sum_{j=1}^{m} p_{A_i}(S_j) \cdot u(O_{ij})$$

Jeffrey interprets this conditional expected utility as an act's 'news value', that is, as measuring how much an agent would appreciate the news that the act is performed.

14 This is a controversial feature of the theory. See Spohn (1977) for criticism of this assumption.

15 Savage's own solution to the problem is that, for his formalism to apply, states and acts need to be specified such that there is no dependence between an action being performed and the likelihood of a state. Jeffrey's response is more elegant in that it requires no such restriction on what kinds of decision problems it can be applied to.

The conditional nature of Jeffrey's decision theory is also what leads to its partition invariance.[16] In Jeffrey's theory, the value of a disjunction is always a function of the value of its disjuncts. For instance, the value of a coarse outcome O_{1-10} which is a disjunction of outcomes $O_1, \ldots, O_{10}$ is a function of the values of the outcomes $O_1, \ldots, O_{10}$. But we could also subdivide the coarse outcome O_{1-10} differently. O_{1-10} is also a disjunction of the coarse outcomes O_{1-5} and O_{6-10}, which are themselves disjunctions of $O_1, \ldots, O_5$ and $O_6, \ldots, O_{10}$ respectively. And so we can also calculate the value of O_{1-10} from the values of O_{1-5} and O_{6-10}. Partition invariance means that we get the same value in either case. The value of O_{1-10} can be represented as a function of the values of any of its subdivisions. This means that, as long as utilities are assigned correctly to disjunctions, Jeffrey's decision theory gives equivalent recommendations no matter how finely we individuate outcomes, states and actions. Joyce argues that for this reason, the use of small-world decision problems is legitimate in Jeffrey's decision theory (see Section 1.4), and that that is a major advantage over Savage's unconditional, and partition variant decision theory.

Jeffrey's and Bolker's representation theorem is less strong than Savage's. It does not pin down a unique probability function. Nor does it result in a utility function that is unique up to positive linear transformations. Instead, it only ensures that probability and utility pairs are unique up to fractional linear transformations.[17]

Joyce (1999) argues that this shows that we need to augment Jeffrey's and Bolker's representation theorem with assumptions about belief, and not merely preference. Unlike von Neumann and Morgenstern, however, he does not propose to simply assume probabilities. Instead, he introduces a 'more likely than' relation, on which we can formulate a number of axioms, just as we did for the preference relation. The resulting representation theorem results in a unique probability function and a utility function which is unique up to positive linear transformations.[18]

We have introduced the most prominent representation theorems for expected utility theory.[19] What do these representation theorems show? Each of them shows that if an agent's preferences abide by certain axioms, and certain structural conditions are met, her preferences can be represented by a utility (and probability) function (or families thereof) such that she prefers an act to another just in case its expected utility is higher.

16 See Joyce (1999), pp. 121-122.

17 A fractional linear transformation transforms u to $\frac{a \cdot u + b}{c \cdot u + d}$, with $a \cdot d - b \cdot c > 0$.

18 Also see R. Bradley (1998), for an alternative way to secure uniqueness.

19 A helpful, more technical and more detailed overview of representation theorems can be found in Fishburn (1981).

Agents who abide by the axioms can thus be represented as expected utility maximisers.

What these kinds of results show depends to some extent on the purpose we want to put our theory to. But it also depends on how we interpret the utilities and probabilities expected utility theory deals with. Section 3 gives an overview of these interpretations and then returns to the question of what the representation theorems can show.

3 INTERPRETATIONS OF EXPECTED UTILITY THEORY

3.1 *Interpretations of Utility*

Some of the earliest discussions of choice under uncertainty took place in the context of gambling. The idea that gamblers maximise some expected value first came up in correspondence between Fermat and Pascal (1654/1929). Pascal, who formulated the expected value function in this context, thought of the value whose expectation should be maximised as money. This is natural enough in the context of gambling. Similarly, in this context it is natural to think of the probabilities involved as objective, and fixed by the parameters of the game.

However, money was soon replaced by the notion of utility as the value whose expectation is to be maximised. This happened for two interrelated reasons. First, the same amount of money may be worth more or less to us depending on our circumstances. In particular, we seem to get less satisfaction from some fixed amount of money the more money we already have. Secondly, the norm to maximise expected monetary value has some counterintuitive consequences. In particular, we can imagine gambles that have infinite monetary value, that we would nevertheless only pay a finite price for. Nicolas Bernoulli first demonstrated this with his famous St. Petersburg Paradox.[20]

In response to these problems, Daniel Bernouilli (1738/1954) and Gabriel Cramer independently proposed a norm to maximise expected utility rather than expected monetary value. However, this raises the problem of how to interpret the notion of utility. One strand of interpretations takes utility to be a real psychological quantity that we could measure. Let us call such interpretations of utility 'realist'. Early utilitarians adopted a realist interpretation of utility. For instance, Bentham (1789/2007) and Mill (1861/1998) thought of it as pleasure and the absence of pain.

20 Bernoulli proposed a gamble in which a coin is thrown repeatedly. If it lands heads the first time, the player gets \$2. If it lands tails, the prize is doubled, and the coin thrown again. This procedure is repeated indefinitely. The expected value of the resulting gamble is thus $\$2 \cdot \frac{1}{2} + \$4 \cdot \frac{1}{4} + \$8 \cdot \frac{1}{8} + ...$, which is infinite. However, most people would only pay a (low) finite amount for it.

Note, however, that these utilitarians were interested in defining utility for the purpose of an ethical theory rather than a theory of rationality. One problem with interpreting utility as pleasure in the context of expected utility theory is that the theory then seems to imply that true altruism can never be rational. If rationality requires me to maximise my own expected pleasure, then I can never rationally act so as to increase somebody else's happiness at my own expense.

For this and other reasons modern realists typically think of utility as a measure of the strength of an agent's desire or preference, or her level of satisfaction of these desires or preferences. I may strongly desire somebody else's happiness, or be satisfied if they achieve it, even if that does not directly make me happy.[21] Jeffrey (1965/1983), for instance, speaks of desirabilities instead of utilities, and interprets them as degrees of desire (p. 63). The corresponding realist interpretation of the probabilities in expected utility theories is usually that of subjective degrees of belief.

The representation theorems described in Section 2 have, however, made a different kind of interpretation of utility (and probability) possible, and popular. These representation theorems show that preferences, if they conform to certain axioms, can be represented with a probability and utility function, or families thereof. And so, encouraged by these results, many decision theorists think of utility and probability functions as mere theoretical constructs that provide a convenient way to represent binary preferences. For instance, Savage (1954) presents his theory in this way. Importantly, on this interpretation, we cannot even speak of probabilities and utilities in the case where an agent's preferences do not conform with the axioms of expected utility theory. Let us call these interpretations of utility and probability 'constructivist'.[22]

3.2 *The Significance of the Representation Theorems*

Whether we adopt a realist or a constructivist interpretation of utility matters for how expected utility theory can serve the three purposes of decision theory described in Section 1.2, and for what the representation theorems presented in Section 2 really establish. Let us first look at the interpretive project. As already mentioned, those interested in the interpretive project have mostly been interested in inferring an agent's beliefs and

21 This is also the interpretation adopted by several later utilitarians, such as Hare (1981) and Singer (1993).

22 See Dreier (1996) and Velleman (1993/2000) for defenses of constructivism. Buchak (2013) draws slightly different distinctions. For her, any view on which utility is at least partially defined with respect to preferences counts as constructivist. Since this is compatible with holding that utility is a psychologically real quantity, she allows for constructivist realist positions. The position that utility expresses strength of desire, for her, is such a position. I will count this position as realist, and not constructivist.

desires from her choice behaviour. If that is the goal, then the probabilities and utilities involved in decision theory should at least be closely related to desires and beliefs. Under the assumption that agents maximise their utility and probability functions, thus understood, we can hypothesise, perhaps even derive, probability and utility functions that motivate an agent's actions.

How could the representation theorems we described in Section 2 help with this project? They go some way towards showing that beliefs and desires can be inferred from an agent's choice behaviour. But the following assumptions are also needed for this project to succeed:

1. The agent's choice behaviour must reflect her preferences, at least most of the time. This assumption is more likely to be met if we think of preferences as a dispositions to choose, rather than as judgements of choiceworthiness.

2. The axioms of the representation theorems must be followed by the agent, at least most of the time. If we want to use expected utility theory to deduce an agent's beliefs and desires, then the agent's preferences have to be representable by an expected utility function. While we can interpret the axioms as rationality constraints, these cannot be the kinds of constraints that people fail to meet most of the time. In particular, if we want to employ expected utility theory for Davidson's 'radical interpretation', then the choice behaviour of agents who fail to abide by the axioms will turn out to be unintelligible.

3. The probabilities and utilities furnished by the representation theorem must correspond to the agent's actual beliefs and desires.

Assumption 2 is controversial for the reasons described in Sections 4 and 5. But assumption 3 is also problematic. The representation theorems only show that an agent who abides by the axioms of the various representation theorems can be represented as an expected utility maximiser. But this is compatible with the claim that the agent can be represented in some other way. It is not clear why the expected utility representation should be the one which furnishes the agent's beliefs and desires.[23]

To answer this challenge, the best strategy seems to be to provide further arguments in favour of expected utility maximisation, and in favour of probabilistic beliefs, apart from the plausibility of the axioms

23 This question was raised, for instance, by Zynda (2000), Hajek (2008) and Meacham and Weisberg (2011). Zynda (2000) argues that the representation theorems alone cannot show that agents do or should have probabilistic degrees of belief. Meacham and Weisberg (2011) provide a number of arguments why the representation theorems alone cannot serve as the basis of decision theory.

of the representation theorems. Suppose we think it is plausible that agents should have probabilistic degrees of belief, and should maximise the expected degree of satisfaction of their desires. And suppose we also think that our preferences are closely related to our desires. Then if, given some plausible axioms, these preferences can be given an expected utility representation, we seem to have good reason to think that the utilities and probabilities furnished by the representation theorem correspond to our degrees of belief and strength of desire.

Setting aside the question of why we might want to have probabilistic degrees of belief, what could such realist arguments for expected utility maximisation be? Note that, for the purposes of the interpretive project, these arguments have to not only be normatively compelling, but also convince us that ordinary agents would be expected utility maximisers. One type of argument appeals to the advantages of being an expected utility maximiser when making decisions in a dynamic context. These will be covered in Section 7. Pettigrew (2014) makes another argument: for most realists, utility is supposed to capture everything an agent cares about. If that is true, then it seems plausible to say that in uncertain situations, I should be guided by my best estimate of how much utility I will get. We can appeal to results in de Finetti (1974) to argue that an agent's best estimate of a quantity is her subjective expectation. This is so because any estimate of the quantity that is a weighted sum different from the expectation will be accuracy dominated by an expectational estimate: the expectational estimate will be closer to the true value no matter what happens. Thus, I should maximise my expected utility.

So far, we have assumed a realist interpretation of utility and probability. Note, however, that expected utility theory could still be explanatorily useful even if a constructivist interpretation of utility and probability are adopted. It is often argued that the representation theorems show that the utility and probability functions allow for a simpler and more unified representation of an agent's preferences: all the agent's preferences can be described with one utility and probability function. This could be seen to make them more intelligible. In fact, Velleman (1993/2000) argues that being an expected utility maximiser makes an agent more intelligible to herself and others, and that this gives her a reason to be an expected utility maximiser.

Let us now turn to the action-guiding and normative projects. These projects will lead to quite different prescriptions depending on whether utility is interpreted in a realist or in a constructivist sense. Suppose that we are constructivists about utility. In that case, there is a sense in which the prescription to maximise expected utility does not make any sense. If one abides by the axioms of one's favourite representation theorem, one's preferences are representable as expected utility maximising. To

maximise expected utility, there is nothing more one needs to do, apart from act according to the preferences over acts one already has. But if one's preferences do not abide by the axioms, on the other hand, one simply does not have a utility function whose expectation one could maximise.

Consequently, constructivists often interpret the prescription of expected utility theory as a prescription to have preferences such that one can be represented as an expected utility maximiser. That is, one should abide by the axioms of expected utility theory. For the action-guiding project, this means that, as an agent, I should have preferences such that they abide by the axioms of expected utility theory. For the normative project, it means that we judge an agent to be irrational if she has preferences that violate the axioms. This is why constructivists often interpret expected utility theory as a theory about what it means to have coherent preferences or ends, rather than as a theory of means-ends rationality.

For realists, however, the prescription to maximise expected utility makes sense even independently of the representation theorems canvassed in Section 2. Consider first the action-guiding project, which aims to interpret expected utility theory as a theory that can guide an agent in deciding what to do. If utility is just my strength of desire, and probability is my degree of belief, and I have introspective access to these, then I can determine the expected utility of the various acts open to me. I can do so without considering the structure of my preferences, and whether they abide by the axioms of expected utility theory. Expected utility theory is then action-guiding without appeal to representation theorems. But note that the advice to maximise expected utility is only useful to agents if they really have such intuitive access to their own degrees of belief and strength of desire.[24]

Similarly, if we are realists and our interests are normative, we can judge an agent to be irrational by considering her utilities and degrees of belief, and determining whether she failed to maximise expected utility. This is because there will be facts about the agent's utilities and probabilities even if she fails to maximise expected utility. Realists about utility and probability can also help themselves to the realist arguments for expected utility maximisation just mentioned. For them, the normative force of expected utility theory does not depend solely on the plausibility of the axioms of expected utility theory. If we adopt a realist interpretation of utility and probability, it is also easier to argue that expected utility theory provides us with a theory of instrumental rationality. Maximising expected utility could be seen as taking the means towards the end of achieving maximum utility. However, realists will also have to provide an argument that this is a goal rational agents ought to have.

24 Also see Bermudez (2009) on this claim.

3.3 *Causal and Evidential Interpretations of Expected Utility Theory*

We have said that the probabilities involved in expected utility theory are usually interpreted as subjective degrees of belief, at least by realists. As we have seen, Jeffrey, Joyce, and others have advocated a conditional expected utility theory. In conditional expected utility theory, agents determine an act's expected utility by weighting utilities by the different states' probabilities conditional on the act in question being performed. Above, we called this probability $p_A(S)$. How this probability is to be interpreted is a further important interpretive question. The main disagreement is about whether it should be given a causal or an evidential interpretation. Jeffrey himself had worked with an evidential interpretation, while causal decision theorists, such as Gibbard and Harper (1978/1981), Armendt (1986), or Joyce (1999)[25] have given it a causal interpretation.

The difference between these two interpretations is brought out by the famous Newcomb Problem, first introduced by Nozick (1969). In this problem, we imagine a being who is very reliable at predicting your decisions, and who has already predicted your choice in the following choice scenario. You are being offered two boxes. One is opaque and either has no money in it, or $1,000,000. The other box is clear, and you can see that it contains $1,000. You can choose to either take only the opaque box, or to take both boxes. Under normal circumstances, it would seem clear that you should take both boxes. Taking the clear box gives you $1,000 more no matter what.

The complication, however, is that the being's prediction about your action determines whether there is money in the opaque box or not. If the being predicted that you will take two boxes, then there is no money in the opaque box. If the prediction was that you will take only the opaque box, there will be money in it. Since the being's prediction is reliable, those who take only one box tend to end up with more money than those who take two boxes.

Note that while this case is unrealistic, there are arguably real-life cases that resemble the Newcomb Problem in its crucial features. In these cases, the acts available to an agent are correlated with good or bad outcomes even though these are not causally promoted by the act. This happens in medical cases, for instance, if a behavioural symptom is correlated with a disease due to a common cause. Before the causal link between smoking and lung cancer was firmly established, interested parties hypothesised that there may be a common cause which causes both lung cancer, and the disposition to smoke. If that were right, smoking would not cause lung

25 Joyce also first showed that the two interpretations can be given a unified treatment in a more general conditional expected utility theory.

cancer, but merely give you evidence that you are more likely to develop it.[26]

Evidential and causal decision theory come apart in their treatment of these cases. Evidential decision theory traditionally interprets $p_A(S)$ as a standard conditional probability:

$$p_A(S) = \frac{p(A\&S)}{p(A)}.$$

According to this interpretation, the probability of the state where there is \$1,000,000 in the opaque box conditional on taking only one box is much higher than the probability of the state where there is \$1,000,000 in the opaque box conditional on taking two boxes. This is because the act of taking only one box provides us with evidence that the prediction was that you would take only one box, in which case there is money in the opaque box. And so expected utility maximisation would tell you to take only one box.

Causal decision theorists take issue with this, because at the time of decision, the agent's actions have no more influence on whether there is money in the opaque box or not. Either there is or there isn't already money in the box. In either case, it is better for you to take two boxes, as Table 4 illustrates. This kind of dominance reasoning speaks in favour of taking both boxes.

	Prediction: one box	Prediction: two boxes
Take one box	\$1,000,000	\$0
Take two boxes	\$1,001,000	\$1,000

Table 4: The Newcomb Problem

Causal decision theory allows for this by giving $p_A(S)$ a causal interpretation. It measures the causal contribution of act A to whether state S obtains. Following a proposal by Stalnaker (1972/1981), Gibbard and Harper (1978/1981) use the probability of a conditional in their causal decision theory, instead of a conditional probability. In particular, they use the probability of the conditional that an outcome would occur if an action was performed.[27]

In the Newcomb Problem, neither the act of taking nor the act of not taking the clear box make any causal contribution to whether there is money in the opaque box. And so, on the causal interpretation, $p_A(S)$

26 See Price (1991) for more examples.

27 Lewis (1981) shows that if the right partition of acts, states and outcomes is used, Savage's decision theory will give the same recommendations as Gibbard and Harper's, and is thus a type of causal decision theory.

just equals the unconditional probability $p(S)$ in both cases. And then dominance reasoning becomes relevant.

Note, however, that it is controversial whether taking both boxes really is the rational course of action in the Newcomb Problem. Those who advocate 'one-boxing', such as Horgan (1981/1985) and Horwich (1987), point out that one-boxers end up faring better than two-boxers. It is also controversial whether evidential decision theory really does yield the recommendation to one-box if the problem is represented in the right way: Eells (1981) argues that evidential decision theory, too, recommends two-boxing.

Jeffrey (1965/1983) himself supplements evidential decision theory with a ratifiability condition, which allows him to advocate two-boxing. The condition claims that an agent should maximise expected utility relative to the probability function she will have once she finally decides to perform the action. In the Newcomb Problem, only two-boxing is ratifiable. If the agent decided to one-box, she would then be fairly certain that there is money in the opaque box, and then she will wish she had also taken the second box. If she decides to two-box, she will be fairly certain that there is no money in the opaque box, and she will be glad that she at least got the $1,000.[28]

4 INCOMPLETENESS AND IMPRECISION

Several important challenges to expected utility theory have to do with the fact that expected utility theory asks us to have attitudes that are more extensive and precise than the preferences ordinary decision makers have. In fact, in many cases it does not seem irrational to have attitudes that are in some way imprecise or incomplete. And so the problems discussed in the following arise both for the interpretive as well as for the action-guiding and normative uses of decision theory.

The challenge takes different forms for constructivists and realists. For constructivists, imprecision and incompleteness will manifest as violations of the axioms of the representation theorems presented in Section 2. As we have seen, all of these representation theorems assume that the agent's preference relation forms a weak ordering of the elements in its domain. This means that the preference relation must be transitive and complete.

28 The status of the ratifiability condition is still a part of the contemporary debate on causal decision theory. One open question is what decision should be favoured in cases of decision instability, where no action is ratifiable, like in Gibbard and Harper's Death in Damascus case (see Gibbard and Harper (1978/1981), and Egan (2007) for further, similar cases). Arntzenius (2008) and Joyce (2012) argue for ways of dealing with this problem. The ratifiability condition also helps to illuminate certain equilibrium concepts in game theory (see Joyce and Gibbard (1998)).

Both assumptions are controversial for related reasons. Completeness is controversial because it asks agents to have a more extensive set of preferences than they actually have. Transitivity is controversial in cases where an agent's desires are coarse-grained, as will be explained below. For realists, a related challenge is that both our degrees of belief and our strength of desire are not precise enough to allow for representation in terms of a precise probability and utility function.

4.1 *Incompleteness*

To start with the completeness condition, the worry here is that agents simply do not have preferences over all the elements of the set the decision theory asks them to have preferences over. For instance, if I have lived in Germany all my life, I might simply have no preference between living in Nebraska and living in in Wyoming. It's not that I have never heard of these places. The question would just never occur to me. It might then neither be the case that I prefer Nebraska to Wyoming nor that I prefer Wyoming to Nebraska. I am also not indifferent between the two. I might simply have no preference. But if these outcomes are part of the set of outcomes the decision theory asks me to have preferences over, then this means that I am violating the completeness condition.

Similar claims are often made about cases of incommensurable values. In a famous example due to Sartre (1945/2007), a young man has to choose between caring for his sick mother and joining the French Resistance. The two options here are often said to involve incommensurable values: on the one hand, responsibility to one's family, and on the other hand, fighting for a just cause. In these kinds of cases, too, we might want to say that the young man is neither indifferent, nor does he prefer one option to the other. And here, this is not because the question of what he prefers has never occurred to the man. He may in fact think long and hard about the choice. Rather, he has no preference because the values involved are incommensurable.

These kinds of examples are more convincing if our notion of preference is that of a judgement of choiceworthiness. In these examples, agents have not made, or are unable to make judgements of choiceworthiness about some of the elements of the relevant set. If one thinks of preference as disposition to choose instead, one might think that even if an agent never thought about a particular comparison of outcomes, there can still be a fact of the matter what she would be disposed to choose if she faced the

choice. Moreover, if this is our notion of preference, we simply draw no distinction between indifference and incommensurability.[29]

However, this alternative notion of preference may get into trouble when some of the acts in the relevant set are ones that the agent could not possibly choose between. The completeness condition in standard expected utility theory may require the agent to have what Broome (1991) calls 'impractical preferences'. For instance, it might require an agent to have a preference between

O_1 : an orange, and

O_2 : an apple when the alternative is a banana

Choosing between these alternatives is impossible in the sense that O_2 will not come about unless the alternative is a banana, not an orange. And so it seems like we cannot determine the agent's choice disposition between them.

Incompleteness in preference is often dealt with by replacing the completeness axiom in the various representation theorems with a condition of *coherent extendibility*.[30] That is, we only require that an agent's preferences are such that we could extend her set of preferences in a way that is consistent with the other axioms of the representation theorem. The problem with this strategy is that any representation in terms of probability or utility that the representation theorem furnishes us with will only be a representation relative to an extension. There will usually be several extensions that are consistent with an agent's incomplete preferences and the axioms of the theorem. And thus, there will be several possible representations of the agent's preferences. The representation theorem will no longer furnish us with a unique probability function, and a utility function that is unique up to positive linear transformations. For this reason, incompleteness of preference is often associated with imprecise probabilities.

4.2 *Imprecise Probabilities*

There is an active field of research investigating imprecise probabilities.[31] These imprecise probabilities are usually represented by families of probability functions. And families of probability functions is exactly what the representation theorems furnish us with if the completeness condition is

29 In fact, Joyce (1999) considers this an important argument against more behaviourist interpretations of preference.

30 This is the strategy taken by Kaplan (1983), Jeffrey (1965/1983), and Joyce (1999).

31 See S. Bradley (2015) and Mahtani (2019) for helpful overviews of the literature. For an introduction to the theory of imprecise probabilities, see Augustin, Coolen, de Cooman, and Troffaes (2014).

replaced by a coherent extendibility condition. While this gives even a constructivist reason to engage with imprecise probabilities, there are also various realist arguments for doing so. Many formal epistemologists agree that sharp degrees of belief that can be expressed with a sharp probability function are both psychologically unrealistic, and cannot be justified in situations where there is insufficient evidence.[32] If we believe that the probabilities in decision theory should accurately describe our belief states, the probabilities in decision theory should then be imprecise.

Another motivation for engaging with imprecise probabilities is that this allows us to treat states or outcomes to which the agent can assign precise probabilities differently from states or outcomes to which the agent cannot assign precise probabilities. This may allow us to make sense of the phenomenon of *ambiguity aversion*. Ambiguity aversion occurs in situations where the probabilities of some states are known, but the agent has no basis for assigning probabilities to some other states. In such situations, many agents are biased in favour of lotteries where the probabilities are known. For instance, take the following example from Camerer and Weber (1992):[33]

> Suppose you must choose between bets on two coins. After flipping the first coin thousands of times you conclude it is fair. You throw the second coin twice; the result is one head and one tail. Many people believe both coins are probably fair ($p(\text{head}) = p(\text{tail}) = .5$) but prefer to bet on the first coin, because they are more confident or certain that the first coin is fair. (p. 326)

Standard expected utility theory cannot make sense of this, since it does not allow us to distinguish between different degrees of uncertainty. In standard expected utility theory, every state is assigned a precise probability. As a result, ambiguity aversion can lead an agent to violate the axioms of the different representation theorems. In particular, ambiguity aversion can result in violations of separability (see Section 5) as in the famous Ellsberg Paradox.[34] Nevertheless, ambiguity aversion is common and does

32 For examples of these claims, see, for instance, Levi (1980) and Kaplan (1996). When an agent cannot assign a sharp probability to states, we sometimes speak of decision-making under indeterminacy or ignorance, as opposed to merely uncertainty.

33 Camerer and Weber (1992) also provide an overview of the empirical evidence of this phenomenon.

34 See Ellsberg (1961).The Ellsberg Paradox runs as follows: you are given an urn that you know contains 90 balls. 30 of them are red. The remaining 60 are either black or yellow, but you don't know what the distribution is. Now first, you are offered the choice between receiving $100 if a red ball is drawn, and receiving $100 if a black ball is drawn. Most people choose the former. Then, you are offered the choice between receiving $100 if a red or yellow ball is drawn, and receiving $100 if a black or yellow ball is drawn. Here,

not seem irrational. Imprecise probabilities may help us to better model ambiguity, and thus hold the promise to help us rationalise ambiguity averse preferences.

There are epistemological objections to using sets of probabilities to represent beliefs.[35] But another common objection to using imprecise probabilities is that they lead to bad decision-making.[36] How could decision-making with imprecise probabilities proceed? We can use each probability function in the family in order to calculate an expected utility for each act open to the agent. But then each act will be associated with a family of expected utilities, one for each member of the family of probability functions. And so the agent cannot simply maximise expected utility anymore. The question then becomes how we should make decisions with these sets of probabilities and expected utilities.

One type of simple proposal that appears in the literature is the following principle, sometimes called *Liberal*: an act which maximises expected utility for every probability function in the family is obligatory. And any act which maximises expected utility for some probability function in the family is permitted.[37] For an overview of other choice rules, see Troffaes (2007).

Elga (2010) raises an important challenge for all such choice rules. If they are permissive, as *Liberal* is, then they will allow us to make choices in a series of bets that leave us definitely worse off. But if they are not permissive, and always recommend a single action, they undercut one main motivation for using imprecise probabilities in the first place. In that case, they will pin down precise betting odds for an agent. But, Elga argues, if we think that the evidence does not license us to use a precise probability, it would be strange if it determined precise betting odds. Moreover, these betting odds, if they abide by the axioms of expected utility theory, could be used to infer a precise probability using the representation theorems discussed above.[38]

Elga's argument bears resemblance to other dynamic arguments against violations of standard expected utility theory, which will be discussed in

---

most people choose the latter. These preferences display ambiguity aversion. They are not consistent with a stable assignment of precise subjective probabilities to the drawing of a yellow or black ball, combined with the assumption of expected utility maximisation.

35 See, for instance, the problem of *dilation*. Dilation occurs when an agent's beliefs become less precise when she updates on a piece of evidence. The phenomenon was first introduced by Seidenfeld and Wasserman (1993) and is argued to be problematic for imprecise probability theory in White (2010). See Joyce (2011), S. Bradley and Steele (2014b) and Pedersen and Wheeler (2014) for critical discussion.

36 See, for instance, Williamson (2010).

37 See White (2010), Williams (2014), Moss (2015).

38 However, note that there are choice rules that determine precise betting odds that do not reduce to expected utility maximisation, such as the one introduced by Sahlin and Weirich (2014).

Section 7. It may be challenged on similar grounds. There may be dynamic choice strategies available to agents that guard them against making sure losses in dynamic choice problems. In fact, Williams (2014) claims that agents using his choice rule can make their choices 'dynamically permissible' by only considering some of the probability functions in the family to be 'live' at any one point. S. Bradley and Steele (2014a), too, argue that agents with imprecise credences can make reasonable choices in dynamic settings.

4.3 *Imprecise Utility and Intransitivity*

One might expect there to be a literature on imprecision with regard to utilities similar to the one on imprecise probabilities. For one, replacing the completeness condition with a condition of coherent extendibility will not only lead to a family of probability representations, it will also result in a corresponding family of utility representations. Moreover, there might be similar realist arguments that could be made in favour of imprecise strength of desire or degree of preference. Some of the examples of incompleteness, such as the cases involving incommensurable values, could be described as examples where it is unclear to what degree an agent desires the goods in question, or how they compare. Such cases are also often described as cases of 'vague preference'. However, imprecise utilities and vague preferences are so far mostly discussed in the mathematical and economic literature. Fishburn (1998) suggests a probabilistic approach to studying vague preferences, while most of the literature uses fuzzy set theory. Salles (1998) provides an introduction to that approach.

There is a certain kind of lack of precision in our attitudes that does not result in vague preferences or incompleteness of preference. Instead, this lack of precision leads to a failure of transitivity, and is thus nevertheless problematic for expected utility theory. Intransitivity arises for outcomes that the agent finds indistinguishable with regard to some of the things she values. The problem is brought out most clearly by the Self-Torturer Problem, introduced by Quinn (1990). It runs as follows: a person has an electric device attached to her body that emits electric current which causes her pain. The device has a large number of settings, such that the person is unable to tell the difference in pain between any two adjacent settings. However, she can tell the difference between settings that are sufficiently far apart. In fact, at the highest settings, the person is in excruciating pain, while at the lowest setting, she is painless. Each week, the person can turn the dial of the device up by one setting, in exchange for $10,000.

Let us call the settings of the dial $D_0, D_1, D_2, ..., D_{1000}$. In this problem, the following set of intransitive preferences seems to be reasonable for a person who prefers less pain to more pain, and more money to less:

$$D_0 \prec D_1 \prec D_2 \prec ... \prec D_{1000} \prec D_0.$$

At the highest settings, the person is in such excruciating pain that she would prefer being at the lowest setting again to having her fortune. At the same time, if turning the dial up by one setting results in a level of pain that is indistinguishable from the previous, it seems that taking the $10,000 is always worth it, no matter how much pain the agent is already in.

An agent who has the self-torturer's preferences is clearly in trouble. In the original example, she can never turn the dial down again once she has turned it up. If she always follows her pairwise preferences, she will end up at the highest setting. This is obviously bad for her, by her own lights: there are many settings she would prefer to the one she ends up at. If, on the other hand, we suppose that the agent can go back to the first setting in the end, the problem is that she could be 'money-pumped'.[39] If the agent has a strict preference for the lowest setting over the highest setting, she should be willing to pay some positive amount of money on top of giving up all her gained wealth for going back to the first setting. She will end up having paid money for ending up where she started.

Advocates of standard expected utility theory may point out that these observations just show why it is bad to have intransitive preferences. However, critics, such as Andreou (2006) and Tenenbaum and Raffman (2012), point out that while these are problematic consequences of having the self-torturer's preferences, there seems to be nothing wrong with the self-torturer's preferences per se. If the agent's relevant underlying desires are those for money and the absence of pain, but the agent cannot distinguish between the levels of pain of two adjacent settings, then there is nothing in the agent's desires concerning the individual outcomes that could speak against going up by one setting. If we think that preferences should accurately reflect our underlying desires concerning the outcomes, the self-torturer's preferences seem reasonable.

Indeed, proponents of expected utility theory acknowledge that it is somewhat unsatisfactory to simply declare the self-torturer's preferences irrational. They have hence felt pressed to give an explanation of why the self-torturer's preferences are unreasonable, despite appearances. Arntzenius and McCarthy (1997), and Voorhoeve and Binmore (2006) have made different arguments to show that rational agents would hold that there

39 Money pumps were first introduced as an argument for transitivity by Davidson, McKinsey, and Suppes (1955).

is an expected difference in pain between two adjacent settings at least somewhere in the chain.

Critics note that it is only in the context of the series of choices she is being offered that the self-torturer's preferences become problematic. And so instead of declaring the self-torturer's preferences irrational, we may instead want to say that in some cases, it is rational for the agent to act against her punctate preferences. Andreou (2006) argues that the intransitive preferences of the self-torturer ought to be revised to be transitive for the purpose of choice only. Tenenbaum and Raffman (2012) note that the underlying problem in the self-torturer's case is that the agent's end of avoiding pain is *vague*. It is not precise enough to distinguish between all the different outcomes the decision theory may ask her to evaluate, and that she in fact may have to choose between. They claim that vague goals that are realised over time may ground permissions for agents to act against their punctate preferences. And so this is another type of imprecision in our attitudes which may call for a revision of standard expected utility theory.

5 SEPARABILITY

5.1 *The Separability Assumption*

The imprecision and incompleteness of our attitudes discussed in Section 4 may be a problem for expected utility theory even in the context of certainty. But another important type of criticism of expected utility theory has to do with the assumptions it makes about choice under uncertainty specifically. All the representation theorems canvassed in Section 2 make use of a similar kind of axiom about choice under uncertainty. These axioms are versions of what Broome (1991) calls *separability*. The idea here is that what an agent expects to happen in one state of the world should not affect how much she values what happens in another, incompatible state of the world. There is a kind of independence in value of outcomes that occur in incompatible states of the world. Separability is largely responsible for the possibility of an expected utility representation. Separability is a controversial assumption, for the reasons explained in Sections 5.2 and 5.3. Here, I present the versions of the separability assumption used in the representation theorems introduced in Section 2.

In von Neumann and Morgenstern's representation theorem (see Section 2.2), separability is expressed by the independence axiom. Let $\mathscr{L}$ be the space of lotteries over all possible outcomes. Then independence requires the following:

INDEPENDENCE: For all $L_x, L_y, L_z \in \mathscr{L}$ and all $p \in (0,1)$, $L_x \succcurlyeq L_y$ if and only if $p \cdot L_x + (1-p) \cdot L_z \succcurlyeq p \cdot L_y + (1-p) \cdot L_z$.

Independence claims that my preference between two lotteries will not be changed when those lotteries become sub-lotteries in a lottery which mixes each with some probability of a third lottery. For instance, suppose I know I get to play a game tonight. I prefer to play a game that gives me a 10% chance of winning a pitcher of beer to a game that gives me a 20% chance of winning a pint of beer. The independence axiom says that this preference will not be affected when the chances of me getting to play at all today change. The possibility of not playing at all tonight should not affect how I evaluate my options in the case that I do get to play.

In Savage's framework (see Section 2.3), separability is expressed by his famous sure-thing principle. To state it, we need to define a set of events, which are disjunctions of states. Let $A_i(E)$ be the act A_i when event E occurs. The sure-thing principle then requires the following:

SURE-THING PRINCIPLE: For any two actions A_i and A_j, and any mutually exclusive and exhaustive events E and F, if $A_i(E) \succcurlyeq A_j(E)$ and $A_i(F) \succcurlyeq A_j(F)$, then $A_i \succcurlyeq A_j$

The idea behind the sure-thing principle is that an agent can determine her overall preferences between acts through event-wise comparisons. She can partition the set of states into events, and compare the outcomes of each of her acts for each event separately. If an act is preferred given each of the events, it will be preferred overall. That is, if a particular act is preferred no matter which event occurs, then it is also preferred when the agent does not know which event occurs.

In Jeffrey's decision theory (see Section 2.4), separability is expressed by the averaging axiom. Remember that for him, acts, states and outcomes are all propositions, and all objects of preference. The averaging axiom claims the following:

AVERAGING: If A and B are mutually incompatible propositions, and $A \succcurlyeq B$, then $A \succcurlyeq (A \text{ or } B) \succcurlyeq B$.

The averaging axiom claims that how much an agent values a disjunction should depend on the value she assigns to the disjuncts in such a way that the disjunction cannot be more or less desirable than any of the disjuncts. When the propositions involved are outcomes that occur in different states of the world, this requirement, too, expresses the idea that there is an independence in value between what happens in separate states of the world. Knowing only that I will end up with one of two outcomes cannot be worse than ending up with any of the individual outcomes.

Assuming separability for preferences in the way that the independence axiom, the sure-thing principle and the averaging axiom do ensures that

the utility representation has an important separability feature as well. As we have seen, in expected utility theory, the overall value of an action can be represented as a probability-weighted sum of the utilities of the outcomes occurring in separate states. This means that the value contribution of an outcome in one state will be independent of the value contribution of an outcome of another state, holding the probabilities fixed. And so the separability of the value of outcomes in separate states is captured by equating the value of an action with its expected utility. If separability is problematic, it is thus problematic independently of any representation theorem. In particular, this means that it is also problematic for realists.

5.2 *Violations of Separability*

To see how separability may fail, consider the following decision problem, known as the Machina Paradox.[40] Suppose you prefer actually going to Venice to staying at home and watching a movie about Venice. You also prefer watching a movie about Venice to doing nothing and being bored. You are now offered the lotteries described in Table 5. Suppose that each lottery ticket is equally likely to be drawn, so that, if we want to apply von Neumann and Morgenstern's framework, each lottery ticket has a probability of 1%.

	Tickets 1–99	Ticket 100
Lottery A	Go to Venice	Bored at home
Lottery B	Go to Venice	Movie about Venice

Table 5: Machina's Paradox

Many people would prefer lottery A to lottery B in this context. Clearly, if I am so unlucky as to draw ticket 100, I'd rather not have to watch a movie reminding me of my misfortune. However, my preferences, as stated, violate the independence axiom and sure-thing principle. It is also clear why this violation of separability occurs. What happens in alternative, incompatible states of the world, that is, what might have been, clearly matters for how I evaluate the outcome of watching a movie about Venice. If there was a big probability that I could have gone to Venice, I will evaluate that outcome differently from when there was no such possibility. In this case, the reason for an interdependence in value between outcomes in alternative states of the world is disappointment: the movie about Venice heightens my disappointment by reminding me of what I could have had.

40 See, for instance, Mas-Colell et al. (1995), chapter 6.

The natural response to this kind of problem is to say that the outcomes in the decision problem as I stated it were under-described. Clearly, the feeling of disappointment is a relevant part of the outcomes of lottery B. There is nothing irrational about wanting to avoid disappointment, and many agents do. Thus, according to all the rules for the individuation of outcomes discussed in Section 1.4, watching a movie about Venice with disappointment should be a different outcome from watching a movie about Venice without disappointment. And then, no violation of separability occurs.

This seems to be a valid response in the case of Machina's Paradox. However, there are other violations of separability that arguably cannot be given the same treatment. One famous case that seems to be more problematic is the Allais Paradox, introduced in Allais (1953). It runs as follows. First a subject is offered a choice between $1 million for certain on the one hand, and an 89% chance of winning $1 million, a 10% chance of winning $5 million, and a 1% chance of winning nothing on the other. What she will get is decided by a random draw from 100 lottery tickets. Many people choose $1 million for certain when offered this choice. Next, the subject is offered the choice of either a 10% chance of $5 million, and nothing otherwise on the one hand, or an 11% chance of $1 million, and nothing otherwise on the other. Again, this is decided by the draw of a lottery ticket. Here, most people pick the first lottery, that is, the lottery with the higher potential winnings.

While this combination of preferences seems sensible, it in fact violates independence and the sure-thing principle, given a natural specification of the outcomes involved. This becomes evident when we represent the two choices in decision matrices, as in Tables 6 and 7.

	Tickets 1–89	Tickets 90–99	Ticket 100
Lottery C	$1 million	$5 million	$0
Lottery D	$1 million	$1 million	$1 million

Table 6: Allais Paradox: First Choice

	Tickets 1–89	Tickets 90–99	Ticket 100
Lottery G	$0	$5 million	$0
Lottery H	$0	$1 million	$1 million

Table 7: Allais Paradox: Second Choice

Choosing lottery D in the first choice, and lottery G in the second choice violates independence and the sure-thing principle. To start with the sure-

thing principle, note that in both choices, the two lotteries to be chosen from are identical with regard to what happens if tickets 1–89 are drawn. And thus, according to the sure-thing principle, the only thing that matters for the overall assessment should be what happens if tickets 90–100 are drawn. But for these tickets, the first choice, between lottery C and lottery D, and the second choice, between lottery G and lottery H are identical. And so, the agent should choose lottery D in the first choice if and only if she chose lottery H in the second choice. Similar reasoning applies for independence, if we regard each lottery as a compound lottery of the sub-lotteries involving tickets 1–89 and 90–100 respectively.

Nevertheless, choosing lottery D in the first choice and lottery G in the second choice is both common[41] and does not seem intuitively irrational. Unless some redescription strategy works to reconcile Allais preferences with expected utility theory, expected utility theory must declare these preferences irrational. Redescribing the outcomes to take account of disappointment (or regret) arguably cannot do away with the violation of separability in the Allais Paradox. Michael Weber (1998) provides an extensive argument to that effect. The Ellsberg Paradox (Section 4.2) is another case that cannot easily be dealt with by redescription. These examples suggest that there are more problematic types of interdependence in value between outcomes in different states of the world that cannot be as easily reconciled with expected utility theory as the Machina Paradox. They have consequently been an important motivation for alternatives to expected utility theory (see Section 6).

There might, however, be good arguments in favour of the verdict that violations of separability, like the Allais preferences, are genuinely irrational. Savage himself, as well as Broome (1991) argue that our reasons for choosing one act or another must depend on states of affairs where the two acts do not yield the same outcome. This seems to speak in favour of the sure-thing principle. However, as Broome acknowledges, this assumes that reasons for action themselves are separable. Somewhat more promisingly, he suggests that, if the kind of rationality we are interested in is instrumental rationality, then all our reasons for action must derive from what it would be like to have performed an action in the various states that might come about.

Buchak (2013), who, as we will see, defends an alternative to expected utility theory, argues that instrumental rationality does not require separability. In any case, note that, even if expected utility theory is right that separability is a requirement of rationality, examples like the Allais Paradox still show expected utility theory to be quite revisionary. Expected utility theory declares preferences that are common and seem intuitively

41 See, for instance Morrison (1967) for experimental evidence that many people choose this way.

reasonable as irrational. While this may not be troubling in the case of the normative and action-guiding projects, this at least seriously calls into question whether expected utility theory can serve the interpretive project.

5.3 *Separability and Risk Aversion*

Examples like the Allais Paradox seem to show that agents actually care about some values that are not separable. The Allais preferences, for instance, make sense for an agent who cares about certainty. Lottery D in the first choice seems attractive because it leads to a gain of $1 million for certain. If the agent does not care merely about the feeling of being certain, but instead cares about it actually being certain that she gets $1 million, then certainty is a value that is only realised by a combination of outcomes across different states.

Buchak (2013) calls agents who are sensitive to values that are only realised by a combination of outcomes across different states (other than expected utility itself) 'globally sensitive'. Agents who are globally sensitive are sensitive to features other than the expected utility of an act. Next to certainty, Lopes (1981, 1996) argues that mean, mode, variance, skewness and probability of loss are further global features of gambles agents may care about. She argues that a normatively compelling theory of decision-making under risk would have subjects weigh off these various different criteria. Buchak (2013), too, argues that global sensitivity can be rational, under certain constraints.[42]

It has been argued that expected utility theory has trouble more generally in accounting for our ordinary attitudes to risk. In expected utility theory, risk averse behaviour, such as preferring a sure amount of money to a risky gamble with a higher expected monetary gain, is always explained by the concavity of the utility function with regard to the good in question. When a utility function is concave, the marginal utility derived from a good is decreasing: any additional unit of the good is worth less the more of the good the agent already has. When the utility function in money is concave in this way, the expected utility of a monetary gamble will be less than the utility of the expected monetary value. And this can mean that the agent rejects gambles that have positive expected monetary value.

Figure 1 illustrates this for an agent with utility function $u(m) = \sqrt{m}$ and current wealth of $100, who is offered a 50/50 chance of either losing $100 or gaining $125. For her, the expected utility of accepting this gamble

42 There is some debate whether global sensitivity can also be made compatible with expected utility theory. Weirich (1986) argues that globally sensitive aversion to risk can be represented with disutilities that are assigned to outcomes. In the context of Buchak's theory, Pettigrew (2014) argues that the global sensitivity allowed for by her theory is compatible with expected utility theory if outcomes are appropriately redescribed.

is $0.5 \cdot \sqrt{0} + 0.5 \cdot \sqrt{225} = 7.5$. This is less than the agent's current utility level of $\sqrt{100} = 10$. The agent would reject the gamble even though it leads to an expected gain of $12.50.[43]

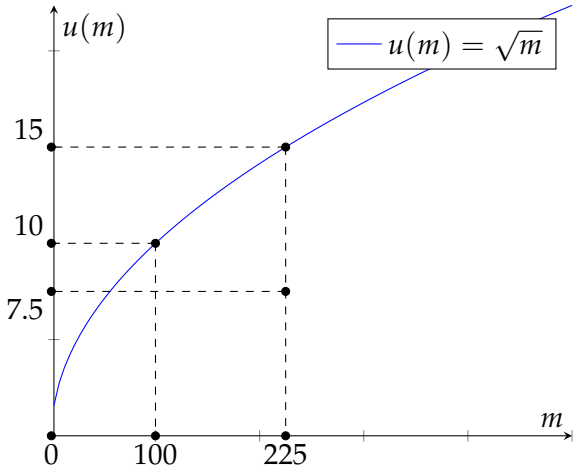


Figure 1: A concave utility function

However, there are results suggesting that decreasing marginal utility alone cannot adequately explain ordinary risk aversion. For monetary gambles, it can be shown that according to expected utility theory, any significant risk aversion on a small scale implies implausibly high levels of risk aversion on a large scale. For instance, Rabin and Thaler (2001) show that an expected utility maximiser with an increasing, concave utility function in wealth who turns down a 50/50 bet of losing $10 and winning $11 will turn down any 50/50 bet involving a loss of $100, no matter how large the potential gain. Conversely, any normal level of risk aversion for high stakes gambles implies that the agent is virtually risk neutral for small stakes gambles.[44] These results are troubling because we are all risk averse for small stakes gambles, and we are all willing to take some risky gambles with larger stakes. Moreover, this does not seem to be intuitively irrational.

Another, more direct line of critique of the way expected utility theory deals with risk aversion is available to realists about utility. If we think of utility in the realist sense, for instance as measuring the strength of our desire, it seems like we can be risk averse with regard to goods for which our utility is not diminishing. But according to expected utility theory, we

43 See Mas-Colell et al. (1995), chapter 6 for more detail on expected utility theory's treatment of risk aversion.

44 See Samuelson (1963) and Rabin (2000) for similar results.

cannot be risk averse with regard to utility itself. For realists, depending on their interpretation of utility, this may be counterintuitive.[45]

6 ALTERNATIVES TO EXPECTED UTILITY THEORY

Most alternatives to expected utility theory have been introduced as descriptive theories of choice under uncertainty, with no claim to capturing rational choice. The most well-known is prospect theory, introduced by Kahneman and Tversky (1979). Its most distinctive features are firstly, that it includes an editing phase, in which agents simplify their decision problems to make them more manageable, and secondly, that outcomes are evaluated as losses and gains relative to some reference point. In prospect theory, losses can be evaluated differently from gains. Since different ways of presenting a decision problem may elicit different reference points, this means that the agents described in prospect theory are sensitive to 'framing'. While real agents are in fact subject to framing effects,[46] sensitivity to framing is commonly regarded as irrational.

Alternatives to expected utility theory in the economic literature, too, have given up the idea that agents maximise a utility function that is independent of some reference point. Generalised expected utility theory, as developed in Machina (1982), for instance, introduces local utility functions, one for each lottery the agent may face. The lack of a stable utility function makes it difficult to interpret these theories as theories of instrumental rationality.

Other non-expected utility theories, in particular rank-dependent utility theory, as introduced by Quiggin (1982), use a stable utility function. In contrast to expected utility theory, however, they introduce alternative weightings of the utilities of outcomes. While in expected utility theory, an outcome's utility is weighted only by its probability, in rank-dependent utility theory, weights depend not only on the probability of an outcome, but also its rank amongst all the possible outcomes of the action. This allows the theory to model agents caring disproportionately about especially good and especially bad low probability outcomes.

Buchak (2013) introduces risk-weighted expected utility theory, in which a 'risk function' plays the role of the weighting function. In contrast to older rank-dependent utility theories, she argues that risk-weighted expected utility theory provides us with utilities and probabilities which can be interpreted as representing the agent's ends and beliefs respectively,

45 See Buchak (2013) for this line of critique, as well as more examples of risk aversion that expected utility has trouble making sense of.

46 See, for instance, Tversky and Kahneman (1981).

and a risk function, which represents the agent's preferences over how to structure the attainment of her ends.[47]

There is a research programme in the psychological literature that studies various heuristics that agents use when making decisions in the context of uncertainty. While these are usually not intended as normative theories of rational choice, they have plausibility as action-guiding theories—theories that cognitively limited agents may use in order to approximate a perfectly rational choice. Payne et al. (1993), for instance, introduce an adaptive approach to decision-making, which is driven by the tradeoff between cognitive effort and accuracy. Gigerenzer et al. (2000) introduce various "fast-and-frugal" heuristics to decision-making under uncertainty.

7 DYNAMIC CHOICE

So far, we have looked at individual decisions separately, as one-off choices. However, each of our choices is part of a long series of choices we make in our lives. Dynamic choice theory models this explicitly. In dynamic choice problems, choices, as well as the resolution of uncertainty happen sequentially. Dynamic choice problems are typically represented as decision trees, like the one in Figure 2. The round nodes in this tree are chance nodes, where we think of the agent as going 'left' or 'right' depending on what state of affairs comes about. The square nodes are decision nodes, where the agent can decide whether to go 'left' or 'right'.

There are a number of interesting cases where an agent ends up making a series of seemingly individually rational choices that leave her worse off than she could be.[48] Dynamic choice theory helps us analyse such cases. Here I want to focus on dynamic choice problems involving agents who violate standard expected utility theory. These cases provide some of the most powerful arguments in favour of expected utility theory, and against the alternatives canvassed in Section 6. We already mentioned Elga's dynamic choice argument against imprecise probabilities in Section 4.2. Here, I turn to arguments involving violations of separability.

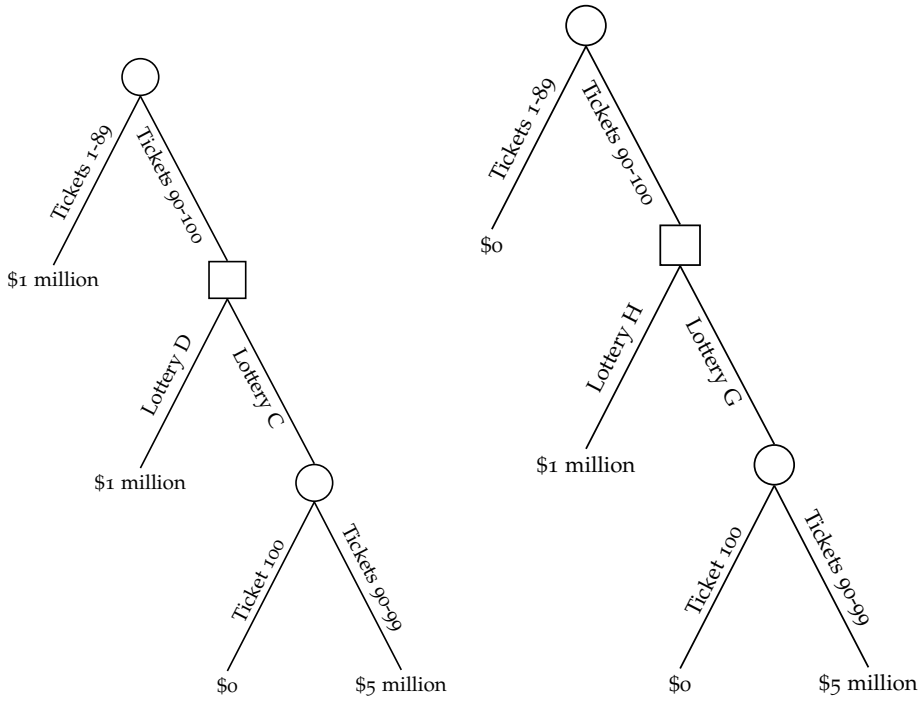
7.1 *Dynamic Arguments in Favour of Separability*

Machina (1989) discusses the following dynamic version of the Allais Paradox. This dynamic version serves as an argument against Allais preferences, and violations of separability more generally. In this dynamic

47 For an overview of other alternatives to expected utility theory in the economic literature, the two most comprehensive surveys are Schmidt (2004) and Sugden (2004).

48 One example is the Self-Torturer Problem discussed in Section 4.3. Andreou (2012) is a helpful overview of more such cases.

version, agents only get to make a decision after some of the uncertainty has already been resolved. They make a choice after they have found out whether one of tickets 1–89 has been drawn, or one of tickets 90–100 has been drawn, as shown in Figure 2.



(a) First Choice (b) Second Choice

Figure 2: Dynamic Allais Problem

The interesting feature of the dynamic case is that at the time where the agent gets to make a decision, the rest of the tree, sometimes called the 'continuation tree', looks the same for the first and second choice. We might think that this means that the agent should decide the same in both cases. But then she will end up choosing in accordance either with lotteries C and G respectively, or with lotteries D and H respectively, but not according to the Allais preferences. That in turn means that for at least one of the choices, an agent with Allais preferences will end up choosing contrary to what she would have preferred at the beginning of the decision problem, before any uncertainty has been resolved.

This has been held to be problematic for a variety of reasons. Firstly, for the agent we are considering, the dynamic structure of the decision problem clearly makes a difference to what she will choose. It can make a difference whether the agent faces a one-off choice or a dynamic version of that choice involving the same possible outcomes. But, it is claimed, for instrumentally rational agents, who care only about the final outcomes,

the temporal structure of a decision problem should not matter. Secondly, suppose the agent anticipates that, after uncertainty has been removed, she will go against the preferences she has at the outset. Such an agent would presumably be willing to pay to either not have uncertainty removed, or to restrict her own future choices. Paying money for this looks like a pragmatic cost of having these kinds of preferences. Moreover, refusing free information has been argued to be irrational in its own right.[49] Thirdly, the agent does not seem to have a stable attitude towards the choice to be made in the dynamic decision problem, even though her underlying preferences over outcomes do not change. All of these considerations have been argued to count against the instrumental rationality of an agent with Allais preferences.

Similar dynamic choice problems can be formulated whenever there is a violation of separability. In Savage's framework, whenever the agent's attitudes are non-separable for two events, one can construct decision problems where the two events are de facto 'separated' by revealing which of the events occurs before the agent gets to decide. And then parallel problems will arise. In fact, if we find the previous argument against Allais preferences convincing, we can formulate a very general argument in favour of expected utility theory. Spelling out the argument from consequentialism in Hammond (1988) in more precise terms, McClennen (1990) shows that, given some technical assumptions, expected utility theory can be derived from versions of the following principles:

NEC (NORMAL-FORM/EXTENSIVE-FORM COINCIDENCE): In any dynamic decision problem, the agent should choose the same as she would, were she to simply choose one course of action at the beginning of the decision problem.

SEP (DYNAMIC SEPARABILITY): Within dynamic decision problems, the agent treats continuation trees as if they were new trees.

DS (DYNAMIC CONSISTENCY): The agent does not make plans she foreseeably will not execute.

A similar argument is made by Seidenfeld (1988). The third condition in McClennen's formulation is fairly uncontroversial. However, those defending alternatives to expected utility theory have called into question both NEC and SEP. Buchak (2013) discusses both the strategy of abandoning SEP and that of abandoning NEC, and argues that at least one of them works.

SEP is characteristic of a choice strategy that was first described by Strotz (1956), and is now known in the literature as 'sophisticated choice'.[50] So-

49 See, for instance, Wakker (1988).

50 See McClennen (1990) for a characterisation of different dynamic choice rules.

phisticated agents treat continuation trees within dynamic choice problems as if they were new tress. Moreover, they anticipate, at the beginning of the dynamic choice problem, that they will do so. Given this prediction of their own future choice, they choose the action that will lead to their most preferred prospect. They thus follow a kind of 'backward induction' reasoning. Sophisticated agents fail to abide by NEC: they can end up choosing courses of action that are dispreferred at the beginning of the choice problem. This can be seen in our example of the dynamic Allais Paradox. Sophisticated agents behave in the way we assumed above. They thus suffer the pragmatic disadvantages we described.[51]

Those who question NEC allow that the dynamic structure of a decision problem can sometimes make a difference, even if that may have tragic consequences. But note that one can question NEC as a general principle and still think that in the particular dynamic choice problems we are considering, the pragmatic disadvantages count against having preferences that violate separability.

Because of the difficulties associated with sophistication described above, many advocates of alternatives to expected utility theory have rejected SEP instead. For instance, Machina (1989) argues that SEP is close enough to separability that accepting SEP begs the question against separability. If SEP is given up, it can make a difference to an agent if she finds herself in the middle of a dynamic choice problem rather than at the beginning of a new one. One choice rule that then becomes open to her is 'resolution', where the agent simply goes through with a plan she made at the beginning of a decision problem. Resolute agents obviously abide by NEC and avoid any pragmatic disadvantages. A restricted version of this dynamic choice rule is advocated by McClennen (1990).[52] Rabinowicz (1995) argues that sophistication and resolution can be reconciled.

7.2 *Time Preferences and Discounting*

While dynamic choice theory is concerned with the temporal sequence of our decisions, there is another branch of decision theory that is concerned with the timing of the costs and benefits that are caused by our actions. This literature studies the nature of our time preferences: do we prefer for an outcome to occur earlier or later? How much would we give up in order to receive it earlier or later?

51 In fact, Seidenfeld discusses cases where sophisticated agents end up making a sure loss.

52 Note that related notions of resolution are also discussed in the non-formal literature in order to deal with problems of diachronic choice, such as the Toxin Puzzle, described in Kavka (1983). See, for instance, Holton (2009) and Bratman (1998), as well as the discussion on the Self-Torturer Problem in Section 4.3 above.

Since most agents prefer for good outcomes to occur earlier, and bad outcomes to occur later, Samuelson (1937) proposed the discounted utility model. According to this model, agents assign the same utility to an outcome (in Samuelson's model these are consumption profiles) no matter when it occurs, but discount that utility with a fixed exponential discount rate. They can then calculate how much a future outcome is worth to them at the time of decision, and maximise their discounted utility. In the case where decisions are made under certainty, let the outcomes occurring at different points in time, up until period t, be $O_1, ..., O_t$. The agent assigns utility $u(O)$ to each of these outcomes. This is an 'instantaneous' utility function, where the timing of the outcome does not matter for the utility assignment. Moreover, let d be the discount factor. The agent's discounted utility $DU(O_1, ..., O_t)$ is then given by:

$$DU(O_1, ..., O_t) = \sum_{i=1}^{t} d^i \cdot u(O_i).$$

This discounted utility describes the current value of the stream of outcomes $O_1, ..., O_t$ to the agent. According to the discounted utility model, agents maximise this discounted utility. When we have $0 < d < 1$, the agent prefers good outcomes to occur sooner rather than later. In that case, it is also true that the value of an infinite, constant stream of benefits will be finite. Koopmans (1960) presents a number of axioms on time preferences, and provides a representation theorem for the discounted utility model.

One main advantage of being the type of agent who abides by the discounted utility model is that for such an agent, there will be no preference reversals as time moves on (this feature is sometimes referred to as 'time consistency'). That is, an agent will never suddenly reverse her preference between two actions as she gets closer in time to a choice. Yet, such preference reversals are common.[53] It has been argued that the hyperbolic discounting model advocated by Ainslie (1992), which allows for such reversals, models the ordinary decision-maker better. Whether the discounted utility model is normatively adequate is controversial, and depends in part on whether we think that time inconsistency is necessarily irrational.[54] In fact, time inconsistent preferences, just like preferences that violate expected utility theory, may lead to problematic patterns of choice in dynamic choice problems, unless the agent adopts the right dynamic choice rule.

The discounted utility model underlies much public decision-making. Discount rates are standardly applied in cost-benefit analyses. This has

53 For empirical evidence of this phenomenon, see, for instance, Thaler (1981).

54 Frederick, Loewenstein, and O'Donoghue (2002) provide a helpful overview of this debate, and the literature on time preferences more generally.

received special philosophical attention in the case of cost-benefit analyses of the effects of climate change. Ethicists and economists have debated whether a strictly positive discount rate is justified when evaluating the costs of climate change.[55] Much recent work on time preference and discounting has focused on how to discount in the context of uncertainty. Again, this question is especially important for evaluating the costs of climate change, since these evaluations are carried out in the context of great uncertainty. Gollier (2002) provides an expected utility based model of discounting under uncertainty that much of this literature appeals to. Weitzman (2009) discusses discounting in a context where our estimates of future climate have 'fat tails', and argues that fat tails make a big difference to our evaluations of the costs of climate change.

8 CONCLUDING REMARKS

This entry started out by introducing decision theories that can be classified under the heading of 'expected utility theory'. Expected utility theory is an enormously influential theory about how we do and should make choices. It has been fruitfully applied in many different fields, not least philosophy. This entry has described expected utility theory, discussed how it can be applied to the choices real agents face, and introduced debates about its foundations and interpretation.

Much recent discussion in decision theory concerns the two main types of challenge to traditional expected utility theory that the latter half of this entry focused on. The first type of challenge claims that traditional expected utility theory requires agents to have attitudes that are too fine-grained and too extensive. According to this challenge, agents have attitudes, and are rationally permitted to have attitudes that are imprecise, or vague, or incomplete. The important question arising for expected utility theory is whether it can incorporate imprecision, vagueness, and incompleteness, or whether it can instead offer a convincing argument that these attitudes are indeed irrational.

The second type of challenge questions the assumption of separability that underlies expected utility theory—that is, the assumption that the value of an outcome in one state of the world is independent of what happens in other, incompatible states of the world. According to this challenge, agents have attitudes to risky prospects that violate this assumption, and are rationally permitted to do so. This challenge, in particular, has inspired alternatives to expected utility theory. Alternatives to expected

55 See, in particular, the debate between Stern (2007) and Nordhaus (2007). For a philosopher who holds that there is no justification for time preference in public decision-making, see Broome (1994).

utility theory face challenges of their own, however, not least the question of whether they can make sense of dynamic choice.

ACKNOWLEDGEMENTS

I am grateful to Seamus Bradley, Richard Pettigrew, Sergio Tenenbaum, and Jonathan Weisberg for many helpful comments on earlier drafts of this entry.

REFERENCES

Ainslie, G. (1992). *Picoeconomics*. Cambridge University Press.

Akerlof, G. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, *84*(3), 488–500.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine. *Econometrica*, *21*(4), 503–546.

Andreou, C. (2006). Environmental damage and the puzzle of the self-torturer. *Philosophy & Public Affairs*, *37*(2), 183–93.

Andreou, C. (2012). Dynamic choice. *Stanford Encyclopedia of Philosophy*. Retrieved from http://plato.stanford.edu/archives/fall2012/entries/dynamic-choice/

Armendt, B. (1986). A foundation for causal decision theory. *Topoi*, *5*, 3–19.

Arntzenius, F. (2008). No regrets, or: Edith piaf revamps decision theory. *Erkenntnis*, *68*, 277–297.

Arntzenius, F. & McCarthy, D. (1997). Self torture and group beneficence. *Erkenntnis*, *47*(1), 129–44.

Augustin, T., Coolen, F., de Cooman, G., & Troffaes, M. (2014). *Introduction to imprecise probabilities*. Wiley Series in Probability and Statistics. Wiley.

Bentham, J. (1789/2007). *An introduction to the principles of morals and legislation*. Dover Publications.

Bermudez, J. L. (2009). *Decision theory and rationality*. Oxford University Press.

Bernouilli, D. (1738/1954). Exposition of a new theory on the measurement of risk. *Econometrica*, *22*(1), 23–36.

Bolker, E. (1966). Functions resembling quotients of measures. *Transactions of the American Mathematical Society*, *2*, 292–312.

Bradley, R. (1998). A representation theorem for a decision theory with conditionals. *Synthese*, *116*, 187–229.

Bradley, R. (2004). Ramsey's representation theorem. *Dialectica*, *58*(4), 483–497.

Bradley, S. (2015). Imprecise probabilities. In E. Zalta (Ed.), *Stanford encyclopedia of philosophy* (Summer 2015). Retrieved from http://plato.stanford.edu/archives/sum2015/entries/imprecise-probabilities/

Bradley, S. & Steele, K. (2014a). Should subjective probabilities be sharp? *Episteme*, *11*, 277–289.

Bradley, S. & Steele, K. (2014b). Uncertainty, learning, and the "problem" of dilation. *Erkenntnis*, *79*(6), 1287–1303.

Bratman, M. (1998). Toxin, temptation, and the stability of intention. In J. Coleman, C. Morris, & G. Kavka (Eds.), *Rational commitment and social justice: Essays for gregory kavka* (pp. 59–83). Cambridge University Press.

Breen, R. & Goldthorpe, J. (1997). Explaining educational differentials: Towards a formal rational action theory. *Rationality and Society*, *9*(3), 275–305.

Broome, J. (1991). *Weighing goods*. Blackwell.

Broome, J. (1994). Discounting the future. *Philosophy and Public Affairs*, *23*, 128–156.

Buchak, L. (2013). *Risk and rationality*. Oxford University Press.

Buchak, L. (2016). Decision theory. In A. Hajek & C. Hitchcock (Eds.), *The oxford handbook of probability and philosophy*. Oxford University Press.

Camerer, C. & Weber, M. [Martin]. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, *5*, 325–370.

Davidson, D. (1973). Radical interpretation. *Dialectica*, *27*, 313–328.

Davidson, D. (1985). A new basis for decision theory. *Theory and Decision*, *18*, 87–98.

Davidson, D., McKinsey, J. C. C., & Suppes, P. (1955). Outlines of a formal theory of value, i. *Philosophy of Science*, *22*, 140–160.

de Finetti, B. (1974). *Theory of probability*. Wiley.

Downs, A. (1957). *An economic theory of democracy*. Harper.

Dreier, J. (1996). Rational preference: Decision theory as a theory of practical rationality. *Theory and Decision*, *40*(3), 249–276.

Eells, E. (1981). Causality, utility, and decision. *Synthese*, *48*, 295–329.

Egan, A. (2007). Some counterexamples to causal decision theory. *Philosophical Review*, *116*, 93–114.

Einav, L. & Finkelstein, A. (2011). Selection in insurance markets: Theory and empirics in pictures. *Journal of Economic Perspectives*, *25*(1), 115–138.

Elga, A. (2010). Subjective probabilities should be sharp. *Philosophers' Imprint*, *10*.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics*, *75*(4), 643–669.

Epstein, L., Landes, W., & Posner, R. (2013). *The behavior of federal judges: A theoretical and empirical study of rational choice*. Harvard University Press.

Feddersen, T. (2004). Rational choice theory and the paradox of not voting. *The Journal of Economic Perspectives*, *18*(1), 99–112.

Fermat, P. & Pascal, B. (1654/1929). Fermat and pascal on probability. In *A source book in mathematics*. McGraw-Hill Book Co.

Fishburn, P. (1981). Subjective expected utility: A review of normative theories. *Theory and Decision*, *13*, 139–199.

Fishburn, P. (1998). Stochastic utility. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory* (Vol. 1). Kluwer.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*(2), 351–401.

Gibbard, A. & Harper, W. (1978/1981). Counterfactuals and two kinds of expected utility. In W. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs: Conditionals, belief, decision, chance, and time* (pp. 153–190). Reidel.

Gigerenzer, G., Todd, P. M., & Group, A. R. (2000). *Simple heuristics that make us smart*. Oxford University Press.

Gollier, C. (2002). Discounting an uncertain future. *Journal of Public Economics*, *85*(2), 149–166.

Hajek, A. (2008). Arguments for – or against – probabilism. *British Journal for the Philosophy of Science*, *59*(4), 793–819.

Hammond, P. (1988). Consequentialist foundations for expected utility. *Theory and Decision*, *25*, 25–78.

Hare, R. (1981). *Moral thinking*. Oxford University Press.

Hausman, D. (2000). Revealed preference, belief, and game theory. *Economics and Philosophy*, *16*(1), 99–115.

Holton, R. (2009). *Willing, wanting, waiting*. Oxford University Press.

Horgan, T. (1981/1985). Counterfactuals and newcomb's problem. In *Paradoxes of rationality and cooperation: Prisoner's dilemma and newcomb's problem* (pp. 159–182). University of British Columbia Press.

Horwich, P. (1987). *Asymmetries in time*. MIT Press.

Jackson, F. (1991). Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics*, *101*(3), 461–482.

Jeffrey, R. (1965/1983). *The logic of decision* (2nd). University of Chicago Press.

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge University Press.

Joyce, J. M. (2011). A defense of imprecise credence in inference and decision. *Philosophical Perspectives*, *24*, 281–323.

Joyce, J. M. (2012). Regret and instability in causal decision theory. *Synthese*, *187*, 123–145.

Joyce, J. M. & Gibbard, A. (1998). Causal decision theory. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory* (Vol. 1). Kluwer.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kaplan, M. (1983). Decision theory as philosophy. *Philosophy of Science*, *50*, 549–577.

Kaplan, M. (1996). *Decision theory as philosophy*. Cambridge University Press.

Kavka, G. (1983). The toxin puzzle. *Analysis*, *43*(1), 33–36.

Koopmans, T. (1960). Stationary ordinal utility and impatience. *Econometrica*, *28*, 287–309.

Levi, I. (1980). *The enterprise of knowledge*. MIT Press.

Lewis, D. (1974). Radical interpretation. *Synthese*, *23*, 331–344.

Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, *59*(1), 5–30.

Lockhart, T. (2000). *Moral uncertainty and its consequences*. Oxford University Press.

Lopes, L. (1981). Decision making in the short run. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 377–385.

Lopes, L. (1996). When time is of the essence: Averaging, aspiration, and the short run. *Journal of Experimental Psychology*, *65*(3), 179–189.

Luce, D. & Suppes, P. (1965). Preference, utility, and subjective probability. In e. a. Luce Duncan (Ed.), *Handbook of mathematical psychology* (Vol. 3, pp. 249–410). Wiley.

Machina, M. (1982). 'expected utility' analysis without the independence axiom. *Econometrica*, *50*(2), 277–323.

Machina, M. (1989). Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, *27*(4), 1622–1668.

Mahtani, A. (2019). Imprecise probability. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*.

Mas-Colell, A., Whinston, M., & Green, J. (1995). *Microeconomic theory* (1st ed.). Oxford University Press.

McClennen, E. (1990). *Rationality and dynamic choice: Foundational explorations*. Cambridge University Press.

Meacham, C. & Weisberg, J. (2011). Representation theorems and the foundations of decision theory. *Australasian Journal of Philosophy*, *89*(641–663).

Mill, J. S. (1861/1998). *Utilitarianism* (R. Crisp, Ed.). Oxford University Press.

Morrison, D. (1967). On the consistency of preferences in allais' paradox. *Behavioral Science*, *12*(5), 373–383.

Moss, S. (2015). Time-slice epistemology and action under indeterminacy. *Oxford Studies in Epistemology*.

Nordhaus, W. (2007). A review of the stern review on the economics of global warming. *Journal of Economic Literature*, *155*, 686–702.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in honor of carl g. hempel* (pp. 114–115). Synthese Library. Reidel.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.

Pedersen, A. & Wheeler, G. (2014). Demystifying dilation. *Erkenntnis*, *79*(6), 1305–1342.

Pettigrew, R. (2011). Epistemic utility arguments for probabilism. *The Stanford Encyclopedia of Philosophy*. Retrieved from http://plato.stanford.edu/archives/win2011/entries/epistemic-utility/

Pettigrew, R. (2014). *Risk, rationality, and expected utility theory*. APA author meets critic session.

Pettit, P. (1991). Decision theory and folk psychology. In M. Bacharach & S. Hurley (Eds.), *Foundations of decision theory: Issues and advances* (pp. 147–175). Blackwell.

Price, H. (1991). Agency and probabilistic causality. *British Journal for the Philosophy of Science*, *42*(2), 157–176.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, *3*(4), 323–343.

Quinn, W. (1990). The puzzle of the self-torturer. *Philosophical Studies*, *59*(1).

Rabin, M. (2000). Risk aversion and expected utility: A calibration theorem. *Econometrica*, *68*(1281–1292).

Rabin, M. & Thaler, R. (2001). Anomalies: Risk aversion. *Journal of Economic Perspectives*, *15*, 219–232.

Rabinowicz, W. (1995). To have one's cake and eat it, too: Sequential choice and expected utility violations. *Journal of Philosophy*, *92*(11), 586–620.

Ramsey, F. P. (1926/2010). Truth and probability. In A. Eagle (Ed.), *Philosophy of probability: Contemporary readings* (pp. 52–94). Routledge.

Sahlin, N.-E. & Weirich, P. (2014). Unsharp sharpness. *Theoria*, *80*, 100–103.

Salles, M. (1998). Fuzzy utility. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory* (Vol. 1). Kluwer.

Samuelson, P. (1937). A note on measurement of utility. *Review of Economic Studies*, *4*, 155–161.

Samuelson, P. (1963). Risk and uncertainty: A fallacy of large numbers. *Scientia*, *98*(108–113).

Sartre, J.-P. (1945/2007). *Existentialism is a humanism* (A. Elkaïm-Sartre, Ed.). Yale University Press.

Savage, L. (1954). *The foundations of statistics*. Wiley.

Schmidt, U. (2004). Alternatives to expected utility theory: Formal theories. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory* (pp. 757–837). Kluwer.

Seidenfeld, T. (1988). Decision theory without "independence" or without "ordering". *Economics and Philosophy*, *4*, 267–290.

Seidenfeld, T. & Wasserman, L. (1993). Dilation for sets of probabilities. *The Annals of Statistics*, *21*(3), 1139–1154.

Sepielli, A. (2013). Moral uncertainty and the principle of equity among moral theories. *Philosophy and Phenomenological Research*, *86*(3), 580–589.

Simon, H. (1976). From substantive to procedural rationality. In T. J. Kastelein, S. K. Kuipers, W. A. Nijenhuis, & G. R. Wagneaar (Eds.), *25 years of economic theory* (Vol. 2, pp. 65–86). Springer US.

Singer, P. (1993). *Practical ethics*. Cambridge University Press.

Spohn, W. (1977). Where luce and krantz do really generalize savage's decision model. *Erkenntnis*, *11*, 113–134.

Stalnaker, R. (1972/1981). Letter to david lewis. In W. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs: Conditionals, belief, decision, chance, and time* (pp. 151–152). Reidel.

Stern, N. (2007). *The economics of climate change*. Cambridge University Press.

Strotz, R. H. (1956). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, *23*(3), 165–180.

Sugden, R. (2004). Alternatives to expected utility theory: Foundations. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory* (pp. 685–755). Kluwer.

Tenenbaum, S. & Raffman, D. (2012). Vague projects and the puzzle of the self-torturer. *Ethics*, *123*(1), 86–112.

Thaler, R. (1981). Some empirical evidence on dynamic inconsistency. *Economic Letters*, *8*(3), 351–401.

Troffaes, M. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, *45*, 17–29.

Tversky, A. & Kahneman, D. (1974). Judgements under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Velleman, D. (1993/2000). The story of rational action. In *The possibility of practical reason*. Oxford University Press.

von Neumann, J. & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Voorhoeve, A. & Binmore, K. (2006). Transitivity, the sorites paradox, and similarity-based decision-making. *Erkenntnis*, *64*(1), 101–114.

Wakker, P. (1988). Nonexpected utility as aversion to information. *Journal of Behavioral Decision Making*, *1*, 169–175.

Weber, M. [Max]. (1922/2005). *Wirtschaft und gesellschaft. grundriss der verstehenden soziologie* (A. Ulfig, Ed.). Zweitausendeins-Verlag.

Weber, M. [Michael]. (1998). The resilience of the allais paradox. *Ethics*, *109*(1), 94–118.

Weirich, P. (1986). Expected utility and risk. *British Journal for the Philosophy of Science*, *37*, 419–442.

Weitzman, M. (2009). On modeling and interpreting the economics of catastrophic climate change. *The Review of Economics and Statistics*, *91*(1), 1–19.

White, R. (2010). Evidential symmetry and mushy credence. *Oxford Studies in Epistemology*, *3*, 161–186.

Williams, J. R. G. (2014). Decision-making under indeterminacy. *Philosophers' Imprint*, *14*(4), 1–34.

Williamson, J. (2010). *In defense of objective bayesianism*. Oxford University Press.

Zynda, L. (2000). Representation theorems and realism about degrees of belief. *Philosophy of Science*, *67*(1), 45–69.

3

IMPRECISE PROBABILITIES

Anna Mahtani

Suppose we take a standard, randomly shuffled pack of cards with no jokers, and ask what the probability is that the top card is a red picture card. We can calculate the probability of this to be $6/52 = 3/26$. And of course if, as many think, people have degrees of belief, or credences, then you—knowing only that the pack is normal and has been randomly shuffled—should have a credence of $3/26$ that the top card is a red picture card.

But many think that you can have credences in all sorts of claims, and not just claims about random events involving cards, dice or coins. In particular, classical Bayesian epistemologists think that you have a credence in every proposition that you can entertain. Thus for example, there is some number between 0 and 1 that is your credence that it will snow in London on New Year's Day 2026; and there is some number between 0 and 1 that is your credence that I have a cup of tea beside my computer as I type. But what exactly are your credences in these claims? Perhaps no particular number springs to mind. Unlike in the playing card scenario above, here there does not seem to be any obvious way to 'work out' what the probability of these events is and so arrive at the precise credence that you ought to have. Cases like these have led some to reject the classical Bayesian epistemologist's claim that people must have precise credences in every proposition that they can entertain. Instead it is claimed people—even rational people—can have imprecise credences in at least some propositions. Hereafter I will use 'imprecise probabilism' as the name for the view that rational people can have imprecise credences.

Imprecise probabilism has some intuitive appeal. Take again the claim (which we can call 'NYD') that it will snow in London on New Year's Day 2026. It is hard to put a precise number on your credence—but there may still be something we can say about your attitude towards this proposition. For example, perhaps you think the claim is not very likely, but far from impossible, and certainly more likely than the claim (which we can call 'MIDSUMMER') that it will snow on Midsummer's day in London in 2026. We might think that your credences in these claims can be represented with ranges, rather than points. For example, perhaps your credence in NYD is the range $(0.1, 0.4)$, and your credence in MIDSUMMER is the range $(0.01, 0.05)$. Versions of this idea—of representing credences by ranges rather than points—can be found in numerous sources, including R. Bradley (2009), Gärdenfors and Sahlin (1982), Jeffrey (1983), Joyce (2005),

Kaplan (1996), Keynes (1921), Kyburg (1983), Levi (1974), Sturgeon (2008), van Fraassen (2006), and Walley (1991).

To explain imprecise probabilism in more depth, I must first set out the classical Bayesian view more precisely. We begin with a set (an *event space*) $\Omega = \{w_1, w_2, \ldots, w_n\}$. Each w_i in Ω is a state of affairs, or possible world. We can then see a proposition (or event) A as a subset of set Ω. For example, suppose we take the proposition that a particular dice throw landed on an even number. This proposition obtains at all those possible worlds where the dice lands on 2, 4, or 6. Thus there will be a set of possible worlds where the proposition obtains. For the purposes of this topic, we assume that the proposition can be identified with that set of possible worlds at which it is true.

Now consider a set $\mathcal{F} = \{A_1, A_2, \ldots, A_m\}$ of these propositions (which are themselves each sets of possible worlds). To make this set a sigma algebra (σ-algebra), the set must be closed under union, intersection, and complementation. For the set to be closed under union, it must be the case that for two propositions A_i and A_j in the set, the union $(A_i \cup A_j)$ is also in the set; similarly, for the set to be closed under intersection, it must be the case that for any two propositions A_i and A_j in the set, the intersection $(A_i \cap A_j)$ must also be in the set; and for the set to be closed under complementation it must be the case that for any proposition A_i in the set, the proposition $\Omega - A_i$ must also be in the set.

Finally we introduce a function p mapping $\mathcal{F}$ to $[0, 1]$. Thus for example, if $\mathcal{F}$ contains some proposition A, then our function p will map that proposition A to some number between 0 and 1. If the function is a probability function, then it will meet these three conditions (the probability axioms):

1. $p(A) \geq 0$, for all A in $\mathcal{F}$,
2. $p(\Omega) = 1$,
3. If $A_i \cap A_j = \emptyset$ then $p(A_i \cup A_j) = p(A_i) + p(A_j)$.

In contrast according to the imprecise probabilist, a rational agent may have an epistemic state that cannot be represented by a single probability function. Instead, the imprecise probabilist typically claims that a rational agent's epistemic state can be represented by a set of probability functions $P = \{p_1, p_2, \ldots, p_k\}$. Thus rather than assigning each proposition in $\mathcal{F}$ some unique number, for each proposition A in $\mathcal{F}$ there will be some least number assigned to it by the probabilities in P (the agent's *lower envelope* of A), and some greatest number assigned to A by the probabilities P (the agent's *upper envelope* of A).

Thus the imprecise probabilist moves away from the classical Bayesian view by claiming that an agent's epistemic state is given not by a single function from propositions to numbers, but by a set of such functions. And

if the agent is rational, then each of the functions in the set that represents the agent's epistemic state will be a probability function. In van Fraassen's terminology, this set of functions is the agent's *representor* (van Fraassen, 1990). On another vivid version of this account, we can see the set as a group of *avatars* (R. Bradley, 2009), each of whom has a precise credence function: these avatars collectively represent the agent's epistemic state.

This view raises some interesting problems, but before I turn to these, I will first explore in more depth the central claims of the view.

1 EXPLORING THE VIEW

How can a set of credence functions represent an agent's epistemic state? Or, to put the point another way, what must be true of a given agent for her epistemic state to be correctly represented by some specific set of credence functions?[1]

The idea is that what holds across all the credence functions in the set, holds for the agent's epistemic state.[2] Thus for example, suppose that every credence function in the set assigns the number 0.5 to the claim that the next fair coin tossed will land heads: then it follows that the agent has a credence of precisely 0.5 in this claim. Or suppose that every credence function in the set assigns a credence of no more than 0.4 to the claim NYD: then it follows that the agent has a credence of no more than 0.4 in this claim. Or suppose that every credence function in the set assigns a higher credence to NYD than it does to MIDSUMMER: then it follows that the agent has a higher credence in the claim NYD than she does in MIDSUMMER.

On this picture, there may be some questions we can ask about the agent's epistemic state which have no answer. For example, we might wonder which of a pair of claims is given the highest credence, or whether they are given equal credence—but there may be no answer to this question if the credence functions that represent the agent's epistemic state conflict over this. Similarly, on learning that an agent has a credence of no more than 0.4 in NYD, we might ask what exactly the agent's credence is in this claim. But there is no answer to this question if the different credence functions that represent the agent's epistemic state assign different values to this claim. In such cases, it is natural to say that the agent's credence in a claim is a range rather than a single unique number—where the range contains all and only those numbers that are assigned to the relevant proposition by some credence function from the set that represents the agent's epistemic state.

1 Richard Bradley argues that for any given epistemic state, there is a *unique* maximal set of such functions that represents that epistemic state (R. Bradley, 2009, p. 242).

2 Or perhaps, is *determinately* true of the agent's epistemic state (Rinard, 2015).

I turn now to consider some variations on this view, and some initial objections and clarifications.

1.1 *Variations on the view*

Here I contrast two different sorts of imprecise probabilist. All proponents of imprecise probabilism agree that agents are sometimes permitted to have imprecise credences in some propositions. They thus stand in contrast to the classical Bayesian epistemologists, according to whom rational agents have precise credences in every proposition which they can entertain. But even amongst those who accept imprecise probabilism, there is disagreement over whether imprecise credences are ever *required* by rationality.

James Joyce, for example, argues that one's degrees of belief should be no sharper than the evidence requires (Joyce, 2005): Joyce requires an agent to have an imprecise credence in a claim where the evidence for that claim does not justify a more precise credence. Thus for example consider again the claim NYD, that it will snow in London on New Year's Day 2026. Given that the evidence for this is (as yet) slight, an agent who had a precise credence in this claim (e.g. a credence of exactly 0.35) would be irrational. In contrast, take the claim that the next fair coin tossed will land heads. Given that the chance of this event is known to be 0.5, it is rational to have a credence of exactly 0.5 in this claim.

To clarify this view, we need to explain what determines the correct imprecise credence for an agent to have in any given situation. One possible answer to this is the *chance grounding thesis*: "one's spread of credence should cover the range of chance hypotheses left open by the evidence" (White, 2009, p. 174).[3] To see what this means, let us consider a few examples. First take an agent who knows that a coin is fair, and is contemplating the claim, HEADS, that on the next toss the coin will land heads. Given that (s)he knows that the chance of HEADS is 0.5, the chance grounding thesis requires that every credence function in the set that represents the agent's epistemic state must assign 0.5 to HEADS—and so the agent must herself have a credence of precisely 0.5 in HEADS. Now suppose instead that the agent has a coin that she does not know to be fair: the chance of its landing heads (HEADS*) is anywhere within the range $(0.2, 0.8)$, for all she knows. Then the chance grounding thesis requires that for each value v within the range $(0.2, 0.8)$, there must be a credence function in the set that represents the agent's epistemic state that assigns v to HEADS*. And furthermore there must be no credence function in the set that assigns to HEADS* some value v that is outside the range $(0.2, 0.8)$.

3 White defines this thesis, but does not endorse it.

This chance grounding thesis generates some counterintuitive results, and Joyce argues that it should be replaced with the less stringent demand that when your *only* relevant evidence is that the chance of some event is within some interval (a, b), then your spread of credence ought to cover this range (Joyce, 2010, p. 289). So for example suppose that in the case above, you know not only that the chance of the coin's landing heads is within the range $(0.2, 0.8)$, but also that the coin was selected at random from a bag which contained a variety of coins with complementary biases: i.e. for each coin in the bag that has a chance v of landing heads, the bag also contained exactly one coin with a chance $1 - v$ of landing heads. In this case, because you have this extra piece of evidence, your "spread of credence" in HEADS is not required to cover the whole range $(0.2, 0.8)$, and a credence of precisely 0.5, say, is permitted. However if you know only that the chance of the coin's landing heads is within the range $(0.2, 0.8)$, then your spread of credence in HEADS is required to cover the whole range $(0.2, 0.8)$.

Now we turn to consider imprecise probabilists who permit, but never require agents to have imprecise credences. For these theorists, an agent is free to have a credence of precisely 0.35 in the claim NYD (that it will snow in London on New Year's Day 2026). To these theorists, we might ask whether there are any rational constraints on an agent's epistemic state, bar the requirement that their state should be represented by some maximal set of credence functions that obey the probability axioms. Such a theorist might require that any rational agent's epistemic state will conform to the *principal principle*—i.e. that the agent's credence in any claim P conditional on the chance of P being some value v, is v (Lewis, 1980). From this, it follows that in the case where an agent is contemplating the claim (HEADS) that on its next toss a coin known to be fair will land heads, the agent's credence in HEADS must be 0.5. But what constraint is placed on the agent in the case where (s)he is contemplating the claim (HEADS*) that on its next toss a coin known to have a chance within the range $(0.2, 0.8)$ will land heads? The principal principle here requires that the agent's credence should not exceed the range $(0.2, 0.8)$, but nothing seems to require that the agent's credence should occupy this entire range.

Having explored this variation in the views of imprecise probabilists, I turn now to contrast the account with an alternative view.

1.2 *Dempster-Shafer Theory*

An alternative approach to modelling our epistemic state involves *belief functions* (Dempster, 1967, 1968; Shafer, 1976). To illustrate this view, we can again take the proposition (NYD) that it will snow in London on New Year's Day in 2026, and suppose that my belief-function assigns a value of

0.6 to this claim: we represent this by writing $Bel(\text{NYD}) = 0.6$. If my belief function was a probability function, then it would follow that the value assigned to the negation of NYD (i.e. to not-NYD) would be 0.4. However a belief function need not be a probability function, and it might assign any value less than or equal to 0.4 to not-NYD. Thus for example, it might assign a value of 0 to not-NYD. This is despite the fact that the value assigned to the tautology (either NYD or not-NYD) must be 1.

More generally, on this view the value assigned to the disjunction of two disjoint propositions A and B, $Bel(A \cup B)$, need not equal the sum of $Bel(A)$ and $Bel(B)$. The requirement is only that the value assigned to the disjunction must be at least as great as the sum of the values assigned to the disjuncts. Thus the belief function is not a probability function, as the third probability axiom (countable additivity) does not apply.

One way to interpret the idea of a belief function, is as a measure of the weight of evidence for each proposition. Thus consider again my belief function that assigns a value of 0.6 to NYD. We can suppose that I have asked a friend whether it will snow in London on New Year's Day 2016, and (s)he assures me that it will. I consider this friend to be reliable in 60% of cases of this sort, and this explains why my belief function assigns a value of 0.6 to this claim. If we suppose that this is all the relevant evidence that I have, then my belief function assigns a value of 0 to not-NYD simply because I have no evidence to support not-NYD. In cases where I have evidence from two different sources (e.g. in a case where I make another friend who also gives me his or her opinion on NYD), then the belief functions that result from these different bodies of evidence need to be combined, and Dempster and others have explored the question of how this combination should be carried out (Dempster, 1967).

In common with imprecise probabilism—and in apparent contrast with classical Bayesianism—this theory has resources designed to model severe uncertainty. To see this, suppose that a coin is about to be tossed, and that you have no information whatsoever about whether the coin is fair or how it might be biased. On the classical Bayesian view, in spite of your severe uncertainty, you will nevertheless have a precise probability that the coin will land head-side-up. This strikes many as counterintuitive. Advocates of both imprecise probabilism and Dempster-Shafer theory take their theories to improve on classical Bayesianism here. According to imprecise probabilism, in the case where you have no information about the bias of the coin, a rational agent may—and on some versions of the theory, must—have a credal range of $(0, 1)$ rather than a precise credence of 0.5. And according to Dempster-Shafer theory, in a case where you have no information about the bias of the coin, you have no evidence in favour of heads, and no evidence in favour of tails, and so your belief function will assign a value of 0 to both HEADS and TAILS.

For more on the Dempster-Shafter theory, and how it differs from both classic Bayesianism and imprecise probabilism, see Halpern (2003) and Yager and Liu (2008).

1.3 *Scoring Rules*

I turn now to an issue for those theorists who want to apply the idea of accuracy *scoring rules* in the context of imprecise probabilism. I begin by outlining a standard proposal for measuring the (in)accuracy of a credence function, and I explain how this sort of scoring rule has been used to construct an argument for probabilism. I then gesture towards some of the challenges that arise when we consider these measures of accuracy in the context of imprecise probabilism.

Let's begin then with the classical Bayesian picture, according to which a rational agent's epistemic state is represented with a single precise credence function. In this context a variety of scoring rules have been proposed for measuring a credence function's (in)accuracy at a given world. One popular such rule is the *Brier score* (Brier, 1950) which I outline here. First we set the truth-value of a proposition at a world to 1 if the proposition is true there, and 0 if it is false. Now we can measure the "distance" between the truth-value of the proposition at a world and the credence assigned to it, by taking the difference between the two and squaring the result. To illustrate this, suppose that you have a credence of 0.8 in the proposition that the world's population is over 7 billion in 2016. In the actual world, this proposition is true, and so has a truth-value of 1. Thus we measure the distance between the credence assigned to this proposition and its truth-value in the actual world as follows: take the truth value of the proposition (1), deduct the value assigned to it by the credence function (0.8), and then square the result (giving 0.04). We get the inaccuracy score for an entire credence function at a world by calculating this distance for each proposition that is assigned a value by the credence function, and summing the lot.

The Brier score is just one suggestion for measuring inaccuracy, and others have been proposed, along with various claims about conditions that any scoring rule ought to fulfil. One such requirement is that a scoring rule ought to be *proper*, which can be defined as follows: any agent with a rationally permissible credence function (i.e. one that obeys the probability axioms), will score her own credence function to be no more inaccurate than every other credence function, if the scoring rule that she uses is proper. The Brier score is one example of a scoring rule that meets this requirement.

Scoring rules of this sort have been used to argue for probabilism—i.e. for the claim that a rational agent's credence function obeys the probability

axioms. The argument works by showing that for any credence function Cr that does not obey the probability axioms, there is an alternative credence function Cr^* which does obey the probability axioms and which dominates Cr in the following sense: the inaccuracy of Cr is at least as great as the inaccuracy of Cr^* at every world, and at some world the inaccuracy of Cr is greater than the inaccuracy of $Cr*$. Thus, the argument goes, it would be irrational to have a credence function such as Cr which does not obey the probability axioms, when an alternative credence function Cr^* is available. Arguments of this sort can be constructed using any scoring rule provided that it meets certain requirements—including the requirement that it be proper (Joyce, 1998). Arguments from accuracy for a variety of other epistemic principles have also been proposed, including an argument for the principal principle (Pettigrew, 2013), and conditionalization (Greaves & Wallace, 2006).

We can now consider how these issues are affected by a switch from precise to imprecise probabilities. If an agent has an imprecise credence function, then how should the inaccuracy of her credence function be measured? We can see at once that the original measures of inaccuracy cannot be straightforwardly carried across—for where an agent's credence in some proposition is imprecise, we have no single number which measures that agent's credence, and so cannot make sense of the idea of deducting the agent's credence in a given proposition from its truth-value at some world. Thus a new way of measuring inaccuracy is needed.

There is not yet any consensus as to what this new way of measuring inaccuracy would be like. Some authors have proposed requirements that any way of measuring the inaccuracy of an imprecise credence function would need to meet, and some have uncovered difficulties for the project. Seidenfeld, Schervish, and Kadane argue that there is no strictly proper scoring rule for imprecise probabilities. See Seidenfeld, Schervish, and Kadane (2012) and Mayo-Wilson and Wheeler (2016) for further discussion on this issue. Schoenfield (2017) argues that if the new accuracy scoring rule meets certain conditions, then the claim that accuracy is all that matters is incompatible with the claim that imprecise probabilities are sometimes rationally required—or even permitted. Thus challenges await those who wish to endorse both imprecise probabilism and accuracy arguments.

Having explored the account of imprecise probabilities, I turn now to some of the most discussed objections and problems for the account. I divide these into two categories: learning and deciding.

2 LEARNING

On the classic Bayesian picture, an agent's epistemic state is represented by a single credence function. If the agent is rational, then she will update (only) by conditionalization. Thus for example suppose that an agent is about to run an experiment at the end of which she will have learnt (just) either E or not-E. At the start of the experiment (at t_0) let's suppose that the agent has a credence of 0.2 in E, and a credence of 0.5 in some hypothesis H. Furthermore, the agent has a conditional credence of 0.9 in H given E: in other words, if we let Cr_0 name the agent's credence at t_0, then $Cr_0(H \mid E) = Cr_0(H \cap E)/Cr_0(E) = 0.9$. Now suppose that the experiment runs, and at t_1 the agent discovers E. The agent's new t_1 credence function (Cr_1) ought rationally to be her old t_0 credence function (Cr_0) conditionalized on the new evidence that she has gained, E. Thus her new credence in H ought to be her old conditional credence in H given E: $Cr_1(H) = Cr_0(H \mid E) = 0.9$.

For the proponent of imprecise probabilities, an agent's epistemic state is represented by a set of credence functions. How will a rational agent adjust her epistemic state in the light of evidence on this account? The idea standardly endorsed by imprecise probabilists is that each credence function in the set will be adjusted in the usual way by conditionalization, and the agent's new, post-evidence epistemic state can be represented by this adjusted set of credence functions. Thus for example, to return to our experiment case above, suppose that every credence function in the set that represents the agent's epistemic state at t_0 assigns a number within the range $(0.4, 0.6)$ to H—and every number within this range is assigned to H by some credence function in the set. And suppose furthermore that for each of these credence functions, the conditional credence assigned to H given E is within the range $(0.85, 0.95)$—and every number within this range is the conditional credence assigned to H given E by some credence function within the set. Then at t_1, when the agent has learnt (just) E, the agent's epistemic state will be represented by the original set of credence functions each conditionalized on E, and thus the agent's new credence in H will be given by the range $(0.85, 0.95)$. I will now turn to two problems—both related to learning—for the proponent of imprecise probabilities.

2.1 *Belief Inertia*

Let us consider a scenario in which you have just selected a coin from a bag, knowing only that the bag contains various coins some of which may be biased to various unspecified degrees. You are going to toss the coin 25 times, and before you begin tossing the coin (a time we can call t_0) you

contemplate claim HEADS25—the claim that the coin will land heads on its 25th toss. According to any proponent of imprecise probabilities, you are permitted to have an imprecise credence in this claim. Now we can consider what will happen to your credence in HEADS25 if you toss the coin a few times, and it lands heads each time. Let HEADS1 be the claim that the coin lands heads on the first toss, HEADS2 be the claim that the coin lands heads on its second toss, and so on. Intuitively, your credence in HEADS25 ought to increase on learning HEADS1, and increase even more on learning (HEADS1 $\cap$ HEADS2), and so on.

For a certain sort of proponent of imprecise probabilism, this scenario is problematic. In particular, consider the sort of imprecise probabilist who claims that an agent's epistemic state should conform to the chance grounding thesis.[4] On this view, all and only those credence functions which are compatible with the known chances must be included in the set that represents the agent's epistemic state. In the scenario that we are considering, at t_0 you can rule out very few chance hypotheses: for all you know, the chance of HEADS25 may be any number strictly between 0 and 1. Thus at t_0 your credence in HEADS ought rationally to be the range $(0, 1)$. What happens if you toss the coin once and it lands heads—i.e. if you learn HEADS1? For any number n within the range $(0, 1)$, you have not learnt that the chance of HEADS25 is not n. For example, you have not learnt that the chance of HEADS25 is not 0.0001. Thus your new credence in HEADS25, after learning HEADS1, ought still to be the range $(0, 1)$. What happens if you toss the coin again, and it again lands heads—i.e. in addition to HEADS1, you also learn HEADS2? You cannot then rule out any additional chance hypotheses. For example, it may still be the case, for all you know, that the chance of HEADS25 is 0.0001. Thus your credence in HEADS25 after learning both HEADS1 and HEADS2 remains the range $(0, 1)$. This pattern continues: even if you toss the coin 24 times and it lands heads on each toss, your credence in HEADS25 should still remain fixed at $(0, 1)$. In this sense, your epistemic state exhibits inertia in the face of evidence. That your epistemic state should rationally exhibit this inertia is very counterintuitive: surely as you toss the coin and it lands heads repeatedly, your credence in HEADS25 ought to increase?

To put the point vividly, we can imagine the credence functions that represent your epistemic state as a group of avatars. The avatars at t_0 will assign various precise credences to HEADS25: for every number in the range $(0, 1)$, there will be some avatar who assigns that value to HEADS25. On learning HEADS1, each avatar ought to update accordingly by conditionalizing. Take an avatar who had a credence of 0.0001 in HEADS25.

4 A similar problem applies to Joyce's adjusted version of this principle mentioned earlier.

It may be[5] that this avatar's conditional credence in HEADS25 given HEADS1 is higher than her unconditional credence in HEADS25, in which case this avatar will increase her credence in HEADS25 on learning HEADS1. But there will be some avatar (perhaps an avatar whose unconditional credence in HEADS25 was even lower than 0.0001) whose credence in HEADS25 conditional on HEADS1 is 0.0001. Thus even after learning HEADS1, there will still be, in the set representing your epistemic state, an avatar whose credence in HEADS25 is 0.0001. Similarly, even if you learn the conjunction of the claims HEADS1 through HEADS24, there will still be an avatar in the set representing your epistemic state whose credence in HEADS1 is 0.0001. Thus your credence in HEADS25 will not shift from the range $(0, 1)$ no matter how much evidence you amass in favour of HEADS25.

This looks like a problem—at least for those imprecise probabilists who accept the chance grounding thesis, or something close to it. For some of the responses available, see R. Bradley (2017), Joyce (2010), Rinard (2013), and Vallinder (2018).

2.2 *Dilation*

Here we turn to another problem for the proponent of imprecise probabilism. The phenomenon I discuss here was first noted by early statisticians of imprecise probabilism Walley (1991) and Seidenfeld and Wasserman (1993), and has recently been prominently discussed by White (2009). Take some claim P, that you have no evidence whatsoever for or against, so that your credence at t_0 in P is the range $[0, 1]$. Suppose that I know whether P is true, and I take a fair coin and paint the heads side over. I write "P' on this heads side iff P is true, and "not P" on the heads side iff P is not true. I similarly paint over the tails side of the coin, and write on this side whichever claim (out of "P" and "not P") is false. You know that I have done this. I then toss the coin before your eyes. Your credence before it lands (i.e. at t_0) that it will land head-side up (HEADS), is 0.5. Then at t_1 you see it land, with the "P"-side up. What then at t_1 is your credence in P and what is your credence in HEADS?

At t_1 you have learnt that the coin has landed "P"-side up. Thus if P is true, then HEADS is also true (i.e. it must have landed heads)—for if P is true then "P" has been painted onto the heads side of the coin, and so given that it has landed "P"-side up it has also landed heads. Furthermore, if HEADS is true, then P is also true—for if it has landed heads then given that it has landed "P"-side up, "P" must have been painted onto the heads side of the coin, which will have happened only if P is true. Thus at t_1 you

5 Though it need not be: perhaps some avatars will stubbornly refuse to adjust their credence in HEADS25 from 0.0001. We might try to avoid this problem by excluding such agents (Halpern, 2003), though this will not solve the problem discussed in the main text.

can be certain that P is true iff HEADS is true. Thus at t_1 you must have the same credence in P as you have in HEADS. Given that at t_0 your credence in HEADS is 0.5, and your credence in P is the range $[0, 1]$, how will your credence adjust between t_0 and t_1? Will your credence in HEADS become the range $[0, 1]$? Or will your credence in P become precisely 0.5? Both options seem counterintuitive.[6] It seems implausible that your credence in HEADS should "dilate" to the range $[0, 1]$: surely (by the principal principle) your credence that a fair coin has landed heads ought to be 0.5, unless you have some evidence as to how it has landed. And knowing that it landed on the "P"-side does not seem to give you any evidence as to whether it has landed heads or tails. And it also seems implausible that your credence in P should sharpen to the number 0.5 (White, 2009), for after all you knew even at t_0 that the coin would either land "P"-side up, or "P"-side down, and we cannot say that learning either of these pieces of information would force your credence in P to become precisely 0.5 without violating van Fraassen's reflection principle (van Fraassen, 1984).

One popular response made by the imprecise probabilist, is to accept that at t_1 your credence in HEADS ought to dilate to $[0, 1]$.[7] Here are two things that might be said in defence of this position.

- It seems as though learning that the coin landing "P"-side up gives you no evidence as to whether it has landed head-side up. But this would not follow if P was a claim that you knew something about. Suppose as a contrast case, then, that P is the claim that you have just won the lottery—a claim in which you have a very low credence indeed. On hearing that I (who know the outcome) am painting the true claim (out of "P" and "not-P") on the heads side, and the false claim on the tails side, you will be almost certain that I am painting "not-P" on the heads side, and "P" on the tails side. Your credence at t_0 in HEADS is 0.5, but when at t_1 you learn that the coin has landed "P"-side up, you will be almost certain that HEADS is false. Thus where you have some evidence concerning P, it is natural to suppose that learning that the coin has landed "P"-side up will alter your credence in HEADS (see Sturgeon, 2010, Joyce, 2010).

  What about in the case where P is a claim about which you have no evidence? In this case, it is tempting to suppose that learning that the coin has landed "P"-side up gives you no reason to adjust your credence in HEADS. But the situation is more complicated than

6 A further option would be for both your credence in HEADS and your credence in P to adjust, but this is no more appealing than the alternatives.

7 As White acknowledges, some statisticians and philosophers (such as Walley, 1991, and Seidenfeld and Wasserman, 1993) had noted this result and "taken it in their stride" (White, 2009, p. 177).

this suggests. Consider again your epistemic state as a set of avatars. For every number in the range $[0, 1]$, there is some avatar in the set that represents your epistemic state that assigns this number to P. Each such avatar, on learning that the coin has landed "P"-side up, will adjust her credence in HEADS accordingly.[8] For example, the avatar whose credence in P is 0.2 will adjust her credence in HEADS downwards; and the avatar whose credence in P is 0.8 will adjust her credence in HEADS upwards. More generally after conditionalizing on the claim that the coin has landed "P"-side up, for every number in the range $[0, 1]$, there will be an avatar who assigns that number to HEADS. We can see then that it is not that learning that the coin has landed "P"-side up gives you no evidence relevant to HEADS, but rather that you are just very uncertain as to in what direction the evidence you have received should pull you, and how far. Thus your credence in HEADS is infected with the imprecision that you assigned to P, and your credence in HEADS dilates to the range $[0, 1]$ (Joyce, 2010).

- It is tempting to object that it is counterintuitive for an increase in evidence to leave your credence function more imprecise than it was before. However it is not obvious that your credence function is more imprecise at t_1 than it was at t_0. To see this, consider that at t_0 though your credence in HEADS was precise, your conditional credence in HEADS given that the coin lands "P"-side up was imprecise. Thus there was imprecision in your credence function even at t_1: this just was not obvious when we focused only on your unconditional credence in HEADS (R. Bradley, 2017).

Further discussion of the problem of dilation can be found in R. Bradley (2017), S. Bradley and Steele (2014), Dodd (2013), Joyce (2010) and Pederson and Wheeler (2014).

3 DECISION-MAKING

On the classic Bayesian picture, a rational agent has a precise credence function assigning some number between 0 and 1 to each proposition, and also a precise utility function assigning some number to each possible outcome representing in some sense how much the agent values each outcome. When faced with a decision problem—i.e. a choice between different actions—on the classic picture the agent must choose an action that has maximum *expected utility*. We can calculate the expected utility of any given action for the agent as follows: for every possible outcome,

8 Those avatars whose credence at t_0 in P is 0.5 need make no adjustment.

	Milk at home (s_1)	No milk at home (s_2)
	$Cr(s_1) = 0.5$	$Cr(s_2) = 0.5$
Stop for milk	9	9
Don't stop	10	5

Table 1: A decision problem

we multiply the agent's credence that the outcome will obtain should she perform the action under consideration, by the utility of that outcome—and then we sum the lot.[9]

Here is an example to illustrate this. Sometimes on the way home from work, I stop to buy a pint of milk, which means that I take a bit longer to get home, but it is certainly better than getting home and finding that there is no milk in the house. Suppose that on this occasion, my credence that there is milk in the house already is 0.5. Table 1 represents my assessment of the possible outcomes.

We can now calculate the expected utility of each available action. The expected utility of stopping to buy milk is $(0.5)(9) + (0.5)(9) = 9$, whereas the expected utility of not stopping to buy milk is $(0.5)(10) + (0.5)(5) = 7.5$. On the classic decision rule "maximise expected utility", I ought to stop to buy milk, because this is the action with the highest expected utility.

The maximise expected utility rule works on the assumption that for every relevant state of the world, the rational agent has a precise credence that that state of the world obtains. But proponents of imprecise probabilities deny this, and so cannot accept this rule. What alternative rule should they put in its place? According to the proponent of imprecise probabilities, what requirements does rationality place on an agent's choice of action? Many different answers have been proposed, and I will briefly outline two of these answers.

PERMISSIVE CHOICE RULES Recall that we can see an agent's epistemic state as represented by a set of avatars, each with a precise credence function. Thus faced with any decision problem, each avatar will have a view as to which action—or actions—will maximise expected utility.[10] According

9 This is a rough and ready sketch of Savage's account (Savage, 1954). Modifications have been made to that account (e.g. in Jeffrey, 1965) but here I will stick to straightforward examples so that the modifications should not be relevant.

10 Here I assume that the agent has a precise utility function which feeds into each avatar's calculation. This of course is also up for debate, and some argue that just as a rational agent can have an imprecise credence function, so (s)he can have an imprecise utility function. I do not discuss this further here however.

to the permissive choice rules,[11] the agent may rationally perform any action provided that at least one of her avatars recommends that action.

To illustrate this, suppose that an agent's credence that it will rain tomorrow is the range $(0.4, 0.8)$. Thus for every number in this range, there is some avatar who assigns that number to the claim that it will rain tomorrow. Suppose then that the agent is offered the following bet: she is to pay out £5, and will get £10 back iff it rains tomorrow. The agent has to choose whether to accept the bet, or reject it. We can assume that the agent values only money, and values it linearly. Some of her avatars would recommend accepting the bet (those whose credence that it will rain is greater than 0.5), some recommend rejecting it (those whose credence that it will rain is less than 0.5), and some rate the expected utility of accepting it equal to the accepted utility of rejecting it (those whose credence that it will rain is 0.5). Thus according to the permissive choice rules, the agent is free to either accept or reject the bet: both actions are permissible. This rule—together with some variations—is discussed under the name 'Caprice' by Weatherson (1998).

MAXIMIN The rule maximin works as follows. Where an agent has an imprecise probability function, we can see her epistemic state as represented by a set of precise functions, or avatars. When considering a possible action, there is an expected utility for that action relative to each precise probability function in the agent's set. Amongst these expected utilities for the action, one will be the lowest—and so each action has a minimum expected utility. According to maximin, when faced with a choice, a rational agent will carry out whichever action has the maximum minimum expected utility.

To illustrate this, take again our agent whose credence that it will rain tomorrow is the range $(0.4, 0.8)$: for every number in this range, there is some avatar who assigns that number to the claim that it will rain tomorrow. Suppose then that the agent is offered the following bet: she is to pay out £5, and will get £10 back iff it rains tomorrow. Each avatar calculates the expected utility of each possible action—i.e. the action of accepting the bet and the action of rejecting the bet. The avatar who assigns the lowest expected utility to accepting the bet is the avatar whose credence that it will rain tomorrow is 0.4: assuming again that the agent values only money and that linearly, we can represent the expected utility of accepting the bet from the perspective of this avatar as $-5 + (0.4)(10) = -1$. Thus the minimum expected utility of accepting the bet is -1. Now we can calculate the minimum utility of rejecting the bet. All avatars assign the same expected utility to this action—namely 0. Thus the minimum expected utility of rejecting the bet is 0. A rational agent will choose from amongst

11 This is Adam Elga's (2010) term.

those actions with the highest minimum expected utility—and as rejecting the bet has a higher minimum expected utility (0) than accepting the bet (-1), the agent if rational will reject the bet.

Variations on this rule have been developed by Gärdenfors and Sahlin (1982), Gilboa and Schmeidler (1989), and others. An analogous maximax rule has been developed by Satia and Lave (1973). Many further rules have been proposed, including those by Arrow and Hurwicz (1972) and Ellsberg (1961). See Troffaes (2007) for a discussion and comparison of some of these rules.

3.1 *Applying These Rules*

In some scenarios, some of the alternative rules developed by imprecise probabilists seem to work better than the classical Bayesian's rule maximise expected utility. Here is a famous case—the Ellsberg paradox—in which this holds (Ellsberg, 1961).

You have an urn before you, which contains 150 balls. 50 are black, and the other 100 are some mixture of red and yellow—but you have no further information as to what the proportions of red and yellow balls are. For all you know, there may be 100 red balls and no yellow balls, or 100 yellow balls and no red balls, or any mixture between these two extremes. Now a ball will shortly be selected at random from the urn, and you have the chance to bet on what colour the ball will be. You can either say 'black', in which case you'll win £100 if it is black, and nothing otherwise; or you can say 'red', in which case you'll win £100 if it is red, and nothing otherwise (Table 2).

	Black (B)	Red (R)	Yellow (Y)
Bet black	£100	£0	£0
Bet red	£0	£100	£0

Table 2: The first scenario in the Ellsberg paradox

Now suppose instead that you have the option of saying 'black or yellow', in which case you'll win £100 if the ball is either black or yellow, and nothing otherwise; or you can say 'red or yellow', in which case you'll win £100 if the ball is either red or yellow, and nothing otherwise (Table 3).

Typically people choose to say 'black' in the first scenario, but 'red or yellow' in the second. Furthermore, many apparently rational people exhibit this betting pattern.[12] The problem is that if we assume that a

12 See Voorhoeve, Binmore, Stefansson, and Stewart (2016) for an analysis and discussion of the prevalence of this betting pattern.

	Black (B)	Red (R)	Yellow (Y)
BET BLACK OR YELLOW	£100	£0	£100
BET RED OR YELLOW	£0	£100	£100

Table 3: The second scenario in the Ellsberg paradox

rational agent has precise probabilities and utilities, and chooses only between those actions that maximise expected utility, then a rational agent cannot exhibit this betting pattern. To see this, let's suppose that some agent who exhibits this betting pattern has precise probabilities, and is maximising expected utility. We let the agent's credence in B, R and Y be given by $Cr(B)$, $Cr(R)$ and $Cr(Y)$ respectively, and we let the utility of winning £100 be given by u_1 and the utility of winning £0 be given by u_2. Then—given that our agent chooses 'black' over 'red' in the first scenario, it follows that

$$Cr(B) \cdot u_1 + Cr(R) \cdot u_2 + Cr(Y) \cdot u_2 > Cr(B) \cdot u_2 + Cr(R) \cdot u_1 + Cr(Y) \cdot u_2,$$

and so that

$$Cr(B) \cdot u_1 + Cr(R) \cdot u_2 > Cr(B) \cdot u_2 + Cr(R) \cdot u_1.$$

But then the agent chooses 'red or yellow' over 'black or yellow' in the second scenario, and so it follows that

$$Cr(B) \cdot u_1 + Cr(R) \cdot u_2 + Cr(Y) \cdot u_1 < Cr(B) \cdot u_2 + Cr(R) \cdot u_1 + Cr(Y) \cdot u_1,$$

and so that

$$Cr(B) \cdot u_1 + Cr(R) \cdot u_2 < Cr(B) \cdot u_2 + Cr(R) \cdot u_1.$$

This contradicts our earlier result. Thus no agent exhibiting this betting pattern can have only precise probabilities and utilities and be guided by the rule maximise expected utility.

What alternative rule might be guiding the agent's behaviour in Ellsberg's scenario? Several of the rules formulated by proponents of imprecise probabilities can explain the agent's behaviour, and so Ellsberg's scenario can be used to argue both for (some of) the alternative rules, and for the claim that rational agents can have imprecise probabilities. To illustrate how some of these rules might handle Ellsberg's scenario, I will run through Ellsberg's own solution to the problem.

In Ellsberg's terminology, a situation can be "ambiguous" for an agent. In an ambiguous situation, more than one probability distribution seems reasonable to the agent. We can gather these probability distributions into a set $P = \{p_1, p_2, \ldots, p_n\}$: these are the distributions that the agent's

information "does not permit him to rule out" (Ellsberg, 1961, p. 661). The agent assigns weights to each of these reasonable distributions, and arrives at a composite "estimated" distribution p_i where p_i is a member of P. The *estimated pay-off* A_{est} of a given action A is the expected utility of the action calculated using p_i (Ellsberg, 1961, p. 661). But when faced with a choice of actions, the rational agent may be guided not just by the expected pay-off of each action, calculated in terms of p_i. The agent may also take into account the lowest expected utility of each action as calculated using any member of P. We let A_{min} denote the minimum expected utility of action A as calculated using any member of P, and we let x denote the agent's degree of confidence in p_i (the "estimated" distribution). Then the *index* of an action A is given by $x \cdot A_{est} + (1 - x) \cdot A_{min}$. Ellsberg's rule for action, then, is as follows: choose the action with the highest index.

In Ellsberg's scenario, the agent is in an ambiguous situation: the agent can be certain that the probability that a ball randomly drawn from the urn will be red is $1/3$, but the agent cannot be certain of the probability of the ball's being yellow or black, because (s)he does not know the proportion of yellow and black balls in the urn. There are a range of probability distributions that seem reasonable to the agent: for every number n between 0 and $2/3$, there is a reasonable probability distribution under which the probability of R is r, the probability of Y is $2/3 - r$, and the probability of B is $1/3$. Let us assume for simplicity that the agent assigns weight evenly across these reasonable probability distributions. Thus on the composite "estimated" distribution, the probability of R is $1/3$, the probability of Y is $1/3$, and the probability of B is $1/3$. Thus the expected payoff of saying 'black' in the first scenario ($1/3 \cdot u_1 + 2/3 \cdot u_2$) is the same as the expected payoff of saying 'red' in that scenario, and the expected payoff of saying 'black or yellow' in the second scenario ($2/3 \cdot u_1 + 1/3 \cdot u_2$) is the same as the expected payoff of saying 'red or yellow' in that scenario.

However a rational agent need not be guided merely by the estimated payoff of each action, but also by the lowest expected utility of each action. For the action of saying 'red' in the first scenario, the lowest expected utility is that given by the probability distribution according to which the probability of R is 0, the probability of Y is $2/3$, and the probability of B is $1/3$: according to this distribution, the expected utility of saying 'red' is 0. In contrast, according to every distribution the expected utility of saying 'black' is $1/3$, and so of course the lowest expected utility of saying 'black' is $1/3$. The 'index' of some action A is given by $x \cdot A_{est} + (1 - x) \cdot A_{min}$, where x is the agent's level of confidence in the 'estimated distribution'. Thus the index of saying 'red' is $1/3 \cdot x + 0 \cdot (1 - x)$, and the index of saying 'black' is $1/3 \cdot x + 1/3 \cdot (1 - x)$. Thus whenever the agent is less than perfectly confident in the estimated distribution—which a rational agent may well be—the value x will be less than 1, and the index of saying 'black' will be

greater than the index of saying 'red'. Thus any agent for whom x is less than 1 will say 'black' rather than 'red' in the first scenario. In the second scenario, however, the very same agents will choose to say 'red or yellow' rather than 'black or yellow'. For it works out that the expected payoff of both of these actions is $2/3$, but the lowest expected utility of saying 'black or yellow' ($1/3$) is lower than the lowest expected utility of saying 'red and yellow' ($2/3$), and so saying 'black or yellow' has a lower index than saying 'red or yellow'.

In short, an agent for whom x is less than 1 is *ambiguity averse*: all else being equal, the agent prefers actions where (s)he knows the chances of the relevant outcomes over actions where (s)he merely estimates those outcomes. In the first scenario, if the agent says 'black' then (s)he will know the chance of winning £100, whereas if she says 'red' then the chance of winning will be unknown. In contrast, in the second scenario, if the agent says 'red or yellow' then (s)he will know the chance of winning £100, whereas if she says 'black or yellow', the chance of winning will be unknown. Thus the betting pattern that is typically displayed in Ellsberg's scenario is permissible.

Here the imprecise probabilist seems to have an advantage over the precise probabilist. The precise probabilist seems forced to claim—counterintuitively—that the typical betting pattern in Ellsberg's scenario is irrational, whereas the imprecise probabilist can account for this betting pattern well.

I turn now to the problem of sequential decision problems, which seem to pose a problem for the imprecise probabilist.

3.2 *Sequential Decision Problems*

Here is a problem posed by Adam Elga (2010).[13] According to the imprecise probabilist, a rational agent may have a credence of, say, $[0.1, 0.8]$ in some claim H. Now consider the following two bets:

> Bet A: If H is true, then you lose £10; otherwise you win £15.
>
> Bet B: If H is false, then you lose £10; otherwise you win £15.

These bets are offered sequentially: first Bet A is offered to the agent, and then Bet B. The agent knows that she will be offered both bets, and has the option of taking both, rejecting both, or taking either one. Intuitively, it would be irrational for an agent to reject both bets, because rejecting both bets leaves the agent with nothing, whereas accepting both bets leaves the

13 The problems that the imprecise probabilist faces over sequential decision problems are widely discussed in the literature from economics, and a puzzle related to Elga's can be found in Hammond (1988).

agent with a sure £5. Surely then a rational agent would not reject both? The challenge that Elga poses to the imprecise probabilist is to put forward a plausible decision rule that entails that a rational agent in this scenario will not reject both bets. Various attempts have been made to meet this challenge.

It seems at first as though the permissive choice rules will not do. To see why, consider that if the agent is presented with just Bet A, there will be avatars who recommend rejection, so it follows that the agent is rationally permitted to reject Bet A. But then when presented with Bet B, there will similarly be avatars (different avatars) who recommend that this bet is rejected. So it follows that the agent is rationally permitted to reject Bet B. Thus it seems that the permissive choice rules would permit the agent to reject both bets, and so this rule cannot be used to meet Elga's challenge. However defenders of this rule may claim either that a sequence of actions is permitted only when that sequence is recommended by a single avatar, or else challenge Elga on his assumption that accepting each bet is a separate action, rather than parts of a single action (Weatherson, 2003; Williams, 2014).

Similarly, it may seem that maximin, Ellsberg's rule, and others will be unable to handle Elga's scenario, for many of these rules would permit a rational agent to reject both bets if offered on separate occasions. However as several authors have pointed out, and as A. Elga (2012) acknowledges, once we call on the resources of game theory, we find that several of these rules do entail that a rational agent in Elga's scenario (in which the agent knows that (s)he will be offered both bets) will not reject both bets. See S. Bradley and Steele (2014), Chandler (2014), and Sahlin and Weirich (2014); see Mahtani (2018) for a response.

A further way of responding to Elga's challenge is to argue that when faced with a series of choices, a rational agent will make a plan and stick to it—and where an agent has an imprecise credence function, that plan will be endorsed as maximising expected utility by at least one of the agent's avatars. For further discussion of this sort of view, see Bratman (2012), Gauthier (1986), and McClenen (1990).

Finally, there are authors who reject the assumption that an agent in an Elga-style scenario who rejects both bets is thereby irrational. For example, Moss (2015) constructs an account of what it is for an agent with imprecise credences to "change his or her mind", and argues that it is permissible in at least some Elga-style scenarios for an agent to reject Bet A while identifying with one of her avatars, and then change her mind and reject Bet B, identifying with a different avatar. Others such as S. Bradley and Steele (2014) also maintain that a rational agent in an Elga-style scenario may reject both bets.

Thus there are a range of interesting ways that the imprecise probabilist might respond to the sort of sequential decision problem that Elga has raised, and the debate over which rule of rationality the imprecise probabilist should endorse is still ongoing.

4 SUMMARY

I began with a natural motivation for accepting imprecise probabilism. I then outlined the most widely discussed account of imprecise probabilities, and considered how the account should be interpreted. I then turned to two categories of objections to the account: objections concerning learning, and objections concerning decision making. Within learning, I discussed two different objections: firstly the problem of belief inertia, and secondly the problem of dilation. Within decision making, I focused on the problems that the imprecise probabilist faces in situations of sequential choice. There has been recent, lively debate about these objections, and while various responses have been put forward by the imprecise probabilists, we are currently far from a consensus.

REFERENCES

Arrow, K. J. & Hurwicz, L. (1972). An optimality criterion for decision making under ignorance. In C. F. Carter & J. L. Ford (Eds.), *Uncertainty and expectations in economics: Essays in honour of g.l.s. shackle*. Oxford: Basil Blackwell.

Bradley, R. (2009). Revising incomplete attitudes. *Synthese*, *171*(2), 235–256.

Bradley, R. (2017). *Decision theory with a human face*. Cambridge University Press.

Bradley, S. & Steele, K. (2014). Should subjective probabilities be sharp? *Episteme*, *11*(3), 277–289.

Bratman, M. (2012). Time, rationality and self-governance. *Philosophical Issues*, *22*(1), 73–88.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3.

Chandler, J. (2014). Subjective probabilities need not be sharp. *Erkenntnis*, *79*(6), 1273–1286.

Dempster, A. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, *38*(2), 325–39.

Dempster, A. (1968). A generalization of bayesian inference. *Journal of the Royal Statistical Society*, *30*(2), 205–247.

Dodd, D. (2013). Roger white's argument against imprecise credences. *The British Journal for the Philosophy of Science*, *64*(1), 69–77.

Elga, A. [A.]. (2012). Errata for subjective probabilities should be sharp.
Elga, A. [Adam]. (2010). Subjective probabilities should be sharp. *Philosophers' Imprint*, *10*.
Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics*, *75*(4), 643–669.
Gärdenfors, P. & Sahlin, N. E. (1982). Unreliable probabilities, risk taking and decision making. *Synthese*, *53*(3), 361–386.
Gauthier, D. (1986). *Morals by agreement*. Oxford: Clarendon Press.
Gilboa, I. & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, *18*(2), 141–153.
Greaves, H. & Wallace, D. (2006). Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind*, *115*(459), 607–632.
Halpern, J. Y. (2003). *Reasoning about uncertainty*. Cambridge: MIT Press.
Hammond, P. (1988). Orderly decision theory. *Economics and Philosophy*, *4*(2), 292–297.
Jeffrey, R. (1965). *The logic of decision*. Chicago: University of Chicago Press.
Jeffrey, R. (1983). Bayesianism with a human face. In J. Earman (Ed.), *Testing scientific theories* (pp. 133–156). University of Minnesota Press.
Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, *65*(4), 575–603.
Joyce, J. M. (2005). How probabilities reflect evidence. *Philosophical Perspectives*, *19*, 153–178.
Joyce, J. M. (2010). A defense of imprecise credences in inference and decision making. *Philosophical Perspectives*, *24*(1), 281–323.
Kaplan, M. (1996). *Decision theory as philosophy*. Cambridge University Press.
Keynes, J. M. (1921). *Treatise on probability*. London: Macmillan.
Kyburg, H. (1983). *Epistemology and inference*. Minneapolis: University of Minnesota Press.
Levi, I. (1974). On indeterminate probabilities. *Journal of Philosophy*, *71*(13), 391–418.
Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability* (pp. 83–132). University of California Press.
Mahtani, A. (2018). Imprecise probabilities and unstable betting behaviour. *Noûs*, *52*(1), 69–87.
Mayo-Wilson, C. & Wheeler, G. (2016). Scoring imprecise credences: A mildly immodest proposal. *Philosophy and Phenomenological Research*, *92*(1), 55–78.
McClenen, F. (1990). *Rationality and dynamic choice: Foundational explorations*. Cambridge University Press.
Moss, S. (2015). Credal dilemmas. *Noûs*, *49*(4), 665–683.

Pederson, A. & Wheeler, G. (2014). Demystifying dilation. *Erkenntnis*, *79*(6), 1305–1342.

Pettigrew, R. (2013). A new epistemic utility argument for the principal principle. *Episteme*, *10*(1), 19–35.

Rinard, S. (2013). Against radical credal imprecision. *Thought*, *2*(1), 157–165.

Rinard, S. (2015). A decision theory for imprecise probabilities. *Philosophers' Imprint*, *15*(7), 1–16.

Sahlin, N. E. & Weirich, P. (2014). Unsharp sharpness. *Theoria*, *80*(1), 100–103.

Satia, J. & Lave, R. (1973). Markovian decision processes with uncertain transition. *Operations Research*, *21*(3), 728–740.

Savage, L. (1954). *The foundations of statistics*. Wiley.

Schoenfield, M. (2017). The accuracy and rationality of imprecise credences. *Noûs*, *51*(4), 667–685.

Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (2012). Forecasting with imprecise probabilities. *International Journal of Approximate Reasoning*, *53*(8), 1248–1261.

Seidenfeld, T. & Wasserman, L. (1993). Dilation for sets of probabilities. *The Annals of Statistics*, *21*(3), 1139–1154.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.

Sturgeon, S. (2008). Reason and the grain of belief. *Noûs*, *42*(1), 139–165.

Sturgeon, S. (2010). Confidence and coarse-grained attitudes. In T. S. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology*. Oxford: OUP.

Troffaes, M. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, *45*, 17–29.

Vallinder, A. (2018). Imprecise bayesianism and global belief inertia. *The British Journal for the Philosophy of Science*, *69*(4), 1205–1230.

van Fraassen, B. C. (1984). Belief and the will. *Journal of Philosophy*, *81*(5), 235–256.

van Fraassen, B. C. (1990). Figures in a probability landscape. In M. Dunn & K. Segerberg (Eds.), *Truth or consequence* (pp. 345–56). Amsterdam: Kluwer.

van Fraassen, B. C. (2006). Vague expectation value loss. *Philosophical Studies*, *127*(3), 483–491.

Voorhoeve, A., Binmore, K., Stefansson, A., & Stewart, L. (2016). Ambiguity attitudes, framing, and consistency. *Theory and Decision*, *81*(3), 313–337.

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman & Hall.

Weatherson, B. (1998). Decision making with imprecise probabilities. Retrieved April 5, 2019, from http://brian.weatherson.org/vdt.pdf

Weatherson, B. (2003). From classical to intuitionistic probability. *Notre Dame Journal of Formal Logic*, *44*(2), 111–123.

White, R. (2009). Evidential symmetry and mushy credence. In T. S. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (pp. 161–186). Oxford University Press.

Williams, J. R. G. (2014). Decision-making under indeterminacy. *Philosophers' Imprint*, *14*(4), 1–34.

Yager, R. & Liu, L. (Eds.). (2008). *Classic works of the dempster-shafer theory of belief functions*. Studies in Fuzziness and Soft Computing. Springer.

4

CONDITIONAL PROBABILITIES

Kenny Easwaran

Conditional probability is one of the central concepts in probability theory. Some notion of conditional probability is part of every interpretation of probability. The basic mathematical fact about conditional probability is that $p(A \mid B) = p(A \wedge B)/p(B)$ where this is defined. However, while it has been typical to take this as a definition or analysis of conditional probability, some (perhaps most prominently Hájek, 2003) have argued that conditional probability should instead be taken as the primitive notion, so that this formula is at best coextensive, and at worst sometimes gets it wrong.

Section 1.1 considers the concept of conditional probability in each of the major families of interpretation of probability. Section 1.2 considers a conceptual argument for the claim that conditional probability is prior to unconditional probability, while Section 1.3 considers a family of mathematical arguments for this claim, leading to consideration specifically of the question of how to understand probability conditional on events of probability 0. Section 1.4 discusses several mathematical principles that have been alleged to be important for understanding how probability 0 behaves, and raises a dilemma for probability conditional on events of probability 0. Section 2 and Section 3 take the two horns of this dilemma and describe the two main competing families of mathematical accounts of conditional probability for events of probability 0. Section 4 summarizes the results, and their significance for the two arguments that conditional probability is prior to unconditional probability.

1 BACKGROUND

1.1 *What is Conditional Probability?*

Before considering the arguments suggesting that conditional probability is a primitive notion (either equal to unconditional probability in fundamentality, or perhaps even more basic), we should consider just what conditional probability is.

Some have argued, following some cryptic remarks of Frank Ramsey, that conditional probability can be understood as the probability of a conditional. However, without a clear interpretation of what a conditional means, this provides little help for clarifying the concept of conditional

probability. There are deep difficulties with this identification, since together with certain plausible logical principles for conditionals, it entails various triviality results about unconditional probability. (Edgington, 1995, summarizes much of this literature and argues that there is some interpretation of the conditional that allows for this identification, and Bacon, 2015, shows how much logic for conditionals can be preserved.) At any rate, the defenders of this principle hope to use conditional probability to clarify the meaning of conditionals, rather than vice versa. Since the meaning of a conditional has so much obscurity, this identification is of no help in trying to analyze the meaning of conditional probability.

Perhaps a more useful (and also Ramsey-inspired) way to think of conditional probability is to look at some of the roles it plays in order to see what features it needs to have. But since there are many different phenomena that have all been said to be interpretations of probability, and conditional probability plays different roles in each, I will break this consideration down into several parts. In this discussion, I will not consider each separate interpretation of probability, but I will instead consider them in three broad families. (For more on specific interpretations, see Hájek, 2007.)

The first family (which I will use as my primary reference point in much later discussion) is the set of broadly "Bayesian" interpretations that treat probability as some sort of informational state. The second family is the set of broadly "physical" interpretations that treat probability as a feature of some part of the world itself, rather than an information state. The third family is the set of "mathematical" applications of probability, some of which I don't think rise to the level of an interpretation, but are worth mentioning separately.

1.1.1 *Bayesian Interpretations*

Among the interpretations I am calling "Bayesian" are both various objective and subjective notions. I mean this class to include "logical probabilities" (Keynes, 1921; Carnap, 1950; Maher, 2006) and "evidential probabilities" (Williamson, 2002), as well as the more familiar objective and subjective Bayesian interpretations of probability as some sort of rational degree of belief (Easwaran, 2011a, 2011b). These interpretations of probability are used in a broad variety of applications in psychology, economics, decision theory, philosophy of science, and epistemology.

However, in all of these applications, it seems that there are three main roles that conditional probability is said to play. First, conditional probability is said to play some sort of fairly direct role in constraining the way that probabilities change over time. Second, conditional probability is used in the analysis of various measures of confirmation (which often

claim to describe the potential value of various pieces of information, whether or not anyone ever gains that information). And third, conditional probability is important in certain accounts of decision theory. If there are roles for conditional probability other than these, then some of my later evaluation of the different mathematical accounts of conditional probability may need to be modified.

The role of conditional probability in updating is perhaps the most familiar one. The traditional notion of Bayesian updating is said to occur when there is some new evidence E that the agent gains with certainty. In this case, the probability function after the update p_{new} and the probability function before the update p_{old} are said to satisfy, for every A, $p_{\text{new}}(A) = p_{\text{old}}(A \mid E)$. Following Jeffrey (1965), many have thought that this sort of update scenario is implausible, because there is never any particular evidence that is gained with certainty. Instead, there is said to be an evidential partition $\mathbf{E}$, which is a set of propositions $\{E_i : i \in I\}$, such that it is antecedently certain that there is exactly one i such that E_i is true. No member of this partition becomes certain, but their probabilities change in a way that drives the change of all other propositions. This notion of "driving the change" is summarized by a constraint known as *rigidity*: for any A, $p_{\text{new}}(A \mid E_i) = p_{\text{old}}(A \mid E_i)$. The specification of these conditional probabilities is said to be enough, in conjunction with the new probabilities for each E_i, to specify the new probability function uniquely, by means of the Law of Total Probability. When the partition is finite, this takes the form $p(A) = \sum p(E_i)p(A \mid E_i)$, though in the infinite case we need to be a bit more careful. As I will discuss in Section 1.4.3, the natural way to generalize this will be notated as $p(A) = \int p(A \mid \mathbf{E_i})\, \mathrm{d}p$, though further complexities arise.

At least since the work of Hosiasson-Lindenbaum (1940), conditional probability has also been very important in analyzing the notion of confirmation. Much of this literature has focused on finding numerical measures of the degree to which particular evidence would support particular hypotheses. Where H is some hypothesis, and E is some potential evidence, some well-known measures are said to take the value $p(H \mid E) - p(H)$, or $p(H \mid E)/p(H)$, or $p(E \mid H)/p(E \mid \neg H)$. (These and other measures are discussed by Fitelson, 1999.) The probabilities that show up in these formulations are of four types. There are two unconditional probabilities, $p(E)$ and $p(H)$, which are called "priors" for the evidence and the hypothesis respectively. (Priors for their negations sometimes appear as well, but since $p(\neg E) = 1 - p(E)$ and $p(\neg H) = 1 - p(H)$ these are not relevantly different.) There are also two types of conditional probability that arise. $p(H \mid E)$ is called the "posterior" of the hypothesis, because (according to the update rule mentioned above), it gives the probability the hypothesis would have after hypothetically learning the evidence. And $p(E \mid H)$ and

$p(E \mid \neg H)$ are called "likelihoods" of the hypothesis and its negation. Some philosophers have focused on measures involving only likelihoods, because they are said to be more objective than priors and posteriors (Royall, 1997). But at any rate, these are the conditional probabilities whose values are relevant to confirmation.

In decision theory, the most traditional analysis of the value of an action doesn't depend on conditional probability at all (Savage, 1954). There are said to be a set $\mathbf{A}$ of actions available to the agent and a set $\mathbf{S}$ of possible states of the world independent of the agent, and together these are said to determine outcomes of the act. The agent has a value $V(A \wedge S)$ for each outcome. When everything is finite, the value of an act $A \in \mathbf{A}$ is given by $V(A) = \sum_{S \in \mathbf{S}} p(S)V(A \wedge S)$. (Again, when $\mathbf{S}$ is infinite, things are more complicated, as will be discussed in Section 1.4.2.) However, Jeffrey (1965) and others have worried about cases in which one can't identify states of the world independent of the agent. In this case, Jeffrey suggests that we should have $V(A) = \sum_{S \in \mathbf{S}} p(S \mid A)V(A \wedge S)$, replacing the unconditional probability of a state with its probability conditional on each action. Joyce (1999) and other "causal decision theorists" have argued that this "evidential decision theory" is wrong for certain cases, and replace the conditional probability $p(S \mid A)$ with something like $p(A \mathrel{\square\!\!\rightarrow} S)$, the probability of the subjunctive conditional. Regardless of how this is to be interpreted, the relevant conditional probabilities for decision theory are what I will call "action probabilities," and they must be defined for states of the world conditional on the possible acts of an agent.

Thus, on the Bayesian interpretations of probability, the conditional probabilities that arise in any relevant application appear to be of three forms—posteriors, likelihoods, and action probabilities. Posteriors must be defined for every hypothesis conditional on every piece of possible evidence (for confirmation theory), or for *every* proposition conditional on every piece of possible evidence (for updating). Likelihoods must be defined for every piece of possible evidence conditional on every hypothesis. And action probabilities must be defined for every state of the world conditional on every possible action. (On Jeffrey's interpretation, action probabilities may just be a special case of posteriors, since the role of an act for him is in some sense as a special piece of evidence, but for Joyce and others the role is somewhat different, though it may not even be a conditional probability in the traditional sense.) In each case, the set of things that may be conditioned on form a "partition"—they are a set of propositions such that it is certain in advance that exactly one of them is true. This fact will be significant for later discussion.

1.1.2 *Physical Interpretations*

Another family of interpretations of probability take probability to be something separate from any sort of information state. One historically influential such interpretation is Popper's account of chance as a sort of "propensity" of the world to evolve in a certain way (Popper, 1959b). Many statisticians have wanted some sort of objective physical notion of probability like this, but without the metaphysical baggage. This has given rise to frequentist statistical practice, described for instance by Mayo and Cox (2006), on which the relevant probabilities are the proportion of cases in which particular outcomes "would arise in a hypothetical long-run of repeated sampling" (p. 79).

These interpretations are possibly more heterogeneous than the Bayesian ones I discussed above, but we can still identify particular families of uses to which conditional probabilities are put. First, conditional probabilities are sometimes said to govern the way in which chances change over time. Second, conditional probabilities are sometimes used to analyze notions of causation or independence. Third, there are various uses conditional probabilities are put to in frequentist statistical practice. And fourth, there may be a relevant notion of expected value computed from physical probabilities.

For changing chances, David Lewis claims that "a later chance distribution comes from an earlier one by conditionalizing on the complete history of the interval in between" (1980, p. 280). That is, if p_{old} is the probability function giving the chances at some earlier time and p_{new} gives the chances at a later time, and H is the history of all events that occur between these two times, then for any A, $p_{\text{new}}(A) = p_{\text{old}}(A \mid H)$. This requires a notion of probability conditional on any $H \in \mathbf{H}$, where $\mathbf{H}$ is the set of all histories that could transpire between one time and another.

Some analyses of causation have said that A is a cause of B iff $p(B \mid A) > p(B)$, where p is some physical notion of probability. There are many obvious problems with this account, turning on cases where there are common causes (the probability of a parent having blond hair given that a child has blond hair is higher than the unconditional probability of a parent having blond hair, even though the child's hair color is not a cause of the parent's), other events intervening (the probability of getting in a car crash given that you've had a drink may be lower than the unconditional probability of getting in a car crash, if drinking makes you less likely to drive, even though drinking does tend to cause car crashes), and similar sorts of problems. Sophisticated versions of this theory have now turned to the sort of "causal modeling" developed by Pearl (2000) and Spirtes, Glymour, and Scheines (2000). On this picture, events A and B are taken to be particular values of variables $\mathbf{A}$ and $\mathbf{B}$, which may have two values

(A occurs or does not occur) or more (if A is seen as one of a class of ways for something to happen). These variables are represented by nodes in a graph with arrows connecting some nodes to others. Physical probabilities are given by a probability distribution for the values of one variable conditional on each specification of the values of the variables with arrows pointing directly to it. There are then *two* notions of conditional probability, depending on whether we "intervene" on one variable or merely "condition" on it (Meek & Glymour, 1994). This difference can be seen by considering the probability of someone having a sun tan given that their vitamin D levels are high—conditioning involves looking at people with *currently* high levels of vitamin D and measuring their tan, while intervening involves artificially *giving* people high levels of vitamin D and measuring their tan. Variable **A** is then said to be a cause of **B** iff intervening on **A** in different ways gives different conditional distributions for **B**, and is said to be independent if the conditional probabilities are the same. (Vitamin D likely turns out not to be a cause of sun tan, but to have correlation due to common cause.) Again, the relevant probabilities always involve conditioning on the elements of a partition. For far more on this, see Hitchcock (2010).

In frequentist statistical practice, there are a variety of conditional probabilities that arise. One of the most well-known such conditional probabilities is the p-value of a piece of evidence. This is the frequency with which evidence *at least as extreme* as the observed value would occur in hypothetical repetitions of the same experimental protocol, assuming that the "null hypothesis" is correct. We might notate this as $p(E^+ \mid H_0)$, where E^+ is the event of evidence at least as extreme being observed, and H_0 is the null hypothesis (though see Section 1.2 for discussion of whether this should really be thought of as a conditional probability). The p-value is often used as a criterion for statistical rejection, and it is common to reject the null hypothesis (in favor of some alternative) if the p-value falls below some pre-arranged threshold. The "power" of a statistical test is said to be the frequency with which the same experimental protocol would result in rejection of the null hypothesis, assuming that the alternative is in fact true. We might think of this as $p(R \mid H')$, where H' is the alternative to the null hypothesis, and R is the event of an experimental result that our protocol recommends rejection on. In statistical tests for which we want to estimate the value of some unknown parameter, our experimental protocol often ends not with rejection, but with specification of a "confidence interval." For instance, a 95% confidence interval is the set of parameter values for which the p-value would be at least .05 if that value were treated as the null—we can think of the confidence interval as the set of values that wouldn't be rejected at a given p-level. These probabilities are not the same as the likelihoods discussed above for Bayesian probabilities (because these

are not probabilities of the actually observed evidence, but rather of the event "an observation at least as extreme would occur"), but they are still probabilities conditional on each hypothesis.

Finally, although many contemporary decision theorists follow Savage (1954) in using some sort of Bayesian probability as the basis of computation of expected value, von Neumann and Morgenstern (1947) use a physical probability as their basis for a theory of rational decisions. Similar issues involving objective correlations between "states of the world" and an agent's actions might motivate some use of conditional probability in calculations of expected value, and these will be like the "action probabilities" I mentioned above.

Again, in all cases, the set of things that can be conditioned on forms a partition.

1.1.3 *Mathematical Interpretations*

There are some other interpretations of probability that don't quite fit in with those mentioned above. The most interesting such interpretation is that of probability as actual relative frequency. For instance, the World Health Organization reports that 68% of deaths worldwide in 2012 were due to non-communicable diseases, such as cancer, diabetes, and cardiovascular diseases. We can interpret this as a sort of probability, and say that the probability that a person who died in 2012 died of a non-communicable disease is .68. On this interpretation, for any descriptions A, B, we can say that $p(B \mid A)$ is the fraction of things fitting description A that also fit description B. Any description whatsoever can be used in either position, provided that there is a meaningful way to count instances of each.

This bears much similarity to the "classical interpretation" of probability attributed by Hájek (2007) to early probability theorists. The idea again is that in many traditional games of chance, physical probabilities or Bayesian probabilities may be usefully approximated by counting all the different possible outcomes of the game and seeing how many of them are of the sort of interest.

Tools like this have also been applied in pure mathematics, in what is called the "probabilistic method." This method was introduced by Erdős (1947) to derive bounds for Ramsey numbers. (These numbers were first investigated by Ramsey, 1930, in an attempt to work on the decision problem for logic, but have since been generalized to the size of any sort of structure that is needed to guarantee the existence of subsets with given complexity.) Erdős considers the complete graph on n vertices where edges are arbitrarily colored in two colors. He then defines a probability function on subsets of this graph, and shows that if n is large enough, then the probability of selecting k vertices at random such that all edges between

them are the same color is non-zero. In particular, this means that for any coloring of the graph on n vertices, there must be k vertices whose edges are all the same color. The importance of Erdős' result is that the bound he arrived at for n is substantially smaller than that arrived at by Ramsey, and is in most cases still the best known. This method has since been deployed in many other problems in combinatorics.

The classic applications of this method don't make any use of conditional probability. More advanced applications might, but in general, the interpretation of the probability function is not really of any interest. Instead, the probabilities (and perhaps any conditional probabilities) are just tools for mathematical computation. Any mathematical account of "conditional probability" could be useful, whether or not it has any application to other interpretations of probability. Thus, this interpretation of probability gives no particular constraint to our theorizing about conditional probability, and if anything, encourages us to explore as many different mathematical accounts as possible, in case one is of use in some mathematical problem or other.

1.2 *Backgrounds vs. Conditions*

There are two main families of argument that all probabilities must really be conditional. One family of argument (considered in this section) is conceptual, and claims that for many different interpretations, some sort of background is essential to even determine probabilities. The second family of argument (considered in Section 1.3) is mathematical, and uses problems involving division by zero to argue that conditional probability must be prior to unconditional probability. Although the mathematical arguments are sometimes clearer and seem more convincing, I will consider the conceptual arguments first, since the mathematical arguments lead more naturally to the issues that arise in the rest of this article. This section is independent of the rest of the article, and can be skipped by readers more interested in the mathematical issues.

This section considers the claim that a background is essential to the possibility of probability. I will consider versions of this argument for each interpretation of probability, and argue that for most interpretations of probability, this "background" is different enough in kind from the sort of thing that one can conditionalize on, that it should be treated separately from conditional probability. I claim that only for probabilities thought of as actual frequencies is it correct to say that every probability requires a background, and that this background makes every probability essentially a conditional probability. For some of the other interpretations, we will at least find that many numbers traditionally thought of as unconditional probabilities may be better thought of as conditional probabilities, but for

all of these other interpretations there is conceptual room to argue that some probabilities really are unconditional.

For this argument, again it will be useful to consider different interpretations of probability in some detail. However, I will skip a few of the most purely mathematical interpretations for which there are no important conceptual requirements, and will consider the other interpretations in somewhat different connections than I did before.

1.2.1 *Degree of Belief*

For subjective degree of belief, some have argued that all probabilities are really conditional on a background. I will argue that the role of the background is different from the role of the conditioning event in conditional probability. De Finetti (1974) says "every evaluation of probability is conditional; not only on the mentality or psychology of the individual involved, at the time in question, but also, and especially, on the state of information in which he finds himself at that moment" (p. 134). That is, rather than representing a subject S's degree of belief at t in a proposition A as $p(A)$, many authors suggest that it should be represented as $p(A \mid K_{S,t})$, where $K_{S,t}$ is the conjunction of all the propositions that S knows at t.

However, if it is possible (or reasonable, or rational) for different subjects with the same knowledge to have different degrees of belief, then including the knowledge as a proposition in an expression of conditional probability doesn't address the fundamental issue. There would not be one single probability function such that conditionalizing it on the knowledge that each subject has at each time yields the degrees of belief that agent does or should have. While the "information" may be a proposition of the same sort as the bearers of probability, the "mentality or psychology of the individual" is not.

Thus, unless we assume that the knowledge an agent has uniquely determines the probabilities that are rationally permitted for her (a thesis known as *Uniqueness*, contrasted with its negation, *Permissivism*; see Kopec and Titelbaum, 2016), it seems more accurate to represent a subject S's degrees of belief at a time t as $p_{S,t}(A)$. There is a separate Bayesian probability function for each subject at each time. This probability function will reflect an agent's knowledge, which may mean that it gives probability 1 to any proposition that is known. If this is the right way to treat knowledge, then $p_{S,t}(A) = p_{S,t}(A \mid K_{S,t})$. But the conditional probability is no more fundamental here.

However, some philosophers, such as Horowitz and Dogramaci (2016), argue that the knowledge or evidence that one has does uniquely determine the rational degrees of belief to have. On this picture, the degrees of belief that are rational for a subject at a time really do turn out to be a

matter of conditional probability, $p_{\text{rational}}(A \mid K_{S,t})$. What the Subjectivist Bayesians think of as a subject-and-time-relative unconditional probability is actually aimed at following an objective conditional probability function. However, even on this interpretation, there is an important theoretical consideration of what the rational degrees of belief would be for an agent with no knowledge whatsoever. The defender of the claim that conditional probabilities are fundamental would represent this as $p_{\text{rational}}(A \mid T)$, where T is some tautology, but it seems just as reasonable to represent this as $p_{\text{rational}}(A)$, so that there are some unconditional probabilities after all. The question then becomes: do the unconditional rational probabilities suffice to determine all the conditional rational probabilities? But this is largely a mathematical question, and not a conceptual one, and this is the fundamental question behind Section 1.3 and Section 1.4, with full theories described in Section 2 and Section 3.

I should also note that there is a view like this one available for a more permissive or subjectivist viewpoint. This viewpoint is associated with the work of Isaac Levi (1980). There is no one objectively rational evidential probability function. Instead, there are just many different "confirmational commitments" that one might have. When this confirmational commitment is conditionalized on the knowledge a subject has, we can find the degrees of belief that the subject is committed to. Thus, what I referred to above as $p_{S,t}(A)$ would instead be referred to as $p_C(A \mid K_{S,t})$, where C is the particular confirmational commitment the agent has. A major advantage this view has, if correct, is that it allows us to extend Bayesian updating to cases in which one revises one's beliefs by giving up something that was taken as evidence, by removing this proposition from one's knowledge. However, this view also requires such hypothetical revisions to yield well-defined commitments for giving up *any* of one's beliefs. And again, there may still be unconditional probabilities on this view (namely, the commitments one has prior to any evidence), though there is still a mathematical question of whether they suffice to determine the conditional probabilities that we usually focus on.

1.2.2 *Chance and Frequentism*

Some have argued that for the chance or frequency interpretation of probability, the role of experimental setup or preconditions for repeatability mean that all chance is conditional. I will again argue that the role of the background here is distinct from the role of the conditioning event in conditional probability, so that these interpretations also have no conceptual reason for making conditional probability prior to unconditional.

On one picture, chances are relative to a world and a time (Lewis, 1980). Thus, the chance of A at a time t in world w is fundamentally given by

$p_{w,t}(A)$. Chances may update by conditionalization, so that if t' is later than t, then $p_{w,t'}(A) = p_{w,t}(A \mid H_{t,t'})$, where $H_{t,t'}$ is the description of the complete history of the world from t to t'. If there is some earliest time 0, then one may even be able to say that $p_{w,t}(A) = p_{w,0}(A \mid H_{0,t})$, so that the chances at all later times are fundamentally given by the conditional chances at the beginning of time. But this still leaves unconditional chances at the earliest time. And if there is no earliest time, then it seems that we must allow unconditional chances at *every* time to count as equally fundamental, because there is no privileged earlier reference point from which they are all conditionalized. And on any of these pictures, the world must enter in as a separate background parameter distinct from the things conditionalized on. The history up to t alone does not suffice to determine the chances at t. (Just consider the following two worlds where nothing happens other than a series of coin flips. In one world the flips are independent and have chance .6 of coming up tails, while in the other they are independent and have chance .5 of coming up tails. It is possible for the first six flips to come up the same way in the two worlds while still maintaining different chances for the seventh flip. This can happen on any view on which chances are determined either by the Humean pattern including the future, or by non-Humean laws.)

On another picture of chance, the chances are determined not by the laws and the world, but by an experimental setup. The chance of a coin coming up heads may be 0.5 when the setup of the coin flipping situation is properly specified. But without a specification that the coin is flipped, that the flip is fair, that the coin is balanced, etc., it just may not be the case that it makes sense to say what the chance is that the coin will come up heads. On some ways of taking this, experimental *outcomes* are the result of chance processes, but experimental *setups* are the result of free choice of the experimenter. Conditional probability is a relationship between two events that are both in the domain of the probability function, while the experimental setup is a *precondition* for the existence of these probabilities at all. As Humphreys points out (Humphreys, 1985, 2004), Bayes' Theorem and other mathematical results allow us to invert conditional probabilities by means of some mathematical calculations. If there were such a thing as $p(\mathit{outcome} \mid \mathit{setup})$, then there would have to be something that is $p(\mathit{setup} \mid \mathit{outcome})$. But the setup is not the sort of thing that has a chance, as it is the result of a free choice, and the outcome is not the sort of thing that characterizes a chance process, so this conditional probability is either senseless or irrelevant. If we want to notate the role of the setup in determining the chance of the outcome, we should write it as $p_{\mathit{setup}}(\mathit{outcome})$, not $p(\mathit{outcome} \mid \mathit{setup})$.

This viewpoint on chance is similar to the one that frequentist statisticians have of probability. The only probabilities that make sense on this

view are the results of repeatable experiments. Scientific hypotheses help specify these probabilities, but do not themselves have probabilities, since they are not the results of repeatable experiments. This sort of thing is often notated by philosophers as $p_{setup}(E \mid H)$, where E is some evidence consisting of experimental outcomes, and H is a scientific hypothesis. The function represents something like the fraction of times that this outcome would occur if one were, hypothetically, to repeat this experimental setup many times, assuming the hypothesis is true. If this is the right way to represent the situation, then every statement of probability must have some scientific hypothesis or other that determines it, so every probability must be conditional.

However, I claim that on the frequentist interpretation, H should not be thought of as being conditioned on, but must instead be part of the background, just like a world, confirmational commitment, or experimental setup. The clearest reason for this is that on the frequentist account, H is from an importantly different ontological category than E, while conditional probability involves pairs of entities of the same ontological category. H is either true or false, and not the outcome of a repeatable experiment. A hypothesis, for the frequentist, is not the sort of thing that has a probability, so it is not the sort of thing that can be conditioned on. In statistical practice, the difference is often indicated by using a semicolon to set off the hypothesis that is the precondition for the probabilities, rather than the vertical line, which is used for conditional probabilities. Thus, we should write "$P(E; H)$" rather than "$P(E \mid H)$".

Furthermore, there *is* a notion of conditional probability that the frequentist *can* talk about, that is quite different. On the hypothesis that an urn has 3 white and 7 black balls, the conditional probability of the second draw (without replacement) being black given that the first is white is $7/9$, while the unconditional probability of the second draw being black is $7/10$. In this case we can calculate the conditional probability as the unconditional probability of a white draw followed by a black one, divided by the unconditional probability of the first draw being white, all given the background of the urn hypothesis, which has no probability of its own for the frequentist. The Bayesian can say that all of these probabilities are conditional on the hypothesis, because the Bayesian thinks that the hypothesis is the sort of thing that has a probability. But the frequentist shouldn't say this. So the frequentist has no special need for primitive conditional probabilities.

1.2.3 *Actual Frequencies*

Some have argued that on the actual frequency interpretation of probability, all probabilities are fundamentally conditional. For this interpretation, I

agree. When probability is interpreted as frequency of some property within an actual reference class, every probability really is conditional.

The interpretation of probability as actual finite frequency says that $p(B \mid A)$ is the fraction of entities with property A that also have property B. There is a particular number that is the frequency of heart attacks among 40-to-50-year-old American males in a given year, which we can calculate by counting how many 40-to-50-year-old American males there were that year, and counting how many of them had heart attacks that year. There is another frequency of heart attacks among all Americans, and another among all humans, calculated similarly. But if there is such a thing as the frequency of heart attacks independent of *any* reference class (even the entire universe), it is just a number, not a probability.

In this case, it looks like the reference class is the same sort of entity as the event whose probability is being measured. We can talk about the frequency of 40-to-50-year-old males among American heart attack victims, by counting how many heart attack victims there were that year, and finding what fraction of them were 40-to-50-year-old American males. Furthermore, if we ask for the *conditional* frequency of heart attacks among 40-to-50-year-old American males *given that* they smoke, this appears to be the same as the "unconditional" frequency of heart attacks among 40-to-50-year-old American males who are smokers. Conditionalizing really just is conjunction with the reference class. Thus, the reference class really is the same sort of thing as a conditioning event. Thus, on the actual finite frequency interpretation, we really do have a good case for every probability being conditional.

1.2.4 *Logical and Evidential Probabilities*

For logical and evidential probabilities (as well as perhaps some objective versions of the degree of belief interpretation of probability), some have argued that all probabilities are fundamentally conditional. For these interpretations, I don't specifically reject this argument. However, there is a special case of "empty background" that might be considered to be an unconditional probability that is equally fundamental to the conditional probabilities, so the upshot of the argument here is more equivocal.

Logical probability is often said to be a relation of partial entailment between two propositions. That is, "$p(B \mid A) = 1$" is said to mean the same thing (or something very similar to) "$A \vdash B$." Saying that $p(B \mid A) = 2/3$ is saying that A "2/3 entails" B. Since entailment is a binary relation, this logical probability is said to be an essentially conditional relation. This is the point of view described, for instance, by Keynes (1921). (A similar viewpoint, though not identical, is expressed with regards to the "evidential probabilities" of Williamson, 2002.)

Both roles here are played by arbitrary propositions, so there are no ontological distinctions between the two sides of the conditional probability. There is no category mistake in reversing a logical entailment (though of course the degree of entailment can differ). Furthermore, just like with actual finite frequencies, there doesn't appear to be any other notion of conditional probability that is interestingly distinct from this one. The probability of A given B, with C as background, doesn't obviously have any interpretation that would be clearly different from the probability of A with $B \wedge C$ as background. Thus, just as with actual frequencies, one might be able to argue on conceptual grounds that all logical probabilities are inherently conditional.

However, unlike with frequencies, the opponent of this view has a response. Deductive logic can be expressed as the study of logical entailment relations, but it can also be expressed as the study of theorems. One can think of theorems either as sentences entailed by a tautology, or as sentences entailed by no premises whatsoever. Similarly, it may be possible to consider the set of logical probabilities conditional on a tautology either as the degree of partial entailment the tautology gives to each sentence, or as the degree of partial theoremhood each sentence has.

If we can interpret $p(B \mid A)$ as the degree to which A partially entails B, we may also be able to interpret $p(A)$ as the degree of partial theoremhood of A. On this account, it may be further possible to recover all the partial entailments from these facts about partial theoremhood through techniques of calculating conditional probabilities, just as it is possible to recover all the deductive entailments from the facts about theoremhood through the deduction theorem. Thus, the opponent of conditional probability as the fundamental notion may have a response to this argument, though it will depend on the extent to which conditional probabilities really can be recovered from the unconditional ones, just as in the case of Objective Bayesianism, or Levi's confirmational commitments.

1.2.5 *Summary*

In summary, degree of belief, physical chance, experimental chance, and hypothetical frequency all have some fundamental ontological distinction between the bearers of probability and the backgrounds that are required for probabilities to even exist. Thus, the necessity of these backgrounds does not motivate the claim that conditional probability is primitive or fundamental. For actual frequencies, logical probability, and evidential probability, the backgrounds are of the same type as the bearers of probability, so this argument does seem to motivate the claim that conditional probability is fundamental. But for logical and evidential probability, there is a possibility of empty background, which can be re-interpreted as a

fundamental notion of unconditional probability. Further mathematical investigation is needed to see whether these unconditional probabilities suffice to determine the conditional probabilities. Only for actual frequencies is it clear that all probabilities really are conditional, because of the necessity of a background for probability.

- All probabilities are non-trivially conditional:
  - Actual frequency
- All are conditional, some conditions are empty:
  - Logical
  - Evidential
  - Unique Degree of Belief
- Background relevant, not all are conditional:
  - Chance
  - Hypothetical Frequency
  - Permissive Degree of Belief

1.3 *Problems for the Ratio*

The previous section considers conceptual arguments that all probabilities are fundamentally conditional. I have argued that this argument works for the interpretation of probability as actual frequency, and is equivocal for logical and evidential probability and related objective epistemic interpretations, but that it does not work for the other interpretations of probability. In this section, I consider arguments for the claim that all probability is fundamentally conditional based on the mathematical features of conditional probability. This set of arguments is the center of Alan Hájek's (2003). Although this argument is perhaps easier to feel the grip of, and is largely independent of the particular interpretation of probability, I put it second, because consideration of it leads naturally to the technical issues considered in the later sections of this article.

The immediate target of Hájek's argument is the common claim that conditional probability is just *defined* as $p(A \mid B) = p(A \wedge B)/p(B)$. As Hájek points out, it appears to be a consequence of this definition that there is no such thing as the conditional probability $p(A \mid B)$ unless $p(B)$ has a precise non-zero numerical value. He then gives a litany of cases in which it seems clear that $p(A \mid B)$ exists, even though $p(B)$ is either zero, imprecise, vague, or non-existent. Thus, we must reject the ratio analysis as a definition of conditional probability. Whether this requires conditional probability to be a (or the) fundamental concept of probability theory is a deep and

difficult question that depends on what alternatives to the ratio analysis exist. The rest of the article after this section is a long consideration of these alternatives. Section 1.4 defines the particular mathematical features of probability and conditional probability that come up in addressing this problem. Section 2 and Section 3 consider the two advanced mathematical characterizations of conditional probability that avoid the problems of the ratio definition, one of which makes conditional probability primary and the other of which allows it to (almost) be calculated from unconditional probability. Evaluation of the merits of these two mathematical accounts is thus essential for deciding whether or not to accept Hájek's argument that conditional probability is prior to unconditional probability.

I will give examples of Hájek's cases shortly. I think that most are not decisive, but there is one family of them that is quite convincing for every interpretation of probability mentioned above, apart from actual frequencies. Thus it is interesting that the two primary arguments for conditional probability being fundamental have this complementary distribution—the one interpretation for which Hájek's argument against the ratio analysis clearly fails is the one interpretation for which all probabilities clearly require a background of the same type as the bearers of probability, so that it can clearly be understood as conditional probability.

1.3.1 *Impossible or Ruled Out Conditions*

I will begin by considering a type of case Hájek considers that is easy to reject. I think it is important to consider how this type of case differs from the others, which are more plausibly relevant. Let H be the proposition that a particular coin flip comes up heads, and T be the proposition that this same flip comes up tails. Hájek claims that $p(T \mid T) = 1$ under any circumstance. In particular, he claims (p. 287) that this should be true even if p is the function encoding physical chances at a time when the flip has already happened and the coin already came up heads, so that $p(T) = 0$. He also suggests that it should be true if p is the function encoding degrees of belief of a rational agent who has already learned that the coin came up heads, so that $p(T) = 0$.

These cases can be rejected because there doesn't appear to be a clear meaning for these conditional probabilities. Although I don't think that conditional probabilities are the probabilities of conditionals, there is a useful analogy to be drawn with conditionals. Conditional probability is intended to capture something more like an indicative conditional, rather than a subjunctive conditional or a material conditional, and indicative conditionals generally aren't considered in cases where the antecedent has already been fully ruled out. It seems correct to say, "if Oswald didn't kill Kennedy then someone else did," but this is because we allow that

our knowledge of the circumstances of the assassination is fallible. If we imagine fully ruling out any possibility that Oswald didn't commit the assassination, then the conditional becomes harder to interpret. We can apply subjunctive or material conditionals even to cases of necessary falsehoods, but it's hard to interpret them as indicative conditionals. Maybe we can make sense of a sentence like, "if 7 hadn't been a prime number, then 8 would have been," but a sentence like "if 7 isn't a prime number, then 8 is" seems only interpretable as a material conditional. Just as indicative conditionals seem not to be acceptable when the antecedent has been fully ruled out, none of the purposes for which conditional probabilities have been proposed makes any use of probabilities conditional on antecedents that have already been ruled out. There is no question of updating on or confirming a hypothesis that has been completely eliminated.

There are processes of belief *revision*, on which one removes a belief that one already has before updating on new information, but this is a different process that uses conditional probability from the revised state rather than the current state.[1] Similarly, the probability of outcomes conditional on acts that weren't done is irrelevant to decision theory.[2] Similarly, there is no question of how the chances of events will evolve when something that didn't occur does occur (though there may be a question of how chances will evolve when something of similar type to that event does occur), and there is no question of the degree of causal relevance of something that didn't occur (though there may be a question of the degree of causal relevance of its *non*-occurrence, which of course is something that *did* occur).

1.3.2 *Vague, Imprecise, or Gappy Conditions*

A second class of cases that Hájek considers involve vague or imprecise probabilities (pp. 293–5). It is controversial whether imprecise probabilities even exist (see the entries by Titelbaum and Mahtani in this volume for further discussion). But if they do, then it's clear that they cause problems. Perhaps one is uncertain about the outcome of the next United States presidential election in such a way that one has imprecise credences about it. Or perhaps it depends on non-deterministic events in a way that leaves it with an imprecise chance. Nevertheless, if D is the proposition

1 Levi's notion of confirmational commitments allows for probability conditional on propositions that are currently known to be false. But in this case, the probability function is not the current degree of belief function, but rather the confirmational commitment—the current degree of belief function is itself conditional on current knowledge. Thus, the probability conditional on something currently known to be false is a prior commitment of an indicative sort—not Hájek's probability conditional on a certain falsehood.

2 Brandenburger (2007) has argued that game theory sometimes needs to consider probabilities conditional on actions that are ruled out by rationality considerations, but these are not ruled out with certainty, the way that tails was in Hájek's examples.

that a Democrat will win the next US presidential election, and H is the proposition that a completely unrelated coin flip will come up heads, it seems clear that $p(H \mid D) = 1/2$.

However, this challenge may not be a fatal objection to the ratio analysis either. One proposal about imprecise probabilities is that, rather than $p(D)$ being an imprecise value (or set or whatever), there are instead multiple precise probability functions p_i that are all part of the representation of degree of belief, or chance, or whichever interpretation of probability we are considering. On each such function, $p_i(H \mid D)$ can be well-defined by the ratio formula, and if they all happen to take value $1/2$, then the conditional probability can be precise even though the unconditional probability is not. (This response is described in slightly greater detail on page 295 of Hájek's paper.)

Hájek puts the most weight on cases where there is *no* unconditional probability, but conditional probabilities are well-defined. He gives a long series of such cases on pp. 295–312. These include cases of free actions (which may be such that they *can't* have credences or chances), mere gaps in the credences or chances, and cases of non-measurable sets.

I think that mere gaps are either best thought of as maximally imprecise probabilities and addressed supervaluationally as above, or as events that are outside of the scope of the relevant probability function. An agent who fails to have a degree of belief in some proposition is an agent who hasn't considered or grasped it, and thus fails to have any degree of belief conditional on it as well (even though there are some facts about what degree of belief she *should* have were she to have them—like $p(A \mid A) = 1$). Similarly with non-measurable sets—if they are outside the bounds of chance or credence, then there are no meaningful conditional probabilities on them either.

There may be some class of events (perhaps the actions of a free agent who is in the process of deliberation) that *can't* have probabilities, but which themselves serve as the conditions for probabilities of other events. However, some of these may in fact be better thought of as the "backgrounds" for probabilities that I considered in Section 1.2. This may be the right way to think of the "action probabilities" of decision theory, for instance, where every probability must depend on a specification of the action of the agent. However, if there were a class of events that can't have probabilities, but which also aren't essential to the specification of other probabilities, even though they can affect them, then this would be a better case.

1.3.3 *Probability 0 Conditions*

At any rate, I think the strongest case is one that Hájek puts less weight on (pp. 289–290). These are cases arising from consideration of infinite probability spaces, where some events have probability 0 *without* being ruled out. Consider a point on the surface of a sphere. Label the sphere with lines of latitude and longitude like those of the Earth. Let N be the proposition that the point is in the northern hemisphere. Let L_θ be the proposition that the point is on the line of longitude at angle θ from the boundary between the eastern and western hemispheres. If the initial probability distribution is uniform, then it is quite plausible that $P(N \mid L_0) = 1/2$, even though $P(L_0) = 0$, so that $P(N \wedge L_0)/P(L_0)$ is undefined. Furthermore, even if the initial probability distribution isn't uniform, it seems that $P(N \mid L_\theta)$ should be defined whenever there is some possibility of L_θ being true. However, there are uncountably many distinct values of θ, and at most countably many of them can have positive probability (because at most n of them can have probability greater than $1/n$, for each of the countably many integers n, and any positive number is greater than $1/n$ for some integer n). Thus, there must be some way to make sense of these conditional probabilities, despite the use of probability 0. This example can be generated for probability interpreted as chances or as degrees of belief or as evidential probability, or any interpretation, as long as there are uncountably many distinct possibilities that aren't ruled out.

There are two methods that have been proposed to block this set of cases. One is to introduce additional non-zero values for the probability function to take that are nevertheless lower than $1/n$ for any positive integer n. I have argued elsewhere that this method is unlikely to be correct for chances or degrees of belief (Easwaran, 2014). (This proposal is discussed in more detail in the contribution to this volume by Sylvia Wenmackers.) Furthermore, this option bears some relationship to one of the proposals described later, in Section 3.1, so I suggest that this is in some sense not really an alternative to the methods considered here—it is effectively equivalent to letting the probability take the value 0.

The other method for blocking this sort of case is to argue that the relevant notion of probability *can't* have uncountably many disjoint possible events. In the case of Bayesian probabilities, this is motivated by some consideration of the finitude of the human mind, while in the case of chances it is motivated by some understanding of quantum mechanics as requiring the universe to be discrete in time, space, and every other meaningful parameter.

However, this sort of interpretation of quantum mechanics is implausible. Although certain parameters like charge and spin are quantized, time and space just enter into "uncertainty" relations. This means that they are

bound to other parameters in a way that interactions depending very precisely on one parameter must allow for exceedingly large variation on the other. However, this does not put any specific lower bound on the precision of any interaction, and doesn't directly motivate the idea that space and time are discrete.

Furthermore, although any particular human mind is finite, there is reason to allow consideration of every hypothesis of the form $V > p/q$, where V is some physical parameter, and p and q are integers. Certainly, science seems to proceed as if each of these hypotheses is meaningful, even if we can never be absolutely sure which are true or false. But these countably many hypotheses together generate a family of uncountably many hypotheses of the form $x = r$ where r is a real number. (The claim that all of the relevant algebras are countably generated, or generated by random variables in this way will be important in Section 2.3.2.) The example with points on a sphere is exactly like this, but so are many others that are more directly relevant in science. To reject these cases is to say that every probability function has some finite limit on the size of examples that are relevant.

This response in terms of finitism is quite effective in the interpretation of probability as actual frequency, if the classes of events one is discussing are always finite. (When the classes may be infinite, it's hard to say how to even *define* the notion of frequency involved.) But this response is no help to the statistical frequentist, who may be interested in scientific hypotheses of the relevant sort. Philosophers often make reference to examples involving a dart thrown at a board, with infinitely many points that its center might hit, or a fair coin being flipped infinitely many times, for which each sequence of heads and tails is a possible outcome. But examples involving infinity are central to much scientific practice as well.

For instance, a statistical frequentist may be interested in some hypothesis about how energetic particles are ejected from an atomic nucleus under a particular sort of process. She may further be interested in the question of how the energy distribution of these particles is correlated to the direction in which they are ejected. If we let E_x be the statement that the energy of the particle is x, and D_θ be the statement that the particle is ejected in a direction at angle θ to the motion of the atomic nucleus, then she could be interested in all probabilities of the form $p(E_x \mid D_\theta)$. But if she hypothetically imagines the process being repeated infinitely many times, the probability of many of the D_θ is likely to be zero, given that there are uncountably many directions in which the particle could be ejected. If we limit consideration to some *actual* set of experiments, then there are likely to be only finitely many such ejections, and so the non-realized D_θ can be ignored. But the statistical frequentist is interested in *hypothetically* repeated experiments, so all of these possibilities must be considered.

To summarize, there may be a way to resist all of these cases. But it would involve some extensive use of special backgrounds for certain types of probability, a particular way of dealing with any kind of imprecision in probability functions, and a rejection of infinity. Most of the mathematical work on alternatives to the ratio analysis only address the issue of infinite probability spaces and probability 0. I think that the other problems can be avoided as in ways that I have suggested along the way. But there is certainly room for further philosophical and mathematical analysis of those suggestions, and perhaps for new alternatives, which may or may not prioritize conditional probability over unconditional probability. But the rest of this article will examine the mathematical theories that have been developed for dealing with the problems that arise around infinite probability spaces and the resulting events of probability 0.

1.4 *Additivity, Disintegrability, and Conglomerability*

Once we consider these infinite families of hypotheses, it seems that we must have some way of making sense of $p(A \mid B)$ even when $p(B) = 0$. There are many different mathematical theories that allow this to work out, and these will be the subject of the further sections of this article. The reason there are so many different theories is due to a fundamental dilemma around infinity, which will take some time to explain.

Every such theory begins with the idea that the "definition" $p(A \mid B) = p(A \wedge B)/p(B)$ should be replaced with an axiom $p(A \mid B)p(B) = p(A \wedge B)$. We can then consider whether further information allows us to define $p(A \mid B)$ from the unconditional values, or at least in some sense ground it in them, or whether we must take $p(A \mid B)$ as a fundamental function separate from the unconditional probability function $p(A)$. However, even allowing for this function, there are difficulties when the set of possibilities is infinite.

In this section I will discuss some of the mathematical properties involved, and show that the idea that conditional probability can be understood as a function $p(A \mid B)$ conflicts with the natural generalization of Additivity in cases of infinity. We must either give up on Additivity (and related principles generalizing the Law of Total Probability), or else accept that conditional probability is given by a function $p(A \mid B, \mathcal{E})$ for a further parameter $\mathcal{E}$. The mathematical theory of conditional probabilities for infinite sets is an interplay between the two horns of this dilemma.

In this section I will formally treat the bearers of probability as sets of possibilities, and will largely bracket concerns about the interpretation of probability until the end.

1.4.1 *Additivity*

When dealing with infinity, a fundamental question for probability theory is whether and how to generalize the notion of Additivity. One of the standard axioms of probability is that if A_1 and A_2 are disjoint events (that is, there is no possibility on which they both occur) then $p(A_1 \cup A_2) = p(A_1) + P(A_2)$. Kolmogorov and others have considered a generalization of this axiom to countable cases.

Definition 1 *The A_i for $i \in I$ form a* partition *of A iff each A_i entails A, and whenever A is true, exactly one of the A_i is true.*

(If no particular A is mentioned, then I am considering a partition of the set of all possibilities.) Thinking of the A_i as sets, that means that they are disjoint, and their union is A. I will refer to this partition with boldface $\mathbf{A_I}$, and with the index set I as subscript, while italic A_i, with a member i of I as subscript, will refer to the member of $\mathbf{A_I}$ that is indexed by element i.

One way to state Countable Additivity is as the requirement that for any countable partition $\mathbf{A_I}$ of A, we have $p(A) = \sum_{i \in I} p(A_i)$. Kolmogorov actually framed his axiom in a slightly different form as a sort of continuity—whenever the B_i for $i \in \mathbb{N}$ are a family of sets whose intersection is empty, we have $\lim_{n \to \infty} p(\bigcap_{i=0}^{n} B_i) = 0$.

However, I think that it is more perspicuous to phrase this generalization in a third way, in order to more clearly demonstrate the further generalizations to uncountable sets. The following is a theorem of standard finitely additive probability, whenever $\mathbf{A_I}$ is a partition of A.

Theorem 1 *If $x \geq p(A)$, then for any finite $I_0 \subseteq I$, $x \geq \sum_{i \in I_0} p(A_i)$.*

We can then define additivity as the converse.

Definition 2 ($\mathbf{A_I}$-Additivity) *If for every finite $I_0 \subseteq I$, $x \geq \sum_{i \in I_0} p(A_i)$, then $x \geq p(A)$.*

The following definition is equivalent.

Definition 3 ($\mathbf{A_I}$-Additivity) *If $x < p(A)$ then there is some finite $I_0 \subseteq I$ such that $x < \sum_{i \in I_0} p(A_i)$.*

Countable Additivity is equivalent to $\mathbf{A_I}$-Additivity for all countable sets of indices I.[3] This is because, for a set of non-negative real numbers,

3 We can also naturally talk about κ-Additivity as $\mathbf{A_I}$-Additivity for all I with cardinality *less* than κ. This is standard notation though it is slightly confusing that Countable Additivity, also known as "σ-Additivity," is $\aleph_1$-Additivity, while $\aleph_0$-Additivity is actually *Finite* Additivity. But this notation is relevant to distinguish between Additivity for all cardinals strictly below $\aleph_\omega$, and Additivity for all cardinals up to and including $\aleph_\omega$, which is called $\aleph_{\omega+1}$-Additivity.

the sum of that set is the smallest real number that is at least as great as every finite sum of those numbers.[4]

Countable Additivity is not entailed by the standard probability axioms, and in fact rules out certain intuitively appealing probability distributions. The classic proposed counterexample to Countable Additivity is often known as the "de Finetti lottery" (de Finetti, 1974; for more detailed discussion see Bartha, 2004, and Howson, 2008). Imagine that some natural number is chosen in such a way that no number is more likely than any other. This intuitively seems possible, and yet it is ruled out by Countable Additivity. Since every number is equally likely to be chosen, each number must have probability less than $1/n$, because otherwise some n of them would exhaust all the probability. The only way for this to be the case is for each number to have probability 0. But this is a violation of Countable Additivity, because the sum of these 0s is strictly less than 1, which is the probability of the countable disjunction of these possibilities.

Considering Definition 3, we can derive a more general set of apparent problems. Let each A_i stand for the event of the number i being picked, and let I be the set $\mathbb{N}$ of all natural numbers, so that $\mathbf{A_I}$ is a partition of the necessary claim that some number or other is picked. In this case, Definition 3 of $\mathbf{A_I}$-Additivity states that for every $x < 1$, there must be some finite I_0 such that $x < \sum_{i \in I_0} p(A_i)$. That is, for every $x < 1$, there is some finite set such that the probability that the number chosen is from that set is at least x. $\mathbf{A_I}$-Additivity doesn't just rule out uniform distributions on the natural numbers—it requires that *every* distribution concentrate most of the probability on some finite set or other.

If $\mathbf{A_I}$-Additivity holds for *all* partitions $\mathbf{A_I}$, then the probability function is said to be Fully Additive. In this case, for any partition $\mathbf{A_I}$ of a set

4 Readers may be familiar with the definition of the sum of a sequence of (non-negative or negative) numbers a_i for $i \in \mathbb{N}$ as

$$\sum_{i \in \mathbb{N}} a_i = \lim_{n \to \infty} \sum_{i=1}^{n} a_i.$$

This definition doesn't work for index sets other than $\mathbb{N}$, and makes essential use of the order of the indices. When some terms are negative, this order can be important—the same set of numbers can have a different sum when added in a different order, if both the negative and positive terms separately sum to infinite values. But when all terms are non-negative, the least upper bound of the sums of finite subsets is the same as the sum of the terms in any order (because every finite initial sequence is a finite subset, and every finite subset is contained within some finite initial sequence, and since there are no negative terms, the sum of any larger subset is at least as great as the sum of any subset contained within it).

For *uncountable* infinite sets of non-negative numbers, it is hard to extend the sequential definition, because we don't have good methods for dealing with uncountably long sequences. However, the least upper bound of the set of all sums of finite subsets is still well-defined.

A, Definition 3 entails that for every n, there is a finite set of A_i whose probability adds up to more than $p(A) - 1/n$. Let $I' \subset I$ be the union of the countably many finite sets of indices of these sets, which is thus countable. By Theorem 1, if we let $A' = \bigcup_{i \in I'} A_i$, then $p(A') \geq p(A) - 1/n$ for each n (since it contains a finite subset adding to this probability). Since $A' \subset A$, we have $p(A') = p(A)$. Thus, the remainder of A that is not in A', $A \setminus A'$, must have probability 0. If A was the set of all possibilities, and each A_i is a singleton set containing a single possibility, then A' is countable. Not only does each element outside of this countable set *individually* contribute probability 0, but even *collectively* they all contribute 0.[5] Thus, if Full Additivity holds, there is a sense in which we can ignore all but countably many possible outcomes, and these countably many outcomes have individual probabilities that add up to 1. A probability function in which the set of all possibilities is countable is said to be *discrete*. While there are many interesting applications of discrete probability, there are also plenty of applications for which no countable set of possibilities should account for all the probability, such as any scientific question for which every real number within some interval is a possible answer. Thus, most probability theorists do not accept Full Additivity.

We can think of different views of probability as along a sort of scale (Figure 1). At the most restrictive end there is the strongly finitistic view that there are only finitely many possibilities that probability is distributed over. Next we get the discrete view, that there are only countably many possibilities that probability is distributed over—this is classical probability theory with Full Additivity for all cardinalities. Next we get the traditional mathematical view on which the set of possibilities can be uncountable, but the probability function is required to satisfy Countable Additivity. Finally, at the most liberal end of the scale, we have the minority view in mathematics but a popular view in philosophy, where the probability space can be uncountable and the probability function is only required to satisfy Finite Additivity. (Some of the popularity of this view among philosophers may stem from confusion with probability over finite spaces, at the opposite end of the scale.) Finite and discrete probability have no problem with Additivity, and in fact allow conditional probability to be uniformly defined by the ratio. However, the consideration of scientific examples where we want to measure the unknown value of some parameter push us towards uncountable spaces. So it is useful to investigate the

5 Another way to see this is to consider the probabilities of each individual possibility. For each n, at most n of the individual possibilities can have probability greater than $1/n$. Thus, at most countably many have non-zero probability. But if Full Additivity holds, then the sum of all the probabilities of the individual possibilities must be 1. So these countably many non-zero probabilities must add up to 1. Thus, the set of all possibilities other than the countably many with non-zero probability must be a set with probability 0.

ways in which probability functions with failures of Additivity can still be well-behaved. I believe that Countable Additivity is the most useful point on this scale, but it is worth considering the mathematical features of all four points.

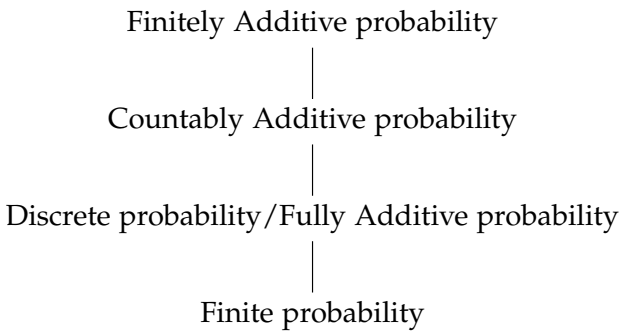


Figure 1: A scale of views

1.4.2 *Disintegrability and Conglomerability*

Although generalizations of Additivity are quite controversial, there are related principles that have been argued to generalize to infinite cases. These principles are defined by using integration in place of addition when infinity arises, to avoid some of the difficulties of adding up zeros. By the end of this section, I will mention some results that show that instances of these principles must fail when instances of Additivity fail. However, in Section 1.4.3, I will show that we can avoid these failures by defining conditional probability relative to a partition.

The starting point for discussion of these principles is the Law of Total Probability.

Theorem 2 (Finite Law of Total Probability) *If A_1 and A_2 are incompatible, and A is the disjunction $A_1 \cup A_2$, then*

$$p(B \cap A) = p(B \mid A_1)p(A_1) + p(B \mid A_2)p(A_2).$$

Given two instances of the conjunction law, $p(B \cap A_i) = p(B \mid A_i)p(A_i)$, this is equivalent to an instance of Additivity: $p(B \cap A) = p(B \cap A_1) + p(B \cap A_2)$. We can state a generalization of this, where $\mathbf{A_I}$ is a partition of some set A.

Definition 4 *The* $B \cap \mathbf{A_I}$-Law of Total Probability *states that*

$$p(B \cap A) = \sum_{i \in I} p(B \mid A_i)p(A_i).$$

Given that $p(B \cap A_i) = p(B \mid A_i)p(A_i)$, it is straightforward to see that the $B \cap \mathbf{A_I}$ Law of Total Probability is equivalent to $B \cap \mathbf{A_I}$-Additivity. Giving up Full Additivity means giving up certain instances of the Law of Total Probability. But there are ways of modifying the Law of Total Probability that don't directly take this additive form.

The Law of Total Probability can be related to considerations of expected value for random variables. Informally, a random variable is some quantity with a potentially unknown real number value, where for each real number x, there are well-defined probabilities $p(V > x)$ and $p(V = x)$. Notably, the set of events $V = x$ form a partition.

Definition 5 *When there are only finitely many possible values for V, the* expected value *of V is given by*

$$exp(V) = \sum_x x \cdot p(V = x),$$

where the sum ranges over all finitely many possible values for V.

This definition would yield strange results if it were applied to a variable V for which Additivity fails on the partition into $V = x$.

Any violation of Additivity must involve some partition $\mathbf{A_I}$ such that $\sum_{i \in I} p(A_i) = 1 - \epsilon$. If I has cardinality at most that of the set of real numbers, then we can generate a random variable whose expected value under an extension of the above definition would be paradoxical. For each $i \in I$, let ϵ_i be a distinct positive value less than $\epsilon/(1-\epsilon)$. Let V be a random variable that takes on the value $1 + \epsilon_i$ iff A_i is true. Then a naive extension of Definition 5 would tell us that $exp(V) = \sum_{i \in I}(1 + \epsilon_i)p(A_i)$. But by choice of ϵ_i, we see that $(1 + \epsilon_i) < (1 + \epsilon/(1-\epsilon)) = 1/(1-\epsilon)$. Thus, $exp(V) < \sum_{i \in I}(1/(1-\epsilon))p(A_i) = (1/(1-\epsilon))(1-\epsilon) = 1$. That is, even though V is a random variable whose value is always strictly greater than 1, this definition of expectation would yield an expected value that is strictly less than 1.

To avoid this problem, it has been standard to define expected value slightly differently in infinite cases. Instead of directly considering the probability of $V = x$ for each possible value that V can take on, mathematicians just directly rule out discontinuities like the one mentioned above. If V is a random variable that only has finitely many possible values, then we follow the old definition and let $exp(V) = \sum_x x \cdot p(V = x)$. If V has infinitely many possible values, but has a lower bound (that is, there is some l such that it is certain that $V > l$), then we can avoid this problem. If V' is a random variable that always takes a value strictly less than V, we will say $V' < V$. We will just directly stipulate that if $V > V'$ then $exp(V) > exp(V')$. This will rule out the problem of the previous paragraph, because we could let V' be the random variable that always takes

the value 1, and see that $exp(V) > exp(V') = 1$. By considering variables V' that only take on finitely many distinct values, we get a set of lower bounds for what $E(V)$ could be. We say that the expectation of V is the least number above all these lower bounds (the "supremum" of this set of lower bounds).

Definition 6 *Let V be a random variable with a lower bound. Then*

$$exp(V) = \sup_{V' < V} exp(V'),$$

where V' ranges over variables that only take on finitely many distinct values.

Similarly, for random variables that have an upper bound, we can define the expectation to be the greatest number below all the upper bounds (the "infimum" of this set). We then deal with unbounded random variables by breaking them into a component with a lower bound and an upper bound. Let V^+ be the random variable that agrees with V when V is positive and is 0 otherwise, and V^- be the random variable that agrees with V when V is negative and is 0 otherwise. Then define $exp(V)$ as follows.

Definition 7

$$exp(V) = \int V \, \mathrm{d}p = \sup_{V' < V^+} \sum_x x \cdot p(V' = x) + \inf_{V' > V^-} \sum_x x \cdot p(V' = x),$$

where V' ranges over random variables that only take finitely many distinct values.

This is the definition of the Lebesgue integral of V with respect to probability function p, and is the final generalized definition of expected value. It agrees with Definition 5 and Definition 6 in the cases where they apply.

With this new definition, we can try to save the Law of Total Probability in a slightly different form. Let $\mathbf{A_I}$ be a partition. We can consider $p(B \mid \mathbf{A_I})$ as a random variable whose value is given by $p(B \mid A_i)$ for whichever proposition A_i is the unique one from $\mathbf{A_I}$ that is true. If $\mathbf{A_I}$ is finite, then the Law of Total Probability takes the form $p(B) = exp(p(B \mid \mathbf{A_I}))$. This motivates the following definition.

Definition 8 *B is* Disintegrable *over the partition $\mathbf{A_I}$ iff*

$$p(B) = \int p(B \mid \mathbf{A_I}) \, \mathrm{d}p.$$

Disintegrability is thus another generalization of the Law of Total Probability, formulated with integrals rather than (potentially infinite) sums.

Let $\mathbf{A_I}$ be any partition, I' be any subset of I and $A' = \cup_{i \in I'} A_i$. Define *Conglomerability* as follows.

Definition 9 $p(B \mid \mathbf{A_I})$ *is* Conglomerable *over* A' *iff*

$$\inf_{i \in I'} p(B \mid A_i) \leq p(B \mid A') \leq \sup_{i \in I'} p(B \mid A_i).$$

It is useful to compare Conglomerability to van Fraassen's principle of "reflection" (van Fraassen, 1984; Briggs, 2009).

It is not hard to see that Disintegrability of B over $\mathbf{A_I}$ entails Conglomerability over each A' with positive probability (because constant functions taking on the infimum or supremum of $p(B \mid A_i)$ are among the set of random variables whose expectation is considered in calculating $exp(p(B \mid \mathbf{A_I}))$). Conversely, Conglomerability of $p(B \mid \mathbf{A_I})$ over all A' with positive probability entails Disintegrability of B over $\mathbf{A_I}$. (Since the integral is defined by comparison to finite sums, this only requires the Finite Law of Total Probability, rather than the generalizations that fail when Additivity fails over infinite partitions.)

We might hope that these new generalizations of the Law of Total Probability in terms of integration rather than summation don't require Countable Additivity. However, this hope turns out to be misplaced. A general theorem is proven by Hill and Lane (1985), verifying that for countable probability spaces, Conglomerability and Countable Additivity are equivalent. That is, any failure of Countable Additivity entails a failure of Conglomerability, and thus Disintegrability, which is the generalization of the Law of Total Probability. (Slightly more general versions of this result were proven earlier by Schervish, Seidenfeld, and Kadane, 1984.)

Instances of this result were noted by de Finetti (1974, pp. 177–8), who also conjectured the general result but hadn't proven it. To see the basic idea, consider something like the de Finetti lottery, where each natural number has equal probability of being chosen. Let E be the event that an even number is chosen. Intuitively, $p(E) = 1/2$. However, if we consider the partition into the sets $A_i = \{2i + 1, 4i, 4i + 2\}$, then intuitively $p(E \mid A_i) = 2/3$, so that the unconditional probability of E, which is $1/2$, is strictly outside the range spanned by its probabilities conditional on each member of the partition, which are all $2/3$. The construction by Hill and Lane notes that even without the assumptions of uniformity underlying the specific probability judgments $1/2$ and $2/3$, if E and its complement are both sets of positive probability, then we can often create each A_i by taking enough elements of E with one element of its complement to make $p(E \mid A_i) > p(A) + \epsilon$. If we can't do this for every element of the complement, we can usually do it by taking enough elements of the complement with one element of E to make $p(E \mid A_i) < p(A) - \epsilon$. The tricky part of the Hill and Lane construction is showing how to create a special partition in the case where neither of these techniques works. These results have been generalized to show that there are failures of Conglomerability for

probability distributions that satisfy Countable Additivity but fail to satisfy Additivity at some cardinality beyond the countable (Seidenfeld, Schervish, & Kadane, 2013, 2014). Thus, Disintegrability and Conglomerability don't let us get quite as much distance from Additivity as we might hope.

1.4.3 *The Fundamental Dilemma*

However, there is a way to separate Disintegrability and Conglomerability from Additivity.

First, we should note that Additivity only makes reference to unconditional probabilities, while Disintegrability and Conglomerability make reference to conditional probabilities. Furthermore, Disintegrability and Conglomerability make reference to conditional probabilities $p(B \mid A_i)$ only in the context of a random variable $p(B \mid \mathbf{A_I})$. In generating a contradiction to Conglomerability from a failure of Additivity, Hill and Lane needed to construct a *new* partition by joining together elements of $\mathbf{A_I}$. (This is also the case for Seidenfeld et al.) Thus, if a given set A is an element of two distinct partitions $\mathbf{A_I}$ and $\mathbf{A'_{I'}}$, we can avoid the problems if we *change* the value of $p(B \mid A)$ when we move from considering $\mathbf{A_I}$ to considering $\mathbf{A'_{I'}}$. That is, we should consider conditional probability as a three-place function, $p(B \mid A_i, \mathbf{A_I})$, so that changing just the partition can change the value of the conditional probability, even if we are considering the same events B and A_i. Some theorists find this repugnant to their sense that conditional probability $p(B \mid A_i)$ must have a single value, but it enables us to avoid the paradoxes.

This move was in fact already made by Kolmogorov (1950). Although he hadn't noticed the connections between Additivity principles and Conglomerability, he had already noticed some problems that Conglomerability apparently led to, and avoided them by turning conditional probability into a three-place function of two events and a partition.[6] (In fact, this problem was already mentioned as early as Bertrand, 1889, though due to Borel's work on this problem, and the existence of another paradox known as "Bertrand's Paradox," this has come to be known as the "Borel Paradox.")

Imagine a point uniformly chosen from the surface of a sphere, labeled with latitude and longitude like the surface of the Earth. Consider the set P of "polar" points—those with latitude greater than 60 degrees north or greater than 60 degrees south. Consider the set E of "equatorial" points—those with latitude between 30 degrees south and 30 degrees north. Let L_θ be the great circle of longitude θ. By symmetry, it seems that $p(P \mid L_\theta)$ should be independent of θ, and so should $p(E \mid L_\theta)$. Conglomerability

6 Strictly speaking, Kolmogorov worked with a "sub-σ-algebra" rather than a partition, but we will discuss the relation of these concepts in Section 2.

over the partition[7] $\mathbf{L}_\theta$ requires that $p(P) = p(P \mid L_\theta)$ and $p(E) = p(E \mid L_\theta)$. But $p(P) = \frac{2-\sqrt{3}}{2} \approx 1/8$ while $p(E) = 1/2$. Note that P and E each cover $1/3$ of the length of L_θ. Thus, conditionalizing a uniform distribution over the sphere in a way that is Conglomerable over the longitudes gives a conditional distribution that is concentrated near the equator and away from the poles.[8]

To force a problem for the two-place conditional probability function, we can fix a given line of longitude and shift which partition it is considered as a member of. Re-describe the sphere so that the poles are still on this line, but where the old equator was. This switches which points on the line are polar and which are equatorial. Conglomerability requires the very same great circle to give rise to different conditional probabilities when considered as a line of longitude for one set of coordinates, rather than as a line of longitude for a different set of coordinates. If we let C be this circle, and $\mathbf{L}_\theta$ be the partition into lines of longitude for the given poles, while $\mathbf{L}_\phi$ is the partition into lines of longitude for poles where C intersects the equator of the original partition, then we get $p(P \mid C, \mathbf{L}_\theta) = \frac{2-\sqrt{3}}{2}$ while $p(P \mid C, \mathbf{L}_\phi) = 1/2$. Conditioning on the same event gives different results when that event is considered as drawn from one partition rather than another.

Thus, Conglomerability already motivates the idea that conditional probability depends not just on the conditioning event, but also on the partition from which that event is drawn. Since the arguments from Conglomerability to Additivity rely on generation of new partitions, we might hope that

7 Strictly speaking, $\mathbf{L}_\theta$ do not form a partition, because every line of longitude includes the poles. However, the example can be slightly modified without making any significant changes to anything by just removing the poles from the sphere, or arbitrarily adding the poles to one particular line of longitude and not any of the others. A slightly cleaner version of the same sort of case exists if X and Y are two independent normally distributed variables with mean 0 and standard deviation of 1. Exercise 33.2 of Billingsley (1995) notes that conditioning on $X - Y = 0$ relative to the partition $\mathbf{X} - \mathbf{Y}$ gives different results from conditioning on $X/Y = 1$ relative to the partition $\mathbf{X}/\mathbf{Y}$. Example 6.1 on pp. 224–5 of Kadane, Schervish, and Seidenfeld (1986) considers the case where $Y = 0$ has been ruled out and notes that conditioning on $X = 0$ relative to the partition $\mathbf{X}$ gives different results from conditioning on $X/Y = 0$ relative to the partition $\mathbf{X}/\mathbf{Y}$.

8 Some have worried that the appeal to symmetry in the argument that $p(P \mid L_\theta)$ should be independent of θ is enough like the appeal to symmetry in the intuition that the conditional probability should be uniform that *both* are suspect. However, if we take the partition into account as part of the description of the problem, then there is a relevant difference. The unconditional probability is symmetric under *any* rotation of the sphere. However, the partition into lines of longitude is only symmetric under rotations of the sphere about the poles—rotating about any other point sends some lines of longitude to great circles that are not lines of longitude. In particular, rotation *along* any particular line of longitude changes the partition, so there is no need for probability conditional on this partition to preserve uniformity under this rotation. See p. 303 of Chang and Pollard (1997) for more discussion of this symmetry breaking.

allowing conditional probability to vary as the partition changes can avoid the worst consequences. And in fact it often can. As shown by Billingsley (1995, Theorem 33.3), if p is a probability function satisfying Countable Additivity over the events involving two random variables, then there is a way to specify the values for $p(B \mid A, \mathbf{A})$ while satisfying Conglomerability, where $\mathbf{A}$ is the partition of possible values of one variable, and B ranges over any proposition involving the two variables. In particular, this means that it is possible to give up on all forms of Additivity beyond Countable Additivity while holding on to Conglomerability.[9]

Thus, we have a choice between allowing conditional probability to be a three-place function $p(B \mid A, \mathbf{A})$ depending on a partition as well as a pair of events, and having unrestricted Conglomerability while only keeping Countable Additivity; or requiring conditional probability to be a two-place function $p(B \mid A)$ just of two events and keeping only as much Conglomerability as we do Additivity. The former option is called *Regular Conditional Probability*, while the latter is called *Coherent Conditional Probability*. ('Coherent' in this sense just means that the same pair of events has the same conditional probability regardless of what algebra it was drawn from, and is not related to the use of the word 'coherent' to mean "satisfying the probability axioms." I don't know where the term 'regular' comes from here, but it is not related to the concept requiring non-zero probabilities.) Mathematical theories of these two types will be the subjects, respectively, of Section 2 and Section 3.

Fuller consideration of the costs and benefits of these proposals will come in Section 2 and Section 3. But I will first mention several arguments for Conglomerability, which defenders of Coherent Conditional Probability must reject.

Recall that Conglomerability (Definition 9) says that for any partition $\mathbf{A}$, $\inf_{A \in \mathbf{A}} p(B \mid A) \leq p(B) \leq \sup_{A \in \mathbf{A}} p(B \mid A)$. By considering either B or its negation as needed, a violation means that there is some value x such that $p(B) < x$, but for every $A \in \mathbf{A}$, $p(B \mid A) > x$. If we consider the role of conditional probability in updating degrees of belief or in measuring confirmation, then this means that if one is about to perform an experiment whose possible outcomes are $\mathbf{A}$, then one can know in advance that one will get evidence confirming proposition B. This possibility seems intuitively costly for statistical or scientific reasoning, though there have been some attempts to mitigate it (Kadane, Schervish, & Seidenfeld, 1996).

For update via Jeffrey Conditionalization, Conglomerability is even more natural. Recall that update via Jeffrey Conditionalization proceeds by taking some partition $\mathbf{E}$ of possible evidence and updating one's old degrees

9 There are some other challenges to Conglomerability raised by Arntzenius, Elga, and Hawthorne (2004), but these also depend on changing partitions while keeping conditional probability fixed.

of belief $p(E)$ to new degrees of belief $p'(E)$ for all $E \in \mathbf{E}$. This then propagates through the rest of one's beliefs by means of "rigidity," the requirement that for any proposition A, we have $p'(A \mid E) = p(A \mid E)$. In the finite case, the Law of Total Probability tells us that $p'(A) = \sum_{E \in \mathbf{E}} p'(A \mid E)p'(E)$, and since these values are specified, so are the probabilities for all other propositions. In the infinite case, we need some version of the Law of Total Probability for this to generalize. The natural thought is that we should have $p'(A) = \int p'(A \mid \mathbf{E})\, \mathrm{d}p'$. But this just is the formulation of Disintegrability for p', which is equivalent to Conglomerability. Thus, giving up Conglomerability would require finding a new version of the Law of Total Probability that doesn't have these features, to use in defining Jeffrey Conditionalization.

Considering the role of conditional probability in decision theory, Conglomerability is also supported by a Dutch book argument. The basic idea is given by Billingsley (1995, p. 431). Basically, any sort of reasoning to a foregone conclusion (as violations of Conglomerability allow) will make for guaranteed changes in one's betting prices that can be exploited by someone who knows one's updating rule. Rescorla (2018) has given a more complete Dutch book argument, including converse theorems proving that Conglomerability suffices for immunity to this sort of Dutch book.

There is also an accuracy-based argument for Conglomerability. Some authors have suggested that the right way to think of degree of belief is as aiming at the truth. Once we have a reasonable notion of "accuracy" that measures closeness to the truth, we can then derive norms for degree of belief from principles of maximizing accuracy (Joyce, 1998; Greaves & Wallace, 2006; Pettigrew, 2016). As it turns out, an update plan for learning which member of a partition is true maximizes expected accuracy iff it satisfies Conglomerability with respect to that partition (Easwaran, 2013a).

None of these arguments is fully definitive. It is possible to reject the importance of Dutch books and accuracy conditions for degree of belief. It is conceivable that an alternative formulation of the Law of Total Probability allows for a generalization of Jeffrey Conditionalization (or that Jeffrey Conditionalization is not the right update rule). And perhaps reasoning to a foregone conclusion is not so bad for updating. And all of these problems are perhaps less bad for physical or chance interpretations of probability than for Bayesian interpretations of one sort or another. Thus, if it is very important that conditional probability really be a two-place function rather than depending on a partition as well, then there is motivation to pursue Coherent Conditional Probability.

Thus the question becomes just how bad the costs are of Regular Conditional Probabilities, with their extra parameter. Some have said that an event alone must be sufficient to determine a posterior probability distribution, and that the fact of the partition from which the event was

drawn can't be relevant. "This approach [Regular Conditional Probability] is unacceptable from the point of view of the statistician who, when given the information that $A = B$ has occurred, must determine the conditional distribution of X_2" (Kadane et al., 1986). This is most plausible for uses of conditional probability in update by conditionalization, where one just learns a new piece of information, and apparently doesn't learn anything about the partition from which this information was drawn.

However, I claim that by considering the situation in a bit more detail, there will always be a partition that is relevant in any application of conditional probability. Billingsley (1995, end of section 33) brings this out with a juxtaposition of three exercises. The first two exercises involve consideration of the Borel paradox with a point on the surface of a sphere, and a version involving two independent normally distributed random variables. The third exercise juxtaposes the effect in these exercises of the same information presented in two different ways (a great circle presented as one from the family of longitudes, or as the equator from a family of latitudes; the fact of two random variables being equal as a piece of information about their difference, or as a piece of information about their ratio) with a classic probability puzzle.

Three prisoners are in a cell and two will be executed in the morning. Prisoner 3 asks the guard to tell him which of 1 or 2 will be executed (since at least one of them will) and on hearing the answer reasons that his chance of survival has gone up from 1/3 (as one of three prisoners, two of whom will be executed) to 1/2 (as one of two prisoners, one of whom will be executed). But of course, as anyone who has considered the similar "Monty Hall" problem can recognize, this reasoning ignores the fact that "Prisoner 1 is executed" and "Prisoner 2 is executed" do not form a partition, since it is possible for both to be true. The relevant learning situation is one in which the partition is "The guard says prisoner 1 will be executed" and "The guard says prisoner 2 will be executed." If these two answers are equally likely conditional on prisoner 3 surviving, then in fact the probability of survival is unchanged by this update.

This sort of example shows that even in elementary cases, we need to be careful about only updating on evidence by conditionalization in cases where it is clear that the evidence is drawn from a partition. To properly take this into account, we must be able to figure out what partition the evidence was drawn from. For Jeffrey Conditionalization, the partition is in fact part of the specification of the update situation, so this is clearer. Thus, I claim that for the first two uses of Bayesian probability (update by conditionalization or Jeffrey Conditionalization) the partition relativity of Regular Conditional Probabilities is no problem. There are some authors who argue that update situations don't always involve evidence that comes from a partition (Schoenfield, 2016; Gallow, 2016). But I think that at least

for scientific cases where evidence comes as the result of the performance of an experiment, the partition is implicit in the experimental setup. This is especially so in cases where the evidence was something that antecedently had probability 0, which are the only cases in which the issue of how to conditionalize arises.

For the uses of conditional probability in the measurement of confirmation, we have to look both at posterior probabilities and likelihoods. That is, we should be looking at probabilities of hypotheses conditional on evidence (as for updating) and for probabilities of evidence conditional on hypotheses. In this case, because of the Problem of Old Evidence (presented by Glymour, 1980, and classified and investigated at length by Eells, 1985), we must be considering conditional probabilities given before the experiment is actually performed. In order to properly compare and contrast the effect of different possible pieces of evidence, or different experiments, on different hypotheses, we must have a sense of the possible experiments, the possible pieces of evidence they could result in, and the possible hypotheses under consideration. This is particularly clear in cases where we are interested in confirmation, disconfirmation, and independence of hypotheses about random variables rather than just single propositions. A scientist who is interested in measuring the value of some physical, social, or biological parameter is going to have a whole family of propositions about its value that each may be confirmed, disconfirmed, or independent of the evidence received, and this family will define a partition for the relevant likelihoods.

For decision-theoretic cases, the relevant conditional probabilities are probabilities of outcomes conditional on actions. Here again it seems plausible that the set of actions available to an agent forms a partition. If this is right, then the relativization to a partition just brings out a feature that is already important to the situation. Thus, just like with the other Bayesian applications of conditional probability, I claim that there is no problem to the three-place formulation of conditional probability required by Regular Conditional Probabilities.

Even once we see that conditional probability depends on the partition from which the conditioning event was drawn, we might worry about how the *description* of events and partitions can affect the actual value. Rescorla (2015) argues that we should think of the same event drawn from a different partition as having something like a different "sense," so that these are just Frege puzzles of a sort. I'm not convinced that this is the right way to understand things, because the difference in conditional probability persists even when everyone involved recognizes that the same conditioning event is a member of multiple partitions. But I think that some reasoning of this sort can dissolve some worries.

Some have also worried that by redescribing the probability space, we might be able to make one partition look like another, so that we can get conflicting requirements for the same conditional probability. But Gyenis, Hofer-Szabó, and Rédei (2016) show that this is impossible—any reparameterization of a set of events and a partition gives rise to some other description on which the mathematical requirements of Conglomerability and Disintegrability give the same results.

In addition to the obvious challenge in terms of relativization, there is also a question of whether Regular Conditional Probabilities require Countable Additivity. Classic results (such as the Radon–Nikodym Theorem, or Theorem 33.3 of Billingsley, 1995) show that when the propositions involved are just about random variables, relativization of conditional probability to a partition as well as a conditioning event is sufficient to allow Conglomerability to hold even when Additivity fails at uncountable cardinalities. However, every existence theorem I know of assumes Countable Additivity. I have not investigated the proofs of Countable Additivity from Countable Conglomerability in enough detail to be sure that they hold up when conditional probabilities are allowed to vary as the partition changes. Thus, if considerations like the de Finetti lottery motivate rejection of Countable Additivity, then there may be further problems for Regular Conditional Probabilities. But as I have argued elsewhere, there are independent reasons to accept Countable Additivity that don't generalize to higher cardinalities (Easwaran, 2013b).

As the reader can probably see, I favor Regular Conditional Probabilities over Coherent Conditional Probabilities. But in the remainder of the paper, I will put forward mathematical theories of both types so that the reader can judge for herself what the appropriate uses of each might be.

2 REGULAR CONDITIONAL PROBABILITIES

2.1 *Formal Theory*

Regular Conditional Probabilities are a central motivation for the Kolmogorov (1950) axiomatization of probability. There is some set Ω of "possibilities," and the bearers of probability are subsets of this set. (Different interpretations of probability will interpret these possibilities and sets of them differently.) Not every subset of the space of possibilities is a bearer of probability, but there is some collection $\mathcal{F}$ of them that are. $\mathcal{F}$ is assumed to be a "σ-algebra" or "σ-field," which means that the empty set is an element of $\mathcal{F}$, the complement of any element of $\mathcal{F}$ is an element of $\mathcal{F}$, and if A_i for $i \in \mathbb{N}$ are any countable collection of elements of $\mathcal{F}$, then $\bigcup_{i\in\mathbb{N}} A_i$ is also an element of $\mathcal{F}$. (This restriction to closure only under countable unions and complements is quite natural for the propositions

implicitly grasped by a finite mind, though one might want to restrict further to computably-definable sets or the like.)

Finally, there is a function p assigning real numbers to all and only the elements of $\mathcal{F}$ subject to the following rules. For any $A \in \mathcal{F}$, $p(A) \geq 0$; $p(\Omega) = 1$; and if A_i for $i \in \mathbb{N}$ are any countable collection of *disjoint* elements of $\mathcal{F}$, then $p(\bigcup_{i\in\mathbb{N}} A_i) = \sum_{i\in\mathbb{N}} p(A_i)$. That is, the probability function satisfies Countable Additivity. We refer to the triple $(\Omega, \mathcal{F}, p)$ as a *probability space*.

For any non-empty set Ω, there are of course multiple different σ-algebras of subsets of that space. Trivially, the set $\{\emptyset, \Omega\}$ is always the minimal σ-algebra on Ω, while the full power set consisting of *all* subsets of Ω is always the maximal σ-algebra on Ω. But usually, $\mathcal{F}$ is some algebra other than these two. We say that a set A is "$\mathcal{A}$-measurable" iff A is an element of $\mathcal{A}$. If $\mathcal{A}$ and $\mathcal{B}$ are any two σ-algebras on Ω, and every element of $\mathcal{A}$ is $\mathcal{B}$-measurable, then we say that $\mathcal{A}$ is a "sub-σ-algebra" of $\mathcal{B}$.

We often consider functions assigning a real number to every element of Ω. If V is such a function, then we say that V is a *random variable*, or that it is *$\mathcal{F}$-measurable*, iff for all rational values x, the set $\{\omega \in \Omega \colon V(\omega) < x\}$ is $\mathcal{F}$-measurable. The set $\{\omega \in \Omega \colon V(\omega) \in S\}$ is often just written as $V(\omega) \in S$ or even $V \in S$, so for V to be $\mathcal{F}$-measurable just is for $p(V < x)$ to exist for all rational values x, just as in Section 1.4.2. Furthermore, since the rational values are a countable and dense subset of the real numbers, the fact that $\mathcal{F}$ is closed under countable unions and complements means that $p(V = x)$, $p(V \geq x)$ and any other probability simply expressible in terms of values of V exist as well.

As in Section 1.4.2, we can define the integral $\int_A V \, \mathrm{d}p$ for bounded random variables V. This definition proceeds in two parts. If V only takes finitely many values on points in A, we say that $\int_A V \, \mathrm{d}p = \sum x \cdot p(A \cap (V = x))$, where the sum ranges over the finitely many values that V takes on. Otherwise, we define $\int_A V \, \mathrm{d}p = \sup_{V' < V} \int_A V' \, \mathrm{d}p$, where the supremum ranges over all random variables V' that take on only finitely many values in A, and such that whenever $\omega \in A$, $V'(\omega) < V(\omega)$.

With these definitions, I can finally give the official definition of a Regular Conditional Probability.

Definition 10 *A* Regular Conditional Probability *is a three-place real-valued function $p(B \mid \mathcal{A})(\omega)$ satisfying the following three conditions:*

1. *Fixing a σ-algebra $\mathcal{A} \subseteq \mathcal{F}$ and $\omega \in \Omega$ defines a function of B satisfying the probability axioms (that is, it is non-negative for all $B \in \mathcal{F}$, it takes the value 1 when $B = \Omega$, and it is Countably Additive).*

2. *Fixing a σ-algebra $\mathcal{A} \subseteq \mathcal{F}$ and a measurable set B defines an $\mathcal{A}$-measurable function of ω.*

3. *For any fixed σ-algebra $\mathcal{A} \subseteq \mathcal{F}$ and a measurable set B, and for $A \in \mathcal{A}$,*

$$\int_A p(B \mid \mathcal{A})(\omega)\, \mathrm{d}p = p(B \cap A).$$

In Section 2.2 I will discuss how this notion relates to the three-place function $p(B \mid A, \mathbf{A})$ of conditional probability mentioned earlier. The basic idea of each condition is as follows. Condition 1 will ensure that conditioning on a single event relative to a single partition yields a probability function. Condition 2 will ensure that we really are conditioning on an event A from the partition $\mathbf{A}$. Condition 3 will ensure that $p(B \mid A, \mathbf{A})$ satisfies Disintegrability (and thus Conglomerability). But for now I will just discuss a few formal features this mathematical function has.

As a first example, consider a probability space defined by a joint probability density for two random variables. That is, we can consider X and Y as two random variables, and let $\Omega = \mathbb{R}^2$, where the element $\omega = (\omega_X, \omega_Y)$ of Ω represents the possibility of $X = \omega_X$ and $Y = \omega_Y$. $\mathcal{F}$ is the σ-algebra generated by the set of sets $X < x$ and $Y < y$. (This algebra is known as the collection of "Borel sets," which is a subset of the Lebesgue-measurable sets, but sufficient for our purposes.) To say that the probability is defined by a joint probability density means that there is a measurable function $d(x, y)$ such that

$$p((x_1 < X < x_2) \cap (y_1 < Y < y_2)) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} d(x, y)\, \mathrm{d}y\, \mathrm{d}x,$$

where the integrals here are ordinary real-valued integrals. (This definition of probability over the rectangular boxes suffices to determine the probability of every measurable set.)[10]

If $\mathcal{X}$ is the σ-algebra generated by the set of sets $X < x$, then we can define a Regular Conditional Probability $p(B \mid \mathcal{X})(\omega)$ as follows. Let

$$p((x_1 < X < x_2) \cap (y_1 < Y < y_2) \mid \mathcal{X})(\omega) = \frac{\int_{y_1}^{y_2} d(\omega_X, y)\, \mathrm{d}y}{\int_{-\infty}^{\infty} d(\omega_X, y)\, \mathrm{d}y},$$

if $x_1 < \omega_X < x_2$ and 0 otherwise. (I use ω_X to represent the fixed value X takes at ω, while I use y as the bound variable of the integral.) Again, because the rectangles $(x_1 < X < x_2) \cap (y_1 < Y < y_2)$ generate the whole σ-algebra, this suffices to define the conditional probability $p(B \mid \mathcal{X})(\omega)$ for all measurable sets B. Note that the values y_1 and y_2 enter on the right as limits of an integral, while the values x_1 and x_2 just determine when

10 Note that since we have assumed there is an unconditional probability function p, then we have assumed that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} d(x, y)\, \mathrm{d}y\, \mathrm{d}x = 1$. In Section 3.2.5, when discussing Rényi's theory of conditional probability, I will allow this integral to be infinite instead, to capture the statistical theory of "improper priors."

the probability is 0. This is because the point (ω_X, ω_Y) with respect to the σ-algebra $\mathcal{X}$ represents the set of all points with $X = \omega_X$ and any value of Y, and the rectangle either intersects this line at all points from y_1 to y_2 or none of them. Intuitively, the numerator of the right side says how much density is concentrated at $y_1 < Y < y_2$ and $X = \omega_X$, while the denominator normalizes this to account for how much density is at $X = \omega_X$ generally. It is tedious, but possible to check that this definition satisfies the three conditions to be a Regular Conditional Probability.[11]

The Borel paradox can be thought of as a special case of this example. If X represents the longitude (from $-\pi$ to π) and Y represents the latitude (from $-\pi/2$ to $\pi/2$), then the uniform unconditional probability is given by the density function $d(x, y) = \frac{\cos y}{4\pi}$ when $-\pi < x < \pi$ and $-\pi/2 < y < \pi/2$, and 0 otherwise. Using the above formula, we calculate that

$$p(y_1 < Y < y_2) \mid \mathcal{X})(\omega) = \frac{\int_{y_1}^{y_2} \frac{\cos y}{4\pi} dy}{1/2\pi} = \frac{\sin y_2 - \sin y_1}{2}.$$

By parallel reasoning, we calculate that

$$p(x_1 < X < x_2) \mid \mathcal{Y})(\omega) = \frac{\int_{x_1}^{x_2} \frac{\cos \omega_Y}{4\pi} \, \mathrm{d}x}{\cos \omega_Y / 2} = \frac{x_2 - x_1}{2\pi}.$$

That is, conditional on lines of longitude, probability is concentrated near the equator, while conditional on lines of latitude, probability is uniform.

If we want to use this sort of technique to figure out other Regular Conditional Probabilities for other sub-σ-algebras, we can often do this, if the new algebra is related to the old one by a change of coordinates. This will work if the probability space is defined by two random variables X and Y, and there are two other random variables f_1 and f_2, such that the values of f_1 and f_2 are uniquely determined by the values of X and Y, and vice versa. For instance, we might have $f_1 = X - Y$ and $f_2 = Y$, or $f_1 = X/Y$ and $f_2 = Y$ (if $Y = 0$ is impossible), or f_1 and f_2 as latitude and longitude in a different set of coordinates than X and Y. In such a case, we can consider f_1 and f_2 as functions of the values of X and Y, and represent points in Ω not as (ω_X, ω_Y), but as $(f_1(\omega_X, \omega_Y), f_2(\omega_X, \omega_Y))$.

Assuming the functions f_1 and f_2 are measurable, we get a new density function given by

$$d(\omega_X, \omega_Y) \cdot \left[\frac{\partial f_1(x, \omega_Y)}{\partial x} \frac{\partial f_2(\omega_X, y)}{\partial y} - \frac{\partial f_2(x, \omega_Y)}{\partial x} \frac{\partial f_1(\omega_X, y)}{\partial y} \right].$$

11 To check the third condition, it's useful to note that the $A \in \mathcal{X}$ are generated by the sets $x_1 < X < x_2$, and the probability of these sets is given by integrals like the denominator of the right-hand-side, so that this denominator cancels in the integration, leaving just the integral of the numerator over X, which is how we defined the unconditional probability in the first place.

This quantity on the right is the Jacobian associated with the relevant change of variables. When $f_1(\omega_X, \omega_Y) = \omega_X$ and $f_2(\omega_X, \omega_Y) = \omega_Y$, so that the "new" variables are the same as the old, the Jacobian is equal to 1, so the density is unchanged, as expected. But the fact that this Jacobian is not generally equal to 1 indicates that corresponding points in the two representations of the probability space will have different densities with respect to the two different sets of variables. Thus, even if one value of one variable occurs exactly when a corresponding value of a different variable occurs (such as $X = 0$ occurring iff $X/Y = 0$, or latitude is 0 in one set of coordinates iff longitude is 0 in another set of coordinates), the densities may have been transformed in some non-uniform way, so the Regular Conditional Probability may take different values.

A slightly different introduction to this sort of method is discussed by Chang and Pollard (1997). They argue that in most cases where Regular Conditional Probabilities are of interest, they can be calculated by a method like this one. Although their discussion is still quite technical, it may be more usable and friendly than some others.

2.2 *Philosophical Application*

As before, I define a "partition" to be a collection $\mathbf{A}$ of subsets of Ω such that every member of Ω is in exactly one member of $\mathbf{A}$. In Section 1.4.3, I argued that in order to maintain Conglomerability, while respecting the roles of conditional probability as posterior for conditionalization, or Jeffrey update, or as likelihood, or as action probability for decision theory, we need a notion of conditional probability that defines $p(B \mid A, \mathbf{A})$ whenever $\mathbf{A}$ is a partition. However, the formal theory given above defined a random variable $p(B, \mathcal{A})(\omega)$, where $\mathcal{A}$ is a sub-σ-algebra rather than a partition, and where ω is an element of Ω rather than a subset of it. In this section, I show that the formal definition of a Regular Conditional Probability is sufficient to give us what we need.

Partitions can be related to σ-algebras in two importantly different ways. One is that we can say that a σ-algebra $\mathcal{B}$ is *generated* by a partition if it is the *smallest* σ-algebra with respect to which every element of $\mathbf{A}$ is measurable. In this case, $\mathcal{B}$ consists of the set of all unions of countably many elements of $\mathbf{A}$, and their complements.[12] However, in many cases, the more useful σ-algebra to consider is a slightly different one. I will say that a σ-algebra $\mathcal{B}$ is *compatible* with a partition $\mathbf{A}$ iff every element of $\mathbf{A}$ is

12 We also talk about σ-algebras generated by collections of subsets other than a partition, and in those cases there can often be much more complex elements of the generated σ-algebra, such as countable unions of complements of countable unions of complements of countable unions of elements. But in the case of a partition, these more complex elements already exist just at the level of countable unions or their complements.

an element of $\mathcal{B}$, and no proper subset of an element of $\mathbf{A}$ is an element of $\mathcal{B}$, except for the empty set.[13] Then, if $\mathcal{B}$ is any σ-algebra and $\mathbf{A}$ is any partition, I will say that the *restriction of $\mathcal{B}$ to $\mathbf{A}$* is the *largest* sub-σ-algebra of $\mathcal{B}$ that is compatible with $\mathbf{A}$. This consists of all elements of $\mathcal{B}$ whose intersection with any element of $\mathbf{A}$ is either empty or the full element of $\mathbf{A}$—it is the set of all $\mathcal{B}$-measurable sets that don't crosscut any element of $\mathbf{A}$.

Given these definitions, for $A, B \in \mathcal{F}$ and $\mathbf{A} \subseteq \mathcal{F}$ a partition containing A, I will define $p(B \mid A, \mathbf{A})$ as $p(B \mid \mathcal{A})(\omega)$, where ω is any element of A and $\mathcal{A}$ is the restriction[14] of $\mathcal{F}$ to $\mathbf{A}$. If A is empty, then $p(B \mid A, \mathbf{A})$ is undefined. This corresponds to the fact that conditional probability is intended to be an indicative conditional for updating rather than revision of beliefs, as discussed in Section 1.3. Otherwise, since $p(B \mid \mathcal{A})(\omega)$, considered as a function of ω, is required to be $\mathcal{A}$-measurable, it must be constant on the atoms of $\mathcal{A}$. But because $\mathcal{A}$ is the restriction of $\mathcal{F}$ to $\mathbf{A}$, the atoms are the elements of $\mathbf{A}$. Since A is an element of $\mathbf{A}$, this means that it doesn't matter which $\omega \in A$ is chosen. Thus, as long as $p(B \mid \mathcal{A})(\omega)$ is a well-defined function, so is $p(B \mid A, \mathbf{A})$, whenever A is non-empty. The stipulations in the definition of a Regular Conditional Probability then mean that $p(B \mid A, \mathbf{A})$ satisfies the probability axioms (including Countable Additivity) when A and $\mathbf{A}$ are fixed, and that Conglomerability is satisfied over $\mathbf{A}$. Thus, if conditional probability should be defined relative to any partition, and Conglomerability must be satisfied, then conditional probability must be related to a Regular Conditional Probability in this way.

2.3 *Existence and Uniqueness of Regular Conditional Probabilities*

The question motivated by the arguments of Section 1.3 is whether unconditional probabilities suffice to determine a notion of conditional probability, or whether conditional probability should be taken as fundamental. The mathematical definition of a Regular Conditional Probability as $p(B \mid \mathcal{A})(\omega)$ is as a function that satisfies some axioms connecting it to the unconditional probability space $(\Omega, \mathcal{F}, p)$. In some cases, we have been able to demonstrate that Regular Conditional Probabilities exist. If they don't exist in probability spaces that are philosophically important, then

13 In more standard terminology, $\mathbf{A}$ consists of the "atoms" of $\mathcal{B}$, where an atom of a σ-algebra is any non-empty element of the σ-algebra such that no non-empty proper subsets are also members of the σ-algebra. Not every σ-algebra has atoms, but if there are any atoms, they are disjoint. The atoms form a partition iff every element of the space is a member of some atom, in which case the σ-algebra is said to be "atomic."

14 Section 2.3.2 will show what goes wrong if we try to use the sub-σ-algebra generated by $\mathbf{A}$ instead of the restriction to it.

Conglomerability must be given up. And if Regular Conditional Probabilities are not unique, then we must either accept that conditional probability is at least as fundamental as unconditional probability, or give some further conditions that suffice to uniquely determine the Regular Conditional Probability uniquely. In this section I will consider some mathematical problems of particular Regular Conditional Probabilities and argue that they don't arise in philosophical application, so they will always exist and have the desired features. Furthermore, I will show that unconditional probability is almost sufficient to define all conditional probabilities in the relevant probability spaces, and give some ideas of what else might suffice to define conditional probability uniquely from unconditional probability.

2.3.1 *In Bad Sub-σ-algebras There Is No Regular Conditional Probability*

It is mathematically well-known that there are probability spaces $(\Omega, \mathcal{F}, p)$ and sub-σ-algebras $\mathcal{A}$ for which there is no Regular Conditional Probability. A classic example is the case where Ω is the set $[0,1]$ of real numbers between 0 and 1, $\mathcal{A}$ is the set of all Borel subsets of this set, $\mathcal{F}$ is generated by $\mathcal{A}$ plus one set that is not Lebesgue-measurable, and p is Lebesgue measure on $\mathcal{A}$ and assigns probability $1/2$ to the additional set generating $\mathcal{A}$. (This example is discussed in Billingsley, 1995, Exercise 33.11.)

However, Theorem 33.3 of Billingsley (1995) states that when $\mathcal{F}$ is the σ-algebra generated by the values of a random variable, this problem can never arise. There will always be a Regular Conditional Probability for every sub-σ-algebra. This result generalizes to cases where $\mathcal{F}$ is the σ-algebra generated by the values of finitely many random variables, as appears to be the case for most scientific applications of probability.

Furthermore, due to the finitistic limits of the human mind, I claim that this in fact includes all epistemically relevant cases. As I suggested near the end of Section 1.3, I think the right interpretation of human finitude doesn't mean that the probability space is finite. Rather, it means that the probability space is generated by the countably many sentences of some finitary language. I claim that the sentences in this language fit within the σ-algebra over this space generated by a particular artificial random variable.

To see this, define the random variable T by enumerating the sentences of the language as ϕ_i and letting

$$T(\omega) = \sum_{\phi_i \text{ is true}} \frac{1}{2^i}.$$

Any possibility ω will make infinitely many sentences true and infinitely many sentences false, and no two such possibilities can result in the same real value, so this random variable distinguishes all possible worlds. We

need to check further that the set of values that are logically consistent is itself measurable. But by the Compactness Theorem of first-order logic, any logically inconsistent set contains one of the countably many logically inconsistent finite sets, and each of these sets is an intersection of finitely many closed sets of values. Thus, the set of consistent values is the complement of a countable union of closed sets, and is thus measurable. Thus, I claim that any epistemically reasonable probability space uses a σ-algebra generated by a random variable, conditionalized on a measurable set. Thus, Theorem 33.3 of Billingsley (1995) entails that Regular Conditional Probabilities exist.

Even without this sort of argument, the existence theorem can be generalized. These generalizations are investigated by Hoffmann-Jørgensen (1971), Faden (1985), Pachl (1978).

2.3.2 *In Bad Sub-algebras, the Regular Conditional Probability Behaves Badly*

Another problem that sometimes arises is highlighted by Blackwell and Dubins (1975) and Seidenfeld, Schervish, and Kadane (2001). They seem to show that in certain partitions $\mathbf{A}$, there is an event A with $p(A \mid A, \mathbf{A}) = 0$, which would seem to be quite bad. However, I claim that this problem only arises in cases where $\mathbf{A}$ is used in a mathematically improper way.

The mathematical result they show is that $p(B \mid \mathcal{A})(\omega) = 0$ even though $\omega \in B$. As an example, let Ω be the set $[0,1]$ of real numbers between 0 and 1, let $\mathcal{F}$ be the collection of all Borel subsets of this set, and let p be the standard Lebesgue measure on $\mathcal{F}$. Let $\mathcal{A}$ be the collection of all countable subsets of $[0,1]$ and their complements. It is straightforward to check that $p(B \mid \mathcal{A})(\omega) = p(B)$ is a Regular Conditional Probability.[15] However, if $B = \{\omega\}$ (or any other countable set containing ω) then $p(B \mid \mathcal{A})(\omega) = p(B) = 0$. Given my translation of $p(B \mid A, \mathbf{A})$, this would seem to mean that $p(\{\omega\} \mid \{\omega\}, \mathbf{A}) = 0$, where $\mathbf{A}$ is the partition into singletons.

However, this is the point at which the distinction between the σ-algebra *generated* by $\mathbf{A}$ and the *restriction* of $\mathcal{F}$ to $\mathbf{A}$ is important. The σ-algebra $\mathcal{A}$ above is the algebra generated by the partition into singletons, but it is *not* the restriction of $\mathcal{F}$ to the partition into singletons. The restriction of $\mathcal{F}$ to the partition into singletons just is $\mathcal{F}$ (as it is for any $\mathcal{F}$—recall that the restriction of $\mathcal{F}$ includes all elements of $\mathcal{F}$ that do not crosscut any element of the partition, and no set crosscuts a singleton). Although $p(B \mid \mathcal{A})(\omega) = p(B)$ is a Regular Conditional Probability, it is straightforward to show that the parallel does not work for $p(B \mid \mathcal{F})(\omega)$. In fact, any Regular Conditional

15 The first two conditions are trivial. The third condition requires that $\int_A p(B \mid \mathcal{A})(\omega)\, \mathrm{d}p = p(A \cap B)$ for all $A \in \mathcal{A}$. However, since $p(B \mid \mathcal{A})(\omega) = p(B)$ for all ω, the left side of the integral just is $p(A)p(B)$. But if A is countable, then $p(A) = 0$, as does $p(A \cap B)$, while if A's complement is countable, then $p(A) = 1$ and $p(A \cap B) = p(B)$.

Probability for this conditioning algebra must have a set C with $p(C) = 1$ such that whenever $\omega \in C$, $p(B \mid \mathcal{F})(\omega) = 1$ if $\omega \in B$ and 0 otherwise, as expected. And Theorem 2 of Blackwell and Dubins (1975) and Theorem 1 of Seidenfeld et al. (2001) show that this is quite general. Whenever $\mathcal{A}$ is countably generated, for any Regular Conditional Probability $p(B \mid \mathcal{A})(\omega)$, there is a set C with $p(C) = 1$ such that whenever $\omega \in C$ and $B \in \mathcal{A}$, $p(B \mid \mathcal{A})(\omega) = 1$.[16] Thus, in my translation, $p(B \mid A, \mathbf{A}) = 1$ if $A \subseteq B$, as expected, whenever the restriction of $\mathcal{F}$ to $\mathbf{A}$ is countably generated. This will automatically be the case if $\mathbf{A}$ is the partition of possible values of a random variable. But I claim that it should hold generally for any partition that is graspable by a finite human mind.

2.3.3 *The Regular Conditional Probability is* Almost *Unique*

Now that we have established that Regular Conditional Probabilities exist and are well-behaved, it remains to see when they are uniquely determined by the unconditional probability space $(\Omega, \mathcal{F}, p)$. It turns out that the answer is *never* in any interesting case. However, the different Regular Conditional Probabilities that exist are *almost* identical in a natural sense. Furthermore, for some sets of niceness conditions, exactly one of them will be nice, and this can be designated as the correct one.

If $p(B \mid \mathcal{A})(\omega)$ is one Regular Conditional Probability, and $S \in \mathcal{F}$ is any set with $p(S) = 0$, then we can let $p'(B \mid \mathcal{A})(\omega) = p(B \mid \mathcal{A})(\omega)$ whenever $\omega \notin S$ and replace the function with any other probability function we like within S, and the result is also a Regular Conditional Probability. This is because the only constraint on the values of a Regular Conditional Probability are through its integrals, and changing a function on a set of probability 0 does not change any of its integrals. Translating to $p(B \mid A, \mathbf{A})$, this means that we can change the values of the conditional probability function on any collection of $A \in \mathbf{A}$ whose total probability is 0 and still satisfy Conglomerability.

Conversely, if $p(B \mid \mathcal{A})(\omega)$ and $p'(B \mid \mathcal{A})(\omega)$ are two Regular Conditional Probabilities for a given unconditional probability, then we can show that for any B and $\mathcal{A}$, the set of ω for which they differ must have probability 0. If it had positive probability, then there would be some ϵ such that the set C of ω on which they differ by at least ϵ would have positive probability, and would be a member of $\mathcal{A}$. But this would contradict the condition that $\int_C p(B \mid \mathcal{A})(\omega)\, \mathrm{d}p = p(B \cap C) = \int_C p'(B \mid \mathcal{A})(\omega)\, \mathrm{d}p$. Thus,

16 Of course, this assumes that a Regular Conditional Probability exists, which requires that $\mathcal{F}$ be a nice algebra, such as the algebra generated by a random variable. See Blackwell (1956) for more on these conditions. In fact, for these sorts of spaces, Yu (1990) proves that existence of the relevant function can be proven in the system "ACA_0" of reverse mathematics, so that strong set-theoretic hypotheses like the Axiom of Choice are not required.

although the Regular Conditional Probability is not exactly unique, it is in a sense "almost" unique. These different Regular Conditional Probabilities are often called "versions" of the Regular Conditional Probability for the given unconditional probability.

This almost uniqueness is not quite enough to satisfy the idea that conditional probability is defined by the unconditional probability function. However, in some cases there is a prospect that by specifying a further condition, we can pick out a unique version of the Regular Conditional Probability. For instance, consider the case of the Borel paradox. As I showed in Section 2.1, one version of the Regular Conditional Probability for this example can be generated by integrals of a probability density that also generates the unconditional probability. In this case, there is a *continuous* density function that generates the unconditional probability (namely, the density function that was given there, with $d(x, y) = \cos y$). Furthermore, it is easy to see that no other continuous density generates the same unconditional probability function. (If two continuous density functions differ at some point, then they must differ on some neighborhood of that point, which would have non-zero probability.) Thus, if an unconditional probability function is generated by some continuous density on the values of some random variables, then we can require that the version of the Regular Conditional Probability used be the one that is generated by this integral calculation from the unique continuous density that generates the unconditional probability.[17]

17 Oddly, if we just consider the partitions into longitudes through various choices of poles, we may be able to take advantage of this non-uniqueness to find a *Coherent* Conditional Probability that satisfies Disintegrability. If we assume the Axiom of Choice and the Continuum Hypothesis (or Martin's Axiom—both assumptions entail that the union of any collection of fewer than continuum-many sets with probability 0 is also a set of probability 0), then we can do the following. Choose some well-ordering of the points on the sphere such that each has fewer than continuum-many predecessors. For any great circle A, find the point $x \in A$ that comes earliest in this ordering. Let $p(B \mid A)$ take the value given by integration with respect to the continuous density where x is chosen as the north pole of the coordinate system.
Now if we consider any particular partition into longitudes with x as a pole, we can see that each line of longitude will give rise to a conditional probability that agrees with the one required for Disintegrability in this partition iff there is no point on the line earlier than x in the chosen ordering. However, because of the way the ordering was set up, there are fewer than continuum-many points earlier than x in the ordering, so the union of all the lines of longitude that contain such a point has probability 0. Thus, enough of the conditional probabilities agree with integration with respect to the relevant continuous density that Disintegrability is satisfied in this partition.
Of course, this particular method only satisfies Disintegrability over partitions into lines of longitude, and not into lines of latitude, or other partitions. Furthermore, the particular Coherent Conditional Probability produced over these conditioning events is highly asymmetrical and requires the Axiom of Choice for its construction. But it is useful to observe that this sort of construction is at least sometimes possible.

However, while I think it is not that implausible to think that all realistic epistemic spaces are generated by some density on the values of some random variables, I don't see any good reason to believe that there must always be a *continuous* density function that generates the unconditional probability. Perhaps there is some similar requirement that could be used to find the "right" Regular Conditional Probability to go along with any unconditional probability function. But I have no idea what that requirement might be. So for now, we have some reason to believe that the existence of uncountable (though countably generated) probability spaces, together with Conglomerability, force us to use Regular Conditional Probabilities, which suggests that conditional probability is in some sense at least as fundamental as unconditional probability. However, if one is only given the unconditional probability function, then for any countably-generated partition $\mathbf{A}$ one can find *some* Regular Conditional Probability $p(B \mid A, \mathbf{A})$ for all propositions B on the elements of $\mathbf{A}$, and one can be sure that *almost all* of the values given by this function will line up with the "correct" conditional probability function. The question is just whether this "almost all" can be turned into "all," or whether conditional probability needs to be specified along with unconditional probability in defining a probability space.

3 COHERENT CONDITIONAL PROBABILITIES

Recall that Coherent Conditional Probability is conditional probability defined as a function just of two events, with no dependence on a partition or sub-σ-algebra or anything else. If Additivity fails at some level (possibly beyond the countable), then Conglomerability and Disintegrability will also fail. There are several different formal theories of Coherent Conditional Probability that have been proposed by philosophers, mathematicians, and statisticians. In this section I will describe three of the most prominent ones.

3.1 *Popper*

The first, which is both oldest and probably most familiar to philosophers, was developed by Karl Popper in his (1955). Popper considered this formulation of conditional probability important enough that he included a revised and simplified version in new appendices *iv and *v to the second edition of *The Logic of Scientific Discovery* (1959a). Popper's axiom system is particularly well-suited to an interpretation of probability as a logical (or even semantic) relation. But I claim that it is not sufficient for general epistemological applications, particularly for scientific purposes.

In this section I will describe Popper's later version of the system, and the features it has.

Popper postulates a finite or countable set of sentence letters $A, B, C, \ldots$, and two uninterpreted connectives—a binary connective '$\wedge$' and a unary connective '$\neg$'. (I have replaced his notation with a more modern one.) He then postulates a two-place conditional probability function mapping pairs of formulas in the language generated by these letters and connectives to real numbers. He then postulates six conditions on the function expressible with these uninterpreted connectives. (I will discuss these conditions later.) Finally, he defines unconditional probability in terms of conditional probability.

One of the important things Popper does along the way is to develop a probabilistic notion of equivalence. He says that two formulas ϕ and ψ of the language are probabilistically equivalent iff replacing ϕ with ψ anywhere in any statement of probability will yield the same value. He then proves that if two formulas are classically logically equivalent, then they are probabilistically equivalent. He doesn't explicitly assume commutativity and associativity for $\wedge$, or the double negation rule, or anything of that sort, but is able to derive probabilistic equivalents of them from his probability axioms.

Popper's axioms entail that some elements ψ are such that for all ϕ, $p(\phi \mid \psi) = 1$. (Among other things, this means that $p(\neg\psi \mid \psi) = 1$!) Following van Fraassen (1976), we call such elements *abnormal* and all others *normal*. Popper's axioms entail that if χ is normal, then $0 \leq p(\phi \mid \chi) \leq 1$, and that $p(\phi \mid \chi) + p(\psi \mid \chi) = p(\neg(\neg\phi \wedge \neg\psi) \mid \chi) + p(\phi \wedge \psi \mid \chi)$, so that conditional on any normal event, we have a standard probability function. Furthermore, they entail that if ψ is abnormal, then for any χ, $p(\neg\psi \mid \chi) = 1$. Finally, they entail that whenever ϕ is a classical logical contradiction, ϕ is abnormal.

Importantly, this means that Popper's notion of conditional probability (like all the others I am aware of) is of no help in using conditionalization to represent belief *revision* rather than just update. Consider an update rule that says $p_{t'}(\phi \mid \psi) = p_t(\phi \mid \psi \wedge \chi)$, where χ is the conjunction of everything that one has learned between t and t'. Now imagine a person who, between time 0 and time 1 learns A, and between time 1 and time 2 learns $\neg A$. If update can include revision of past learning (which implicitly means that learning is fallible), then this should result in something reasonable. However, what we see is that for any ϕ and ψ, $p_2(\phi \mid \psi) = p_1(\phi \mid \psi \wedge \neg A) = p_0(\phi \mid (\psi \wedge \neg A) \wedge A)$. But since $(\psi \wedge \neg A) \wedge A$ is a contradiction, it is abnormal. Thus, $p_0(\phi \mid (\psi \wedge \neg A) \wedge A) = 1$. So by updating on the negation of something that one previously learned, one's degrees of belief have become unusable, because all probabilities are equal to 1. This is why I focused in Section 1.3 on the role of infinity in generating events of

probability 0, rather than Hájek's examples of conditionalizing on the negation of something that has already been learned.

However, one important thing to note for Popper's system is that $p(\psi) = 0$ does *not* entail that ψ is abnormal. However, if $p(\psi) = 0$ but ψ is normal, then unconditional probabilities alone do not suffice to determine the probabilities conditional on ψ. Thus, conditional probability really is primitive in this system. For instance, consider models of Popper's axioms with sentence letters $A \wedge B$, with $p(A) = 1/2$ and $p(B) = 0$. Every formula of the language is classically equivalent to a contradiction, or to a disjunction of some of $A \wedge B$, $A \wedge \neg B$, $\neg A \wedge B$, $\neg A \wedge \neg B$. The stipulated values determine all the unconditional probabilities, and thus all the probabilities conditional on formulas of positive unconditional probability. However, it is consistent with these values that $A \wedge B$ and $\neg A \wedge B$ be either normal or abnormal. If both are abnormal, then so is B, and probabilities conditional on any of the three of them are all equal to 1. If one is abnormal and the other is normal, then probabilities of any formula conditional on the normal one are 1 or 0 depending on whether the formula is entailed by it or not. If both are normal, then any value for $p(A \mid B)$ is possible, but this value then suffices to determine the rest of the probabilities in the model.

And in fact, Kemeny (1955) proves that something like this holds fairly generally for finite languages. If we only have n sentence letters, then there are 2^n "state descriptions" in the language (conjunctions of each sentence letter or its negation), and every formula is either a contradiction or equivalent to a disjunction of some of these. The Popper axioms are then equivalent to the following stipulation. There are k functions m_i for $i < k$, and each of these functions assign a non-negative real number to each state description. For each m_i, the sum of the values it assigns to the state descriptions is 1. For each state description X, there is at most one m_i such that $m_i(X) > 0$. A proposition is abnormal iff it is either a contradiction, or it is a disjunction of state descriptions that are assigned value 0 by every m_i. If ψ is normal, then let i be the lowest number such that there is a state description X with $m_i(X) > 0$ and X entails ψ. Then

$$p(\phi \mid \psi) = \frac{\sum_{X \text{ entails } \phi \wedge \psi} m_i(X)}{\sum_{X \text{ entails } \psi} m_i(X)}.$$

In this system, unconditional probabilities are just equal to the sums of the values of m_1, but they put no constraints on the values of the succeeding functions, which are needed to define the full conditional probability function.

For infinite languages, things can be slightly more complicated. Consider a language with sentence letters A_i for natural numbers i. Consider just the models M_i where M_i satisfies sentence A_i and none of the others. It

is not hard to check that every formula of the language is either true in finitely many of these models and false in the rest, or false in finitely many of these models and true in the rest. If ψ is true in infinitely many models, then let $p(\phi \mid \psi) = 0$ if ϕ is true in finitely many models and 1 otherwise. If ψ is true in none of these models, then ψ is abnormal. Otherwise, if ψ is true in finitely many models, then define $p(\phi \mid \psi)$ as the ratio of the number of models in which $\phi \wedge \psi$ is true to the number of models in which ψ is true. This definition satisfies Popper's axioms, but cannot be represented by a lexicographically ordered set of probability functions as Kemeny shows in the finite case. (This example is one that Halpern, 2009 attributes to Stalnaker.) Halpern also discusses a slight variant of this case where the probability function agrees with this one in all cases except where ψ is true in finitely many models. In the variant, $p(\phi \mid \psi) = 1$ if ϕ is true in the *highest numbered* model in which ψ is true, and 0 otherwise. This probability function also satisfies Popper's axioms but cannot be represented by a lexicographically ordered set of probability functions. But again, these functions have the same unconditional probabilities and the same abnormal propositions, but different conditional probabilities, so that conditional probability must be specified separately from unconditional probabilities.

Popper's six conditions are the following (Popper, 1959a, Appendix iv*).

1. For all ϕ, ψ there are χ, θ with $p(\phi \mid \psi) \neq p(\chi \mid \theta)$.
2. If for all χ, $p(\phi \mid \chi) = p(\psi \mid \chi)$, then for all θ, $p(\theta \mid \phi) = p(\theta \mid \psi)$.
3. For all ϕ, ψ, $p(\phi \mid \phi) = p(\psi \mid \psi)$.
4. $p(\phi \wedge \psi \mid \chi) \leq p(\phi \mid \chi)$.
5. $p(\phi \wedge \psi \mid \chi) = p(\phi \mid \psi \wedge \chi) p(\psi \mid \chi)$.
6. For all ϕ, ψ, either $p(\phi \mid \psi) + p(\neg\phi \mid \psi) = p(\psi \mid \psi)$, or for all χ, $p(\psi \mid \psi) = p(\chi \mid \psi)$.

In Appendix v* of Popper (1959a), he derives a sequence of consequences of these postulates. Importantly, he doesn't assume any logical features of $\wedge$ and $\neg$ in these derivations—he only uses the explicit probabilistic assumptions made above.

First, using condition 3, he defines $k = p(\phi \mid \phi)$ for any formula ϕ. Using 4 and 5 he then proves that $k^2 \leq k$, so $0 \leq k \leq 1$. After a few more steps, he then proves that $0 \leq p(\phi \mid \psi) \leq k$ for any ϕ, ψ. From this, he is then able to derive that $k = k^2$, so $k = 0$ or $k = 1$, but condition 1 rules out $k = 0$. Condition 4 then tells us that $1 = p(\phi \wedge \psi \mid \phi \wedge \psi) \leq p(\phi \mid \phi \wedge \psi)$, so $p(\phi \mid \phi \wedge \psi) = 1$. With condition 5 this then proves that $p(\phi \wedge \phi \mid \psi) = p(\phi \mid \psi)$. A bit more manipulation allows him to derive that

$p(\phi \wedge \psi \mid \chi) = p(\psi \wedge \phi \mid \chi)$, and that $p(\phi \wedge (\psi \wedge \chi) \mid (\phi \wedge \psi) \wedge \chi) = 1$, and after several more steps, that $p(\phi \wedge (\psi \wedge \chi) \mid \theta) = p((\phi \wedge \psi) \wedge \chi \mid \theta)$. Thus, he has derived that $\wedge$ is commutative and associative, up to probabilistic equivalence.

He then turns his attention to negation and derives several important results. First, he derives that $p(\neg(\phi \wedge \neg\phi) \mid \psi) = 1$. Then he derives that $p(\neg(\neg\phi \wedge \neg\psi) \mid \chi) = p(\phi \mid \chi) + p(\psi \mid \chi) - p(\phi \wedge \psi \mid \chi)$. If we introduce an abbreviation $\vee$ such that $\phi \vee \psi$ just stands for $\neg(\neg\phi \wedge \neg\psi)$, this becomes $p(\phi \vee \psi \mid \chi) = p(\phi \mid \chi) + p(\psi \mid \chi) - p(\phi \wedge \psi \mid \chi)$, which is a version of the standard law of Additivity. He then derives that $p(\phi \wedge (\psi \wedge \chi) \mid \theta) = p((\phi \wedge \psi) \wedge (\phi \wedge \chi) \mid \theta)$, and $p(\phi \wedge (\psi \vee \chi) \mid \theta) = p((\phi \wedge \psi) \vee (\phi \wedge \chi) \mid \theta)$. Using this, he derives that $p(\neg\neg\phi \wedge \psi \mid \chi) = p(\phi \wedge \psi \mid \chi)$ and that if $p(\phi \mid \chi) = p(\psi \mid \chi)$ then $p(\neg\phi \mid \chi) = p(\neg\psi \mid \chi)$. He then derives that $p(\phi \vee \phi \mid \psi) = p(\phi \mid \psi)$. And finally, he proves that if for all κ, $p(\phi \mid \kappa) = p(\psi \mid \kappa)$, and $p(\chi \mid \kappa) = p(\theta \mid \kappa)$, then for all κ, $p(\phi \wedge \psi \mid \kappa) = p(\chi \wedge \theta \mid \kappa)$.

With these conditions, he is then able to show that logically equivalent formulas are probabilistically equivalent, and derive the facts I mentioned above about abnormal formulas, and probabilities conditional on normal formulas.

For Popper, one of the important features of this characterization is that probability can play the role of giving the meanings of the logical symbols. This is quite a natural desideratum for a logical interpretation of probability, though it may not be as natural for other interpretations. This program is developed further by Field (1977), who gives a method for giving meanings to quantifiers (though this is substantially more clumsy than Popper's method for the connectives).

One thing to note about Popper's formalism is that infinitary versions of Additivity (and Conglomerability, and Disintegrability) can't even be *stated*, much less satisfied or violated. First, every formula is finite, so that even if the language is expanded by adding a disjunction symbol, there are no infinite disjunctions explicitly expressible in the language. Second, by the Compactness Theorem of propositional logic, no formula in this language is logically equivalent to an infinite disjunction of formulas expressible in the language unless it is also logically equivalent to a disjunction of finitely many of those disjuncts. One might wonder whether this holds for probabilistic equivalence, but probabilistic equivalence is only defined for formulas within the language, and infinite disjunctions aren't in the language, so the question doesn't arise.

While some might find this to be an advantage of the sentential formulation of probability, many have found it to be a limitation and have given what they call versions of Popper's system where the bearers of probability are sets rather than formulas of a language, and the operations are set intersection and complement rather than (uninterpreted) $\wedge$ and $\neg$

(Roeper & LeBlanc, 1999; Hájek & Fitelson, 2017). But since Popper's goal was at least partly to characterize the sentential operations in terms of probability, rather than using facts about sets to prove some results about probability, I think of these systems as significantly different.

Versions of these systems are given by van Fraassen (1976), Spohn (1986), McGee (1994), and Halpern (2009), among others. Because the bearers of probability are sets, these authors are able to prove more general characterizations than Kemeny. In particular, Spohn shows that if we add Countable Additivity to Popper's axioms, then these probabilities can always be represented as a lexicographically-ordered set of Countably Additive measures m_i. However, because of the results mentioned in Section 1.4.2, there must be failures of Conglomerability and Disintegrability in certain partitions, even if Countable Additivity is assumed. These authors also show several results relating these set-theoretic versions of Popper's system to probabilities involving infinitesimals (as discussed in Sylvia Wenmackers' contribution to this volume). However, while McGee claims that the two systems are equivalent, Halpern shows that there are some subtleties to consider. But once we start looking at Countably Additive set-based systems that are like Popper's it is useful to consider a slightly more general formalization that includes all of the above as special cases.

3.2 *Rényi*

Alfréd Rényi gave the first English-language version of his system for conditional probability in his (1955), though it also appears briefly in the second chapter of the posthumous textbook (1970a) and is developed in somewhat greater detail in the second chapter of his (1970b). I will generally follow his (1955) in my discussion, though the structural requirements on $\mathcal{B}$ only appear in the later books. Some of the theory appears slightly earlier in publications in German or Hungarian.

Although philosophers often lump Popper and Rényi together, Rényi's early theory is much more flexible than Popper's. It does include a set-based version of Popper's system as a special case, but it also includes a version of Kolmogorov's Regular Conditional Probability as a special case as well. However, Rényi's major aim in developing his theory is to account for a very different application from either of these (and in fact, his later theory explicitly rules out non-trivial versions of Popper and Kolmogorov's systems in favor of these other applications). In statistical practice it is sometimes relevant to work with an "improper prior"—something much like a probability function, that can turn into a probability function by conditioning on some event, but for which the unconditional "probabilities" are infinite. This flexibility also allows Rényi's theory to include

actual relative frequencies, as a system where there is no unconditional probability and all probabilities are conditional.

3.2.1 *Overview*

The background theory for Rényi's conditional probabilities (just like for Regular Conditional Probabilities) is the traditional Kolmogorov axiomatization of probability. There is some set Ω of "possibilities," and the bearers of probability are subsets of this set. (Different interpretations of probability will interpret these possibilities and sets of them differently.) Not every subset of the space of possibilities is a bearer of probability, but there is some collection $\mathcal{A}$ of them that are. $\mathcal{A}$ is assumed to be a *σ-algebra* or *σ-field*, which means (as before) that the empty set is an element of $\mathcal{A}$, the complement of any element of $\mathcal{A}$ is an element of $\mathcal{A}$, and if A_i for $i \in \mathbb{N}$ are any countable collection of elements of $\mathcal{A}$, then $\bigcup_{i\in\mathbb{N}} A_i$ is also an element of $\mathcal{A}$.[18]

$\mathcal{A}$ is the set of bearers of probability. But unlike in Popper's theory, not every bearer of probability can be conditioned on. Instead, Rényi considers a collection $\mathcal{B} \subseteq \mathcal{A}$ subject to the following conditions. For any B_1 and B_2 that are both in $\mathcal{B}$, $B_1 \cup B_2 \in \mathcal{B}$. There exists a countable sequence B_i for $i \in \mathbb{N}$ of elements of $\mathcal{B}$ such that $\bigcup_{i\in\mathbb{N}} B_i = \Omega$. And $\emptyset \notin \mathcal{B}$. While $\mathcal{A}$ is a σ-algebra, $\mathcal{B}$ is a "bunch," that may lack complements and infinite unions, as well as Ω, and definitely lacks the empty set.

He then defines a conditional probability function $p(A \mid B)$ for $A \in \mathcal{A}$ and $B \in \mathcal{B}$ to be any function satisfying the following conditions. For all $A \in \mathcal{A}$ and $B \in \mathcal{B}$, $p(A \mid B) \geq 0$ and $p(B \mid B) = 1$. For any countable sequence of disjoint sets $A_i \in \mathcal{A}$, $p(\bigcup_{i\in\mathbb{N}} A_i \mid B) = \sum_{i\in\mathbb{N}} p(A_i \mid B)$—conditional on any fixed element B, probability is Countably Additive. Finally, if $B, C, B \cap C \in \mathcal{B}$, then $p(A \cap B \mid C) = p(A \mid B \cap C)p(B \mid C)$. (In the later book he adds one more condition, which I will discuss later.) Although there is no official notion of unconditional probability, if $\Omega \in \mathcal{B}$, then we can use $p(A \mid \Omega)$ as a surrogate for $p(A)$. (The fact that $\mathcal{B}$ may lack Ω may make this formalism of particular interest for interpretations of probability where some positive amount of information is needed to generate any probabilities, like actual relative frequency, and perhaps logical and evidential probability. See Section 1.2.)

18 In the previous section, '$\mathcal{F}$' was used for the field of all bearers of probability and '$\mathcal{A}$' was used for the sub-field that we are conditioning on. In this section I follow Rényi in using '$\mathcal{A}$' for the field of all bearers of probability, and '$\mathcal{B}$' for the subset that can be conditioned on. I hope that the change in notation is not too confusing—readers should expect still other choices of letters in other sources on this topic.

3.2.2 *Simplest Examples*

Rényi gives several basic examples of conditional probability spaces satisfying these axioms. Many of these examples use the notion of a "measure," which is very much like a probability function. A measure is just a Countably Additive function μ assigning non-negative extended real numbers to elements of a σ-algebra $\mathcal{A}$ of subsets of some set Ω. To say that the values are "extended real numbers" is just to say that in addition to all the non-negative real numbers, $+\infty$ is also a possible value of the function, with Countable Additivity defined to include this value in the obvious ways (as the sum of any non-convergent series of positive real numbers, or as the sum of any set including $+\infty$). The difference between a measure and a probability function is that for a standard probability function, $p(\Omega) = 1$, while for a measure, $\mu(\Omega)$ can be any non-negative extended real number. A measure is said to be *finite* if $\mu(\Omega)$ is a positive real number, and *σ-finite* if there is a countable collection of sets S_i for $i \in \mathbb{N}$ with each $\mu(S_i)$ finite and $\Omega = \bigcup_{i\in\mathbb{N}} S_i$.

The most basic example of a Rényi conditional probability space is to let μ be any finite measure, and let $\mathcal{B}$ be the collection of all elements of $\mathcal{A}$ whose measure is positive. Then define $p(A \mid B) = \mu(A \cap B)/\mu(B)$, and it is straightforward to see that all axioms apply. Of course, this example is of no help to the problems discussed in Section 1.3, because it leaves probabilities conditional on many elements of $\mathcal{A}$ undefined, and in particular on any element whose measure is 0, which are exactly the elements that have unconditional probability 0.

A slightly more general example is to let μ be any measure at all on Ω, and let $\mathcal{B}$ be the collection of all elements of $\mathcal{A}$ whose measure is positive and finite. Then define $p(A \mid B) = \mu(A \cap B)/\mu(B)$. Interestingly, if $\mu(\Omega) = +\infty$, then this means that there is no notion of unconditional probability—all probability is conditional probability. However, in addition to leaving out probabilities conditional on Ω, this sort of example also still leaves out $p(A \mid B)$ when $\mu(B) = 0$. However, this sort of example is the one that motivated Rényi's development of the theory, and in his later books he adds an axiom that entails that every conditional probability space is of this type, with μ being σ-finite. I will come back to the features of this class of examples later.

3.2.3 *Popper and Kolmogorov*

In the slightly more general system defined in his earlier paper, he also gives several other interesting examples. Instead of a single measure μ we can consider a countable *set* of measures μ_i for $i \in \mathbb{N}$. Then we let $\mathcal{B}$ be the collection of all members of $\mathcal{A}$ such that there is exactly one α with $\mu_\alpha(B) > 0$, and no α such that $\mu_\alpha(B) = +\infty$. If we define

$p(A \mid B) = \mu_\alpha(A \cap B)/\mu_\alpha(B)$ for this unique α, then we have another example of a Rényi conditional probability function. By Spohn's result mentioned in Section 3.1, this means that every Countably Additive Popper function is an example of a Rényi conditional probability function (where we leave probability conditional on abnormal sets undefined, rather than saying it is uniformly equal to 1).

Rényi also considers cases in which Disintegrability or Conglomerability might be satisfied. Starting on p. 307 of his (1955), he discusses both what he calls "Cavalieri spaces" and then "regular probability spaces." These are spaces in which $\mathcal{A}$ is the σ-algebra generated by a random variable V, and $\mathcal{B}$ contains all the sets of the form $x < V < y$ as well as the sets of the form $V = x$, and in which the probability function satisfies Conglomerability with respect to the partition in terms of $V = x$. As he notes, his basic definition of a conditional probability space allows for Conglomerability over $\mathcal{A}$ to fail. However, he gives several examples in which it holds, including an instance of the Borel paradox where $\mathcal{B}$ is the set of longitudes and wedges built up from longitudes. This shows a case where he allows for non-trivial probabilities conditional on some events of probability 0. But it leaves conditional probability undefined for *any* event that is not composed of longitudes.

As I discussed in Section 1.4.3, if we consider not just one conditional probability function, but have many, each with its own $\mathcal{B}$, such that every non-empty set is in one of the $\mathcal{B}$, then we can get an adequate notion of conditional probability that responds to the problem of conditioning on events of probability 0 (from Section 1.3) while satisfying Conglomerability. However, $p(A \mid B)$ will then depend on which probability function is being used, which corresponds to the question of which bunch $\mathcal{B}$ of sets is the base of conditioning. Regular Conditional Probability is a special case of Rényi's theory, where $\mathcal{B}$ ranges only over sub-σ-algebras and Conglomerability is required to hold.

Thus, Rényi's theory is mathematically more general than the theory of Regular Conditional Probability. However, this generality leaves many choices open to us. If the philosophical interest is in preserving a unique notion of conditional probability that doesn't depend on $\mathcal{B}$ at all, then most of this generality is unwanted. Restricting to the case where $\mathcal{B}$ just is the set of all non-empty sets is the subject of Section 3.3.

3.2.4 *Infinite Measure*

Despite the interest of these sorts of conditional probability spaces, Rényi's primary interest is in the second example from Section 3.2.2, where the conditional probability is defined from a single measure μ that is σ-finite but not finite. This is made clear by the discussion in the first two pages

of his (1955) of the importance of unbounded measures in statistical practice. In his (1970a) he adds an extra axiom to the definition of a conditional probability space, requiring that for any $B, C \in \mathcal{B}$ with $B \subseteq C$, $p(B \mid C) > 0$.[19] And in his (1955), most of his discussion is confined to spaces that satisfy it.

As Theorem 8 in his (1955), and as Theorem 2.2.1 of his (1970a), he proves that for every conditional probability space satisfying this further condition, there is a σ-finite measure μ such that $p(A \mid B) = \mu(A \cap B)/\mu(B)$ for all $A \in \mathcal{A}$ and $B \in \mathcal{B}$, and that this measure is unique up to constant multiple.

The proof is not terribly difficult. Recall that there is a countable sequence $B_i \in \mathcal{B}$, for $i \in \mathbb{N}$ with $\bigcup_{i \in \mathbb{N}} B_i = \Omega$. Without loss of generality, we can assume that $B_i \subseteq B_j$ for any $i \leq j$. (If they don't already satisfy this condition, just replace B_j with the finite union $\bigcup_{i \leq j} B_i$.) Now we can define $\mu(B_1) = 1$, and $\mu(B_n) = 1/p(B_1 \mid B_n)$. Then, for any $A \in \mathcal{A}$, we can define $\mu(A) = \lim_{n \to \infty} \mu(B_n) p(A \mid B_n)$. Verifying that this definition of μ is well-defined and gives a measure is somewhat tedious, but not terribly difficult. It is substantially easier to verify that any other measure giving the same conditional probability function must be a constant multiple of this one, and that this one is σ-finite.

By restricting consideration to this sort of probability space, Rényi eliminates all of the non-trivial Popper functions. This is because under this new characterization, whenever $p(A \mid B)$ is defined, $p(B \mid C)$ will be positive whenever it is also defined, unless $C \cap B = \emptyset$. However, Popper's notion of conditional probability was intended to capture cases where $p(B) = 0$ and yet B is normal.

Some philosophers have grouped Popper and Rényi together as giving similar notions of primitive conditional probability. However, Rényi requires Countable Additivity where Popper can't even state it, and Rényi's mature theory rules out all interesting Popper functions, as well as ruling out any resolution to the problem of conditioning on events of probability 0. Although Rényi's theory even more so than Popper's makes conditional probability the basic notion (because Ω can fail to be in $\mathcal{B}$), it addresses only the motivating problem from Section 1.2 (the conceptual requirement that all probabilities are conditional) and not the one from Section 1.3 (conditioning on events of probability 0).

This mature theory works well for the actual relative frequency interpretation of probability. In fact, one of the standard examples that Rényi considers has exactly this form. Let Ω be some countable set, let $\mathcal{A}$ be the collection of all subsets of this set, and let $\mu(A)$ be the number of

19 He appears to have this same restriction in mind in his (1970b), though he writes the requirement in a way that is *conditional* on $p(B \mid C) > 0$ rather than requiring it. But that book develops very little of the theory.

elements of A. (Since Ω is countable, we see that μ is σ-finite, since Ω is the union of countably many sets with finitely many elements each.) If we let $\mathcal{B}$ be the set of all non-empty finite subsets of Ω, and define $p(A \mid B) = \mu(A \cap B)/\mu(B)$, then this just is the definition of finite relative frequency.

3.2.5 *Improper Priors*

Another more characteristic example lets Ω be the set $\mathbb{R}^2$ of pairs of real numbers. Let $\mathcal{A}$ be the collection of all Lebesgue measurable subsets of this set, and let μ be standard Lebesgue measure. Then let $\mathcal{B}$ be the set of all Lebesgue measurable subsets of this set with positive finite measure. The resulting probability measure is uniform conditional on any finite region, and undefined on infinite or null regions.

If we return to the generality of the early theory (so that we allow $\mathcal{B}$ to contain elements whose probability is 0 conditional on large elements of $\mathcal{B}$), we can generalize to a slightly more interesting set $\mathcal{B}$ as follows. Let $R^{x_2,y_2}_{x_1,y_1}$ be the rectangle of points $\{(x,y)\colon x_1 \leq x \leq x_2, y_1 \leq y \leq y_2\}$. Let $\mathcal{B}$ be the set of all such rectangles. When $x_1 < x_2$ and $y_1 < y_2$, we define $p(A \mid R^{x_2,y_2}_{x_1,y_1})$ as before, as the ratio of the standard two-dimensional Lebesgue measure of $A \cap R^{x_2,y_2}_{x_1,y_1}$ to the measure of $R^{x_2,y_2}_{x_1,y_1}$, which is just $(x_2 - x_1)(y_2 - y_1)$. However, when $x_1 = x_2$ or $y_1 = y_2$, the "rectangle" is actually a line segment. In such a case we use the relevant *one*-dimensional Lebesgue measure to define the conditional probability. (This is effectively an example where we have a sequence of three measures—two-dimensional Lebesgue measure $\mu_{x,y}$, one-dimensional Lebesgue measure μ_x with respect to x, and one-dimensional Lebesgue measure μ_y with respect to y.) Again, our probability is uniform conditional on finite rectangles of positive size, but it is also uniform conditional on finite line segments parallel to the x or y axis. But again, there is no unconditional probability, because the space as a whole has infinite measure.

The motivation for this sort of example comes when we generalize it still further. Instead of using Lebesgue measure, we use a measure with a non-uniform density. Then the formulas for calculating conditional probabilities are exactly those given in Section 2.1 for Kolmogorov's Regular Conditional Probabilities, except that some of the integrals might be infinite, and we only officially allow for probabilities conditional on sets where the integrals are finite. In that section, since there was an unconditional probability function, the integrals were always guaranteed to be finite, but here we allow for them to be infinite. When they are infinite, it is standard to say that the conditional probability function arises from an "improper prior," which is not itself a probability function.

This is the foundation of much Bayesian statistical practice. For instance, one might be interested in estimating the distribution of values of V in some population. One might antecedently be sure that, over the relevant population, V is distributed according to a normal distribution with some unknown mean μ and variance σ^2. In the absence of information one wants an "uninformative prior," which should be invariant under changes of measuring scale of V. (For instance, we might convert feet to meters, or Fahrenheit to Celsius.) It turns out that the only such prior is one where the probability that $x_1 < \mu < x_2$ and $0 < y_1 < \sigma^2 < y_2$ is proportional to $(x_2 - x_1) \log \frac{y_2}{y_1}$. But without antecedent bounds on how large μ and σ^2 might be, this gives rise to an improper prior. In particular, since

$$\int_{x_1}^{x_2} \int_{y_1}^{y_2} \frac{1}{y} \, \mathrm{d}y \, \mathrm{d}x = (x_2 - x_1) \log \frac{y_2}{y_1},$$

this means that we can do the calculations with a density given by $d(\mu, \sigma^2) = 1/\sigma^2$.

In this case, in addition to the population mean and variance, there are further random variables given by the observed values of V on samples from the population. We have assumed that each of these samples is taken from the same normal distribution with mean μ and variance σ^2. If we represent the density of the normal distribution by $N_{\mu,\sigma^2}(x)$, then our overall density is given by $d(x, \mu, \sigma^2) = N_{\mu,\sigma^2}(x)/\sigma^2$. Interestingly, although this density yields an improper prior, it turns out that conditional on any possible observed value of x, the integral over all values of μ and σ^2 is finite (because the normal distribution dies off fast enough in each direction). It is a classic result of Bayesian statistics that the posterior distribution of μ conditional on observed x values is given by Student's t-distribution. There are many other cases like this, where a density function over some parameters gives rise to an improper prior, but the natural likelihood function for some observable evidence yields a proper posterior conditional on any possible observation.

Of course, all of this Bayesian analysis only works when it is possible to calculate probabilities by integrating densities. This only works when the conditional distributions satisfy Conglomerability (and thus Countable Additivity) wherever they are defined. Thus, this sort of statistical application requires both Rényi's idea that "unconditional probabilities" can be unbounded, and Kolmogorov's idea that conditional probabilities might be relativized to a partition.

However, the notion of an improper prior is also in some ways closely conceptually related to *failures* of Countable Additivity. This can be seen by looking back at the first example we gave of an improper prior. This was the conditional probability space given by finite counting over a countable set. There is some sense in which this conditional probability

space is aiming to represent a uniform unconditional probability over the countable set, like the de Finetti lottery that (for some) motivates rejection of Countable Additivity. By the technique of improper priors, Rényi is able to represent this distribution in a way that captures much that is important, though it does not give any notion of unconditional probability. Because the total space is σ-finite, there is a countable sequence of sets $B_i \in \mathcal{B}$ for $i \in \mathbb{N}$ such that $\Omega = \bigcup_{i \in \mathbb{N}} B_i$. We can define a merely Finitely Additive probability function over Ω by defining $p(A) = \lim_{i \to \infty} p(A \mid B_i)$, though for many sets A this limit is undefined, and in general the limit will depend on the specific choice of the sequence B_i.

3.3 *De Finetti/Dubins—Full Coherent Conditional Probabilities*

The final theory of Coherent Conditional Probabilities to be considered here takes seriously the motivation in these cases to have well-defined unconditional probabilities while giving up on Countable Additivity. This theory arises from de Finetti (1974) and Dubins (1975, section 3). However, it may be useful for many readers to also consult the expositions of this theory by Seidenfeld (2001), Seidenfeld et al. (2013), or the book length treatment by Coletti and Scozzafava (2002).

The basic background system is the same as that of Kolmogorov and Rényi, but I repeat the definitions here so that readers don't have to flip back. There is a set Ω of possibilities, and we consider some collection $\mathcal{A}$ of subsets of Ω. If $\mathcal{A}$ contains the empty set, as well as complements and pairwise unions of its members, then $\mathcal{A}$ is said to be an *algebra*. If it also contains unions of any countable set of its elements, then it is said to be a *σ-algebra*. An algebra $\mathcal{B}$ is said to be a *sub-algebra* of $\mathcal{A}$ iff every member of $\mathcal{B}$ is a member of $\mathcal{A}$, and a *sub-σ-algebra* of $\mathcal{A}$ if $\mathcal{B}$ is a σ-algebra.

Unconditional probability for an algebra $\mathcal{A}$ is assumed to be given by a function $p(A)$ defined for $A \in \mathcal{A}$ subject to the three basic principles. $p(\Omega) = 1$, $p(A) \geq 0$ for all $A \in \mathcal{A}$, and $p(A \cup B) = p(A) + p(B)$ when A and B are disjoint members of $\mathcal{A}$. If $\mathcal{B}$ is a sub-algebra of $\mathcal{A}$, then a conditional probability for $(\mathcal{A}, \mathcal{B})$ is a two-place function $p(A \mid B)$ defined for $A \in \mathcal{A}$ and non-empty $B \in \mathcal{B}$ subject to the following constraints. For any $A \in \mathcal{A}$ and non-empty $B \in \mathcal{B}$, $p(A \mid B) \geq 0$ and $p(B \mid B) = 1$. For any $A_1, A_2 \in \mathcal{A}$ and non-empty $B \in \mathcal{B}$, if $A_1 \cap A_2 \cap B$ is empty, then $p(A_1 \mid B) + p(A_2 \mid B) = p(A_1 \cup A_2 \mid B)$. For any $B, C \in \mathcal{B}$ with $B \cap C$ non-empty, and any $A \in \mathcal{A}$, $p(A \cap B \mid C) = p(A \mid B \cap C) p(B \mid C)$.

These axioms are much like Popper's axioms, but formulated in terms of sets rather than sentences of a language. They are much more like Rényi's axioms, but without Countable Additivity (and without the requirement that $\mathcal{A}$ be a σ-algebra), and with the additional requirement that $p(A \mid \Omega)$ be defined (since Ω is a member of any algebra $\mathcal{B}$).

One further notion is of great interest here. If $\mathcal{B} = \mathcal{A}$, then the Coherent Conditional Probability is said to be *Full*. The central results in the relevant section of Dubins' paper show that for any probability function on an algebra $\mathcal{A}$ there is a Full Coherent Conditional Probability agreeing with it, and that for any conditional probability function on $(\mathcal{A}, \mathcal{B})$ there is an extension to a Full Coherent Conditional Probability. In fact, he shows that the same is true for any partial function, each of whose finite fragments can be extended to a Full Coherent Conditional Probability function on its finite algebra. In particular, this applies to any Rényi conditional probability function, and even allows us to extend to the case in which $\mathcal{A}$ is the full power set of Ω. Thus, we are able to get what Popper was after—a notion of conditional probability that is defined for every non-empty set.

However, the techniques for proving that these Full Coherent Conditional Probabilities exist are non-constructive. Dubins uses Tychonov's theorem (which is equivalent to the Axiom of Choice), and cites similar results by Krauss (1968) arrived at using non-principal ultrafilters (whose existence is proven using the Axiom of Choice). Similar results extending linear (i.e., finitely additive) functions on subspaces to full spaces often appeal to the Hahn-Banach Theorem, which is also independent of Zermelo-Fraenkel set theory without the Axiom of Choice. Given a Full Coherent Conditional Probability on the surface of a sphere, one can generate the paradoxical Banach-Tarski sets (Pawlikowski, 1991). Thus, we are not usually able to work with these Full Coherent Conditional Probabilities in any explicit way, if we really want them to be defined on *all* subsets of a reasonably-sized probability space. I have argued elsewhere (Easwaran, 2014) that mathematical structures depending on the Axiom of Choice in this way cannot be of epistemic or physical relevance, though they are surely of mathematical interest.

Given the results of Section 1.4.3, Full Coherent Conditional Probabilities fail to satisfy Conglomerability when some Additivity fails. For instance, let Ω be the set of pairs (m, n) of natural numbers. Let S_m be the set of all pairs whose first coordinate is m and let T_n be the set of all pairs whose second coordinate is n. Let p be any probability function such that $p(S_m \mid T_n) = p(T_n \mid S_m) = 0$ for all m and n. (We can think of this probability function as describing two independent de Finetti lotteries.) Let E be the event that $m > n$. Then we can see that for any m, $p(E \mid S_m) = 0$ (since, conditional on S_m, only finitely many values of n will satisfy E), but for any n, $p(E \mid T_n) = 1$ (since, conditional on T_n, only finitely many values of m will *fail* to satisfy E). Since the S_m and the T_n are both partitions, *any* value of $p(E)$ will fail to satisfy Conglomerability in at least one of these partitions. This sort of failure of Conglomerability is inevitable if one allows failures of Countable Additivity and requires that sets like E nevertheless have both unconditional and conditional probabilities.

However, these Finitely Additive Full Coherent Conditional Probabilities have the advantage of existing even for algebras that are not countably generated, avoiding the problems for Regular Conditional Probabilities mentioned in Section 2.3.1. They also always satisfy $p(A \mid A) = 1$, even in the bad algebras where Countably Additive conditional probabilities are forced to allow for $p(A \mid A) = 0$, as mentioned in Section 2.3.2 (Seidenfeld et al., 2001). In particular, in addition to the case where one adds a non-measurable set to the collection of Borel sets, one might also consider the algebra of "tail events," defined as follows.

Let Ω be the set of all countable sequences $(a_0, a_1, a_2, \dots)$ of 0s and 1s (which can be taken to represent the set of all countable sequence of coin flips). Let $\mathcal{A}$ be the σ-algebra generated by the sets of the form

$$A_i = \{(a_0, a_1, a_2, \dots) : a_i = 1\}.$$

Say that an element $A \in \mathcal{A}$ is a "tail event" if, for any element of A, changing any finitely many places in the sequence results in another element of A. (The tail events are exactly those that depend only on the long-run behavior of the sequence and not on any short-term behavior.) Let $\mathcal{B}$ be the set of all tail events. It is clear that $\mathcal{B}$ is a sub-σ-algebra of $\mathcal{A}$.

A classic result of Kolmogorov shows that if the unconditional probability is that on which each event A_i ("the i-th flip results in heads") is independent with probability $1/2$, then every event in $\mathcal{B}$ has probability 1 or 0. A further generalization by Hewitt and Savage shows that if the unconditional probability is *any* "exchangeable" probability (in the sense of de Finetti), then the events in $\mathcal{B}$ all have probability 1 or 0. As a consequence of these results, and a theorem about algebras in which all probabilities are 1 and 0, it turns out that any element $B \in \mathcal{B}$ whose unconditional probability is 0 must also have $p(B \mid B) = 0$, if conditional probability is Countably Additive. (See Blackwell and Dubins, 1975, or Seidenfeld et al., 2001. This is possible because the algebra of tail events is not countably generated.) But if conditional probability is allowed to be merely Finitely Additive, then we can have $p(B \mid B) = 1$ for these tail events. Dubins and Heath (1983) show how to construct such a Full Coherent Conditional Probability. However, this construction assumes a particular merely Finitely Additive probability distribution over all subsets of the natural numbers, and thus indirectly appeals to the Hahn-Banach Theorem, and thus the Axiom of Choice.

Since these functions are defined on the full power set, there is a sense in which we no longer need to limit ourselves to an algebra $\mathcal{A}$ of "measurable" sets. Even the unmeasurable sets are assigned some probability. We aren't able to pin down precisely what the probability is of any such set, but since the non-measurable sets themselves are only proved to exist by non-constructive means using the Axiom of Choice, this may not be

such a problem. The Banach-Tarski Paradox shows that if Ω contains 3-dimensional (or higher) Euclidean space, then any such Finitely Additive probability function must fail to be invariant under rotations and translations. But again, the sets under which these invariances must fail are only proven to exist by means of the Axiom of Choice.[20]

Thus, provided that one is not worried about working with non-constructive methods, Full Coherent Conditional Probabilities can be of interest when dealing with algebras that aren't countably generated.

4 CONCLUSION

There are two main families of arguments that conditional probability should be taken as the basic notion of probability, or at least as equally fundamental to unconditional probability. One set of arguments (Section 1.2) is based on conceptual grounds, but apart from the interpretation of probability as actual frequency, it doesn't appear to be decisive. For logical, evidential, and perhaps even subjective probabilities (if we follow Levi), we may be able to argue that nearly all probabilities are conditional. But if we can make sense of conditioning on a tautology, then again the argument is not decisive. Instead, this argument points out that many probability functions depend on some background condition that is of a different type than the events that have probabilities.

The other set of arguments (Section 1.3) is based on mathematical grounds. Depending on how we treat vague or indeterminate probabilities (if there even are any), these problem cases may not motivate anything beyond a supervaluational treatment. I believe that supposed cases of conditioning on an event with undefined unconditional probability are either cases of maximally vague probability, cases where the "event" is actually part of the background for a probability function rather than a condition, or are cases where the conditional probability also does not exist.

Instead, it is cases of probability 0 (and particularly those where the 0 arises from an infinite partition) that motivate a reconsideration of the mathematics of probability theory the most strongly. To deny that these cases exist is to assume something much stronger than Finite Additivity or Countable Additivity—it is either to assume Full Additivity for all cardinalities (and thus discrete probability, distributed only over countably many possibilities) or else the even stronger assumption that there are only

20 If we replace the Axiom of Choice by the Axiom of Determinacy, then we lose the Hahn-Banach theorem and the other means by which these Finitely Additive functions were proven to exist, but Lebesgue measure turns out to already be defined—and Countably Additive!—over all subsets of Euclidean space. See Bingham (2010, Section 8).

finitely many possibilities. This seems to go against the meaningfulness of scientific vocabulary discussing numerical parameters in the world.

I have discussed four different mathematical theories for conditioning on events of probability 0. Regular Conditional Probabilities may allow us to say that unconditional probability is prior to conditional probability, while Popper's theory, Full Coherent Conditional Probabilities, and the most general version of Rényi's theory require conditional probability to be prior.

Popper's theory is the one most familiar to philosophers. This theory has the advantage of deriving the relations of deductive propositional logic as special consequences of the probability axioms, so it may be particularly well-suited to the logical interpretation of probability. But because the bearers of probability are sentences in a language rather than sets of possibilities, it can't even express the circumstances that give rise to the problem of probability 0, much less say anything useful about them. In any case, it is effectively an instance of the more general Dubins/de Finetti Full Coherent Conditional Probability.

Rényi's theory is the most general, having versions of the others as special cases (though some require dropping Countable Additivity). Rényi's theory is particularly well-suited to the account of probability as actual relative frequency, and may well be particularly suited to interpretations of probability where not every proposition can be conditionalized upon, particularly if the tautology is one of these propositions (so that there is no such thing as unconditional probability). It also has advantages for certain calculations in a Bayesian statistical framework that depend on the use of "improper priors."

The Dubins/de Finetti Full Coherent Conditional Probabilities, and the Regular Conditional Probabilities descending from Kolmogorov, have competing mathematical virtues. Regular Conditional Probabilities can satisfy Conglomerability in each partition, as well as Countable Additivity, which appears to be the most well-motivated level of Additivity. However, Full Coherent Conditional Probabilities allow each conditional probability to be defined in a unified and coherent way (rather than one depending on a partition in addition to a conditioning event). I suggested in Section 1.1 that actual applications of conditional probability always come with some clear sense of the partition that is relevant, so this is not a cost of the theory of Regular Conditional Probabilities. Full Coherent Conditional Probabilities avoid some problem cases that arise on badly behaved algebras. However, I claim these algebras are too complicated for a finite human mind to grasp, so I think they don't arise in epistemic application in any case. Regardless, Full Coherent Conditional Probabilities are themselves so complex that they can't be proved to exist without some version of the

Axiom of Choice, while Regular Conditional Probabilities can be given constructively when the unconditional probability is defined by a density.

The Regular Conditional Probabilities associated with an unconditional probability are generally only unique up to measure 0. Perhaps there could be some constraint like continuity, or computability, that might uniquely define conditional probabilities for each partition given unconditional probabilities on countably generated algebras. If this is right, then we may be able to say that unconditional probability is basic after all, and conditional probability defined in terms of it. But otherwise, there must be some sense in which conditional probability is either primitive, or at least equally fundamental to unconditional probability. Or else we can follow Myrvold (2015) and allow that we can't always get what we want in a theory of conditional probability.

Rényi's fully general theory must be used in a few situations where conditional probability is required to be independent of unconditional probability (namely, for actual relative frequency in infinite worlds, and in applications requiring "improper priors"). For other applications, the situation is summarized in Table 1 (page 193).

REFERENCES

Arntzenius, F., Elga, A., & Hawthorne, J. (2004). Bayesianism, infinite decisions, and binding. *Mind*, *113*(450), 251–283.

Bacon, A. (2015). Stalnaker's thesis in context. *The Review of Symbolic Logic*, *8*(1), 131–163.

Bartha, P. (2004). Countable additivity and the de Finetti lottery. *British Journal for the Philosophy of Science*, *55*, 301–321.

Bertrand, J. (1889). *Calcul des probabilités*. Gauthier-Villars.

Billingsley, P. (1995). *Probability and measure*. Wiley.

Bingham, N. H. (2010). Finite additivity versus countable additivity: De Finetti and Savage. *Electronic Journal of the History of Probability and Statistics*, *6*(1), 1–33.

Blackwell, D. (1956). On a class of probability spaces. In *Proceedings of the third Berkeley symposium on mathematics, statistics, and probability* (Vol. 2, pp. 1–6).

Blackwell, D. & Dubins, L. (1975). On existence and non-existence of proper, regular, conditional distributions. *The Annals of Probability*, *3*(5), 741–752.

Brandenburger, A. (2007). The power of paradox: Some recent developments in interactive epistemology. *International Journal of Game Theory*, *35*, 465–492.

Briggs, R. (2009). Distorted reflection. *Philosophical Review*, *118*(1), 59–85.

	Finite/Discrete Probability	Regular Conditional Probabilities	Full Coherent Conditional Probabilities
Pros	ratio definition Full Additivity mathematically simple	Conglomerability Countable Additivity mathematically standard	two-place $p(B \mid A)$ defined for all non-empty A exist for all algebras
Cons	only countable spaces no continuous random variables	three-place $p(B \mid A, \mathbf{A})$ problems in bad algebras "almost" uniqueness	only Finitely Additive require Axiom of Choice non-Conglomerability
Priority	unconditional probability	uncond. probability (almost)	conditional probability

Table 1: Summary of views and their features

Carnap, R. (1950). *Logical foundations of probability*. University of Chicago Press.

Chang, J. & Pollard, D. (1997). Conditioning as disintegration. *Statistica Neerlandica*, *51*(3), 287–317.

Coletti, G. & Scozzafava, R. (2002). *Probabilistic logic in a coherent setting*. Kluwer.

de Finetti, B. (1974). *Theory of probability*. Wiley.

Dubins, L. (1975). Finitely additive conditional probabilities, conglomerability, and disintegrations. *The Annals of Probability*, *3*(1), 89–99.

Dubins, L. & Heath, D. (1983). With respect to tail sigma fields, standard measures possess measurable disintegrations. *Proceedings of the American Mathematical Society*, *88*(3), 416–418.

Easwaran, K. (2011a). Bayesianism I: Introduction and arguments in favor. *Philosophy Compass*, *6*(5), 312–320.

Easwaran, K. (2011b). Bayesianism II: Applications and criticisms. *Philosophy Compass*, *6*(5), 321–332.

Easwaran, K. (2013a). Expected accuracy supports conditionalization — and conglomerability and reflection. *Philosophy of Science*, *80*(1), 119–142.

Easwaran, K. (2013b). Why countable additivity? *Thought*, *1*(4), 53–61.

Easwaran, K. (2014). Regularity and hyperreal credences. *The Philosophical Review*, *123*(1).

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329.

Eells, E. (1985). Problems of old evidence. *Pacific Philosophical Quarterly*, *66*, 283–302.

Erdős, P. (1947). Some remarks on the theory of graphs. *Bulletin of the American Mathematical Society*, *53*, 292–294.

Faden, A. M. (1985). The existence of regular conditional probabilities: Necessary and sufficient conditions. *The Annals of Probability*, *13*(1), 288–298.

Field, H. (1977). Logic, meaning, and conceptual role. *The Journal of Philosophy*, *74*(7), 379–409.

Fitelson, B. (1999). The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, *66*(3), S362–S378.

Gallow, J. D. (2016). *Diachronic Dutch books and evidential import*. ms.

Glymour, C. (1980). *Theory and evidence*. Princeton University Press.

Greaves, H. & Wallace, D. (2006). Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind*, *115*(459), 607–632.

Gyenis, Z., Hofer-Szabó, G., & Rédei, M. (2016). *Conditioning using conditional expectations: The Borel-Kolmogorov paradox*. ms.

Hájek, A. (2003). What conditional probability could not be. *Synthese*, *137*, 273–323.

Hájek, A. (2007). Interpretations of probability. *Stanford Encyclopedia of Philosophy*.

Hájek, A. & Fitelson, B. (2017). Declarations of independence. *Synthese*, *194*(10), 3979–3995.

Halpern, J. (2009). Lexicographic probability, conditional probability, and nonstandard probability. *Games and Economic Behavior*.

Hill, B. M. & Lane, D. (1985). Conglomerability and countable additivity. *Sankhyā: The Indian Journal of Statistics, Series A*, *47*(3), 366–379.

Hitchcock, C. (2010). Probabilistic causation. *Stanford Encyclopedia of Philosophy*.

Hoffmann-Jørgensen, J. (1971). Existence of conditional probabilities. *Mathematica Scandinavica*, 257–264.

Horowitz, S. & Dogramaci, S. (2016). Uniqueness: A new argument. *Philosophical Issues*, *26*.

Hosiasson-Lindenbaum, J. (1940). On confirmation. *Journal of Symbolic Logic*, *5*(4), 133–148.

Howson, C. (2008). De Finetti, countable additivity, consistency and coherence. *The British Journal for the Philosophy of Science*, *59*(1), 1–23.

Humphreys, P. (1985). Why propensities cannot be probabilities. *The Philosophical Review*, *94*(4), 557–570.

Humphreys, P. (2004). Some considerations on conditional chances. *British Journal for the Philosophy of Science*, *55*, 667–680.

Jeffrey, R. (1965). *The logic of decision*. McGraw-Hill.

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, *65*(4), 575–603.

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge University Press.

Kadane, J. B., Schervish, M. J., & Seidenfeld, T. (1986). Statistical implications of finitely additive probability. In P. K. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti*. North-Holland.

Kadane, J. B., Schervish, M. J., & Seidenfeld, T. (1996). Reasoning to a foregone conclusion. *Journal of the American Statistical Association*, *91*(435), 1228–1235.

Kemeny, J. (1955). Fair bets and inductive probabilities. *The Journal of Symbolic Logic*, *20*(3), 263–273.

Keynes, J. M. (1921). *A treatise on probability*. Macmillan and co.

Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. Chelsea.

Kopec, M. & Titelbaum, M. (2016). The uniqueness thesis. *Philosophy Compass*.

Krauss, P. H. (1968). Representation of conditional probability measures on Boolean algebras. *Acta Mathematica Hungarica*, *19*(3-4), 229–241.

Levi, I. (1980). *The enterprise of knowledge*. MIT Press.

Lewis, D. (1980). A subjectivist's guide to objective chance. In R. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. 2). University of California Press.

Maher, P. (2006). A conception of inductive logic. *Philosophy of Science*, *73*, 518–523.

Mayo, D. & Cox, D. (2006). Frequentist statistics as a theory of inductive inference. *IMS Lecture Notes–Monograph Series 2nd Lehmann Symposium — Optimality*, *49*, 77–97.

McGee, V. (1994). Learning the impossible. In E. Eells & B. Skyrms (Eds.), *Probability and conditionals*. Cambridge University Press.

Meek, C. & Glymour, C. (1994). Conditioning and intervening. *British Journal for the Philosophy of Science*, *45*, 1001–1021.

Myrvold, W. (2015). You can't always get what you want: Some considerations regarding conditional probabilities. *Erkenntnis*, *80*, 573–603.

Pachl, J. K. (1978). Disintegration and compact measures. *Mathematica Scandinavica*, 157–168.

Pawlikowski. (1991). The Hahn-Banach theorem implies the Banach-Tarski paradox. *Fundamenta Mathematicae*, *138*(1), 21–22.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford University Press.

Popper, K. (1955). Two autonomous axiom systems for the calculus of probabilities. *The British Journal for the Philosophy of Science*, *6*(21), 51–57.

Popper, K. (1959a). *The logic of scientific discovery*. Harper & Row.

Popper, K. (1959b). The propensity interpretation of probability. *The British Journal for the Philosophy of Science*, 25–42.

Ramsey, F. P. (1930). On a problem of formal logic. *Proceedings of the London Mathematical Society*, *30*(1), 264–286.

Rényi, A. (1955). On a new axiomatic theory of probability. *Acta Mathematica Hungarica*, 285–335.

Rényi, A. (1970a). *Foundations of probability*. Holden-Day.

Rényi, A. (1970b). *Probability theory*. North-Holland.

Rescorla, M. (2015). Some epistemological ramifications of the Borel-Kolmogorov paradox. *Synthese*, *192*, 735–767.

Rescorla, M. (2018). A Dutch book theorem and converse Dutch book theorem for Kolmogorov conditionalization. *The Review of Symbolic Logic*, *11*(4), 705–735.

Roeper, P. & LeBlanc, H. (1999). *Probability theory and probability logic*. University of Toronto.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. Chapman and Hall.

Savage, L. J. (1954). *The foundations of statistics*. Dover.

Schervish, M. J., Seidenfeld, T., & Kadane, J. B. (1984). The extent of non-conglomerability of finitely additive probabilities. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *66*, 205–226.

Schoenfield, M. (2016). *Conditionalization does not (in general) maximize expected accuracy*. ms.

Seidenfeld, T. (2001). Remarks on the theory of conditional probability: Some issues of finite versus countable additivity. In V. Hendricks (Ed.), *Probability theory: Philosophy, recent history and relations to science* (pp. 167–178). Kluwer.

Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (2001). Improper regular conditional distributions. *The Annals of Probability*, *29*(4), 1612–1624.

Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (2013). Two theories of conditional probability and non-conglomerability. In *8th international symposium on imprecise probability: Theories and applications, compiègne, france*.

Seidenfeld, T., Schervish, M. J., & Kadane, J. B. (2014). *Non-conglomerability for countably additive measures that are not κ-additive*. Carnegie Mellon University.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction and search*. MIT Press.

Spohn, W. (1986). The representation of Popper measures. *Topoi*, *5*, 69–74.

van Fraassen, B. (1976). Representation of conditional probabilities. *Journal of Philosophical Logic*, *5*(3), 417–430.

van Fraassen, B. (1984). Belief and the will. *The Journal of Philosophy*, *81*(5), 235–256.

von Neumann, J. & Morgenstern, O. (1947). *Theory of games and economic behavior, second edition*. Princeton University Press.

Williamson, T. (2002). *Knowledge and its limits*. Oxford.

Yu, X. (1990). Radon-Nikodym theorem is equivalent to arithmetical comprehension. In W. Sieg (Ed.), *Logic and computation, proceedings of a workshop held at Carnegie Mellon University, June 30-July 2, 1987* (pp. 289–297). American Mathematical Society.

5

INFINITESIMAL PROBABILITIES

Sylvia Wenmackers

> *Suppose that a dart is thrown, using the unit interval as a target;*
> *then what is the probability of hitting a point?*
> *Clearly this probability cannot be a positive real number,*
> *yet to say that it is zero violates the intuitive feeling that,*
> *after all, there is some chance of hitting the point.*
>
> —Bernstein and Wattenberg (1969, p. 171)

> *It has been said that to assume that* $0 + 0 + 0 + ... + 0 + ... = 1$ *is absurd,*
> *whereas, if at all, this would be true if*
> *'actual infinitesimal' were substituted in place of zero.*
>
> —de Finetti (1974, p. 347)

Infinitesimals played an important role in the seventeenth century development of the calculus by Leibniz and—to a lesser extent—by Newton. In the twentieth century, calculus was applied to probability theory. By this time, however, Leibnizian infinitesimals had lost their prominence in mainstream calculus, such that "infinitesimal probability" did not become a central concept in mainstream probability theory either. Meanwhile, nonstandard analysis (NSA) has been developed by Abraham Robinson, an alternative approach to the calculus, in which infinitesimals (in the sense of equation 1 below) are given mathematically consistent foundations. This provides us with an interesting framework to investigate the notion of infinitesimal probabilities, as we will do in this chapter.

Even taken separately, both infinitesimals and probabilities constitute major topics in philosophy and related fields. Infinitesimals are numbers that are infinitely small or extremely minute. The history of non-zero infinitesimals is a troubled one: despite their crucial role in the development of the calculus, they were long believed to be based on an inconsistent concept. For probabilities, the interplay between objective and subjective aspects of the concept has led to many puzzles and paradoxes. Viewed in this way, considering infinitesimal probabilities combines two possible sources of complications.

This chapter aims to elucidate the concept of infinitesimal probabilities, covering philosophical discussions and mathematical developments (in as far as they are relevant for the former). The introduction first specifies what it means for a number to be infinitesimal or infinitely small and it addresses some key notions in the foundations of probability theory. The remainder of the chapter is devoted to interactions between these two notions. It is divided into three parts, dealing with the history, the

mathematical framework, and the philosophical discussion on this topic, followed by a brief epilogue on methodological pluralism. The appendix reviews the literature of 1870–1989 in more detail.

Infinitesimals

In an informal context, infinitesimal means extremely small. The word 'infinitesimal' is formed in analogy with 'decimal': decimal means one tenth part; likewise, infinitesimal means one infinith part. As such, the word 'infinitesimal' suggests that infinitesimal quantities are reciprocal to infinite ones, and that infinitely many of them constitute a unit. In Wenmackers (2018), I have introduced the term 'harmonious' as a property of number systems such that "each infinite number is the multiplicative inverse of a particular infinitesimal number, and vice versa." In other words, an harmonious number system does justice to the etymology of 'infinitesimal.' Moreover, in such a number system, "neither the infinite nor the infinitesimal numbers are conceptually prior to or privileged over the other in any way."

These suggestions can be formalized in non-standard analysis (NSA), which allows us to work with so-called hyperreal numbers. The set of hyperreal numbers, ${}^*\mathbb{R}$, contains positive (and negative) infinite numbers, larger than any (standard) number, as well as their multiplicative inverses, which are strictly positive (or strictly negative, respectively) infinitesimal numbers, smaller than any positive real number yet greater than zero.[1] The hyperreals are harmonious in the sense just defined.

Let us now state the formal definition for infinitesimals that we consider in this chapter. A number x is infinitesimal if

$$\forall n \in \mathbb{N} : \ |x| < \frac{1}{n}. \tag{1}$$

According to this definition, *zero is an infinitesimal* and it is the only real-valued infinitesimal.[2] Number systems that do not contain strictly positive or strictly negative infinitesimals, such as $\mathbb{R}$, are called *Archimedean*; number systems that do contain non-zero infinitesimals, such as ${}^*\mathbb{R}$, are called *non-Archimedean*. NSA is certainly not the only framework for dealing with infinitesimals,[3] but currently it is the most common one for representing infinitesimal probabilities, so that is what this chapter focuses on.

1 Actually, it is more accurate to write '*a* set of hyperreal numbers,' rather than 'the set,' since the definition is not categoric (unlike that of $\mathbb{R}$) and there is no canonical choice among the ${}^*\mathbb{R}$'s. See section 16.2 of the appendix for details.

2 Some authors exclude zero in their definition of infinitesimals, but for the exposition in this chapter it will turn out to be beneficial to include it.

3 Section 11 mentions two alternative frameworks that deal with infinitesimal numbers.

What is an infinitesimal probability value? The answer depends on which number system you are using: we already observed that zero is *the* infinitesimal number within the real numbers, whereas the hyperreal numbers contain (infinitely many) strictly positive infinitesimals, which could serve as strictly positive infinitesimal probability values.

One way to obtain a new number system is by considering a suitable quotient space. In general, the definition of a quotient space relies on the definition of some equivalence relation on a collection of objects, which can be (generalized) sequences.[4] Informally, the equivalence relation expresses a condition for two objects to be "indistinguishable" from each other or for their difference to be "infinitesimal' or "negligible." In the case of (generalized) sequences, this condition has to specify (i) *a criterion* to compare corresponding positions by and (ii) *a selection rule* that specifies at which collections of indices said criterion has to hold. Both the construction of the real numbers and that of the hyperreal numbers fits this general description, but the relevant equivalence relations impose different conditions for sequences to be indistinguishable from each other:

(1) The negligibility of a sequence can be formalized as "converging to zero": the sequence gets (i) *arbitrarily close to* the (rational) number zero (ii) *eventually*.

(2) Another way to define negligibility of a sequence is as being (i) *exactly equal to* the (real) number zero (ii) *except for a small index set*.

We will define the criteria and selection rules in italics later in this chapter (see section 8.5). For now, it suffices to know that two sequences can be defined to be equivalent if they differ only by a negligible sequence (in a well-defined sense). Using this equivalence relation, we can define equivalence classes of sequences; the structure of the collection of these equivalence classes is a quotient set. For some choices, this set may be isomorphic to that of the set of real or hyperreal numbers. In particular, the equivalence class of rational-valued Cauchy sequences that are negligible in the sense of (1) is the real number zero ($0_{\mathbb{R}}$) and the equivalence class of real-valued sequences that are negligible in the sense of (2) is the hyperreal number zero ($0_{*\mathbb{R}}$).

Since being exactly equal to zero implies being infinitely close to zero, but not vice versa, we may think of $0_{\mathbb{R}}$ as *the* infinitesimal in the set of the real numbers, which corresponds with an infinite equivalence class of sequences, many of which belong to that of non-zero infinitesimals in the hyperreal context. In this sense, the hyperreal numbers are capable of representing finer distinctions (among sequences) than the real numbers are.

4 For generalized sequences, see section 9.2.

After this brief introduction to infinitesimals, let us now give an even briefer intro to probabilities.

Probabilities

In an informal context, probable means plausible or likely to be true. Similar words were available in medieval Latin ('*probabilis*' for probable and '*verisimilis*' for likely). As such, probability can be seen as a shorthand for 'probability of truth' and likelihood is a measure of appearing to be true. This suggests that probability is a hybrid concept that combines *objective* chances and *subjective* degrees of belief (or credences). We may picture it as a two-layered concept with an objective ground layer, which represents the objective state of affairs (truth), and an epistemic cover layer, that deals with evidence presented to an agent and quantifying the possibility of it being misleading concerning what is underneath it (appearance).

Many authors have tried to capture this duality that is inherent in the probability concept. Hacking (1975) describes it very aptly as the Janus-faced nature of probability and Gaifman (1986) paints a colourful picture of probability as living on a spectrum from purely objective to purely epistemic forms. It may be helpful to imagine both layers as allowing for different degrees of opacity. For an agent with limited epistemic (cognitive and empirical) resources, the outer layer acts as a veil. First assume that the underlying system is purely deterministic, such that there are no probabilities "out there," or, put differently, they are all zero or one. However, the agent does not see things exactly as they are—only approximately so. Hence, the probabilities that are relevant to such an agent may be other than just zeros and ones.[5] If the underlying system is indeterministic, on the other hand, even an agent with unlimited epistemic resources (such as Laplace's demon), who could see right through the outer layer, would still need probabilities to describe the system.

Apart from its interpretation, the topic of this chapter also requires us to pay attention to the mathematical representation of probabilities. Probability is usually formalized as a function from the event space—a collection of subsets (often a sigma-algebra) of a given set, the sample space—to the unit interval of the real numbers or a non-standard extension thereof. A probability distribution is called *fair* or *uniform* if the same probability is assigned to any singleton from the domain. Depending on other background assumptions, this may imply slightly stronger properties, such as translation invariance.

5 This viewpoint helps us to understand that Laplace (1814) was strongly involved in the development and popularization of probability theory, while also popularizing the idea of a deterministic universe.

In this chapter, we will encounter infinitesimals both in the context of subjective probability (infinitesimal credences or degrees of belief) and in the context of objective probability (infinitesimal chances), as well as in contexts that are intermediate on this continuum.

PART I
HISTORICAL OVERVIEW

In this part, we review some essential mathematical developments that allow us to represent infinitely small probabilities as positive infinitesimals in a hyperreal field. We also review philosophical discussions of the topic. A much more detailed list of contributions from the period 1870–1989 can be found in the appendix. More recent contributions are discussed in Part IV.

The concept of infinitesimals was thought to be intrinsically problematic and inconsistent for most of European history. An important exception is the work of Archimedes, who allowed infinitesimals as a method to find new results, though he did not regard them sufficient for establishing rigorous proofs of those results. In the sixteenth century, a Latin translation of many of the works of Archimedes was published in Europe, which led to a revival of scholarly interest in infinitesimals, especially in Italy. (See Alexander, 2014, for an overview of the seventeenth century response to infinitesimals in Europe.)

In the second half of the seventeenth century, infinitesimals played a crucial role in the development of the calculus, especially in the work of Gottfried Wilhelm Leibniz (see, *e.g.*, Katz & Sherry, 2012; Katz & Sherry, 2013). Whereas the guiding notion in Newton's calculus was the "fluxion" (the derivative of a continuous quantity), Leibniz developed his version of the calculus starting from infinite sums (integrals). Newton's and Leibniz's usage of infinitesimals was criticized early on, famously by Berkeley (1734), who called them "ghosts of departed quantities." Around the 1870s, the calculus received its formalization in terms of real numbers and standard limits, which do not allow non-zero infinitesimals. This further consolidated the general belief that infinitesimals do not live up to the rigour of modern mathematics, but we will see that a formalization of this concept was discovered later on, in the 1960s.

The current standard approach to calculus, which is used for instance in college physics, is based on the nineteenth century formalization, in which the epsilon-delta definition of the limit operation takes a central place (see appendix 16.1). As a result, our standard calculus differs from both the Newtonian and the Leibnizian version of it. The core idea of a limit operation is closer in spirit to the Newtonian version, while Leibnizian notation proved to be more enduring, with, for instance, $\mathrm{d}x/\,\mathrm{d}t$

for the derivative of x to t. (For Leibniz, this signified an actual ratio of infinitesimals, whereas our standard calculus defines it as the limit of a ratio of real numbers.)

As we will see below, measure and probability theory was developed based on the standard calculus. The non-standard approach, based on the alternative formalization of the calculus from the 1960s, is more recent. (Hence the unfortunate name 'non-standard'.) But, like infinitesimals in general, also the more specific notion of infinitesimal probability was in use long before its formal definition. For instance, in his famous wager argument (*Pensées* L418/S680), Pascal specifically excluded them from his argument.[6]

1 THE PRE-ROBINSONIAN ERA: 1880–1959

Around 1880, the current foundations of the real numbers and the standard calculus, with the epsilon-delta definition of the limit, were well in place. Non-standard analysis was not developed yet.

Standard measure theory was being developed by mathematicians such as émile Borel, Henri Lebesgue, Johann Radon, Maurice Fréchet, Giuseppe Vitali, and many others. In response to the sixth problem of David Hilbert (1900), also the first axiomatization of probability theory was developed: Kolmogorov (1933) presented an approach that embedded probability theory into standard measure theory. (His axioms are included in section 7.)

After the foundational work by Kolmogorov, the measure-theoretic approach to probability became the standard formalism, which represents probabilities as real numbers. Strictly speaking, non-zero infinitesimal probabilities (defined as non-Archimedean quantities) are incompatible with this formalism. Nevertheless, informal usage of the term has remained in fashion in at least two ways. First, in some contexts it is used to discuss events that have zero probability but that are logically possible to occur. Second, the phrase 'infinitesimal probability' is also used in the context of continuous probability distributions, to refer to $\mathrm{d}p$.[7]

At about the same time, Bruno de Finetti (1931) was developing a qualitative theory for ranking events in terms of their probability. He discovered that, in general, these rankings are non-Archimedean. His rankings can be said to be more fine-grained than what is expressible

6 In Krailsheimer's translation, the relevant sentence reads as follows (Pascal, 1670/1995, p. 151, my emphasis): "[W]herever there is infinity, and where there are not *infinite chances of losing against that of winning*, there is no room for hesitation, you must give everything."

7 The notation stems from Leibniz, for whom $\mathrm{d}p$ indicated an infinitesimal increment of a quantity p. In contemporary standard analysis, however, there are no non-zero infinitesimals and $\mathrm{d}p$ merely indicates that the variable of differentiation or integration is p.

by the real-valued probability functions in Kolmogorov's theory. Five years later, de Finetti (1936) specifically addressed logically possible events that receive probability zero in Kolmogorov's theory. Here, we see that de Finetti explicitly entertained the notion of infinitesimal probabilities, but he ultimately chose to stick to real-valued probabilities and to reject countable additivity.

Working on the subjective interpretation of probability, Frank P. Ramsey and Bruno de Finetti developed the notion of coherence: in order for an agent's degrees of belief to be rational (at a given point in time), they have to conform to Kolmogorov's axioms for probability. Abner Shimony (1955) aimed to strengthen this notion to strict coherence (now often called regularity): it requires that the degree of confirmation of an hypothesis h given a piece of evidence e is 1 if and only if h logically entails e. Shimony was aware that strict coherence required infinitesimal betting quotients—and thus was incompatible with Archimedean values—if the sample space was infinite. Inspired by this proposal, Rudolf Carnap (1980) set out to develop a theory for non-Archimedean credences. Although this interesting approach was written before Robinson's work, it was only published afterwards. As a result, it has not been very influential.

Meanwhile, Thoralf Skolem (1934) had discovered non-standard models of the natural numbers (Peano arithmetic), which we now call hypernatural numbers. By applying similar model-theoretic techniques to the real numbers, Robinson would be able to develop non-standard analysis. This brings us to the next period.

2 ROBINSON'S NON-STANDARD ANALYSIS: 1960S

Abraham Robinson (1961, 1966) founded the field of NSA: he applied earlier results from mathematical logic (such as that of Skolem) to real closed fields in order to develop an alternative framework for differential and integral calculus based on infinitesimals and infinitely large numbers. This allowed for a formal and consistent treatment of infinitesimal numbers and provided a harmonious number system (as defined in the introduction). Soon enough, NSA was applied to measure theory in general and to probability theory in particular.

For our current purposes, it is good to be aware of two modes of operation of NSA: in one, the hyperreal numbers merely serve as a means to prove results about the real numbers, but in the other, obtaining a hyperreal-valued function or some other non-standard object is the final goal.[8] The first mode of operation represents the oldest and still the *most*

8 This situation is similar to that of the complex numbers. On the one hand, as Painlevé (1967, pp. 1–2) writes: "*entre deux vérités du domaine réel, le chemin le plus facile et le plus court passe bien souvent par le domaine complexe*" ("between two truths of the real domain,

common application of NSA, which is to make proofs about standard analysis shorter, easier, or both—mainly by alleviating epsilon-delta management (Tao, 2007).[9] Although the most common one, this is *not the only* application of NSA. The second mode of operation allows us to investigate non-standard objects in their own right, including those that (roughly speaking) do not have standard counterparts.[10] In particular, if we are interested in developing a probability theory that allows us to assign non-zero infinitesimal probabilities to some events, we cannot achieve this if we move back to the real domain in the final step.

An early example of a non-standard measure was provided by Bernstein and Wattenberg (1969), who attempted to measure the infinitesimal probability of hitting a particular point when playing (infinitely precise) darts on the unit interval of the real numbers. This result was a very important first step in the development of probability theories in which the numerical values respect the non-Archimedean ordering of the events (as studied by de Finetti, 1936). Hence, Bernstein and Wattenberg (1969) have often been cited by philosophers who work on the foundations of probability theory. However, since they focused on a particular case, their result is not fully general: they did not present a non-standard probability theory, although their approach can be generalized and does in fact contain many of the essential ingredients present in later developments.

3 POST-ROBINSONIAN DEVELOPMENTS: 1970–1989

Seminal contributions to non-standard measure theory were obtained by Peter A. Loeb (1975). The dominant line of research in non-standard measure and integration theory is based on real-valued functions that have a non-standard domain and the main application (like for all of NSA) is finding new results in standard measure and integration theory. Although the well-developed theory of Loeb measures has proven fruitful in many applications, and therefore should not go unmentioned, it is not of immediate interest to the topic of this chapter (but see Herzberg, 2007, 2010). For, although infinitesimal probabilities do occur in the construction

---

the easiest and shortest route quite often passes through the complex domain"). On the other hand, complex numbers are also useful by themselves (for instance, to represent phasors in physics). This analogy is also employed by Bartha and Hitchcock (1999, p. 416), who write: "Just as imaginary numbers can be used to facilitate the proving of theorems that exclusively concern real numbers, our use of nonstandard analysis will be used to facilitate and motivate the construction of purely real-valued measures."

9 An early expression of this (prior to the development of NSA) can be found with Joseph-Louis Lagrange, as cited in Błaszczyk, Katz, and Sherry (2013, p. 63). Recent examples are given by Terence Tao in his blog posts (see, *e.g.*, Tao, 2007–2012).

10 These are "external" objects, as will be defined in section 4.

of Loeb measures, the end goal is to obtain real-valued measures, thereby eliminating all non-zero infinitesimal probabilities.

Although de Finetti lived long enough to see the advent of NSA and was aware of its existence, he never used it to continue his 1936 observations regarding infinitesimal probabilities and he did not show much interest in applying it in his own work on probability.[11]

To make the earlier, often technical, work accessible to a larger audience, including philosophers, it was important to summarize and interpret it. Brian Skyrms played an important role in this regard. For instance, in Skyrms (1980, appendix 4), he discussed the trade-off between four demands—additivity, translation invariance, everywhere-definedness, and regularity—for standard and non-standard measures. In the same year, David Lewis (1980) discussed infinitesimal credences, in the same spirit as Shimony and Carnap had done prior to Robinson's work. Later on, Lewis (1986a) also mentioned infinitesimal chances, in wordings very reminiscent of Bernstein and Wattenberg (1969).

Observe that at this point, there still was no non-Archimedean alternative to parallel Kolmogorov's Archimedean probability theory. It was Edward Nelson (1987), who provided the first axiomatic approach for a probability theory with infinitesimal values. His "radically elementary probability theory" is indeed very simple, but it requires an entirely different mind set than, for instance, Loeb's approach. In particular, Nelson's theory cannot be used to assign probability measures to any standard infinite set. Instead, one has to go one step back in the modelling process and represent the set of possibilities by an infinite hyperfinite set rather than a standard infinite set. We will introduce the notion of hyperfinite sets in section 4.3. Since hyperfinite sets are very similar to discrete finite ones, after that choice, everything resembles Kolmogorov's theory for finite sample spaces.

At this point, we end our historical overview. Some of the more recent approaches and debates will be discussed in sections 8, 9, and 14.

11 See section 16.3 of the appendix for details.

PART II
MATHEMATICAL PRELIMINARIES

In this part, we will briefly review some common non-standard tools and the dual notions of filters and ideals. We will apply these notions in the ultrafilter construction of the hyperreals. We also present the axioms of standard probability theory. After that, we will be properly equipped to address infinitesimal probabilities in the context of countable lotteries as well as other cases.

4 COMMON NON-STANDARD TOOLS

In this section, we review some common tools that appear in (nearly) all approaches to non-standard analysis.[12]

4.1 *Universe*

By a *universe*, we mean a non-empty collection of mathematical objects, such as numbers, sets, functions, relations, *etc*.—all of which can be defined as sets by working in Zermelo–Fraenkel set theory with the Axiom of Choice (ZFC). This collection is assumed to be closed under the following relations and operations on sets: $\subseteq$, $\cup$, $\cap$, $\setminus$, $(\cdot,\cdot)$, $\times$, $\mathcal{P}(\cdot)$, $\cdot^{\cdot}$. Furthermore, we assume that the universe contains $\mathbb{R}$ and that it obeys transitivity (*i.e.*, elements of an element of the universe are themselves elements of the universe).

In particular, we are interested in the standard universe, which is the superstructure $V(\mathbb{R})$, and a non-standard universe, ${}^*V(\mathbb{R})$.

4.2 *Star-map*

The star-map (or hyperextension) is a function from the standard universe to the non-standard universe.

$$\begin{aligned} * : V(\mathbb{R}) &\to {}^*V(\mathbb{R}) \\ A &\mapsto {}^*A \end{aligned}$$

We assume that $\forall n \in \mathbb{N}$, ${}^*n = n$ and that $\mathbb{N} \neq {}^*\mathbb{N}$.

In the literature, two notations occur for the star map: before or after the standard object. In this chapter, I have opted for the former notation, because it allows us to read the *-symbol as the prefix 'hyper-'. For instance, ${}^*\mathbb{R}$ are the hyperreals.

12 For further information, see also Benci, Di Nasso, and Forti (2006, section 1) and Cutland (1983, section 1.2).

4.3 *Internal and External Objects*

It is important to realize that the star-map does *not* produce all the objects in the superstructure of ${}^*\mathbb{R}$; it only maps to the internal objects, which live in ${}^*V(\mathbb{R}) \subsetneq V({}^*\mathbb{R})$.

Some examples of internal objects ($\in {}^*V(\mathbb{R})$):

- any element of ${}^*\mathbb{R}$, so in particular any element of $\mathbb{N}$ or $\mathbb{R}$;
- any hyperfinite set, such as $\{1, \ldots, N\}$ with $N \in {}^*\mathbb{N}$ (which can be obtained via the hyperextension of a family of finite sets);
- the hyperextensions of standard sets, such as ${}^*\mathbb{N}$ and ${}^*\mathbb{R}$;
- the hyperpowerset of a standard set, A: ${}^*\mathcal{P}(A)$, which is the collection of all *internal* subsets of ${}^*A$.

Some examples of external objects ($\in V({}^*\mathbb{R}) \setminus {}^*V(\mathbb{R})$):

- elementwise copies of standard, infinite sets (notation for the elementwise copy of A in the non-standard universe: ${}^\sigma A$), such as ${}^\sigma\mathbb{N}$ or ${}^\sigma\mathbb{R}$ (due to the embedding of $\mathbb{N}$ and $\mathbb{R}$ in ${}^*\mathbb{R}$, the ${}^\sigma$-prefix is often dropped);
- the complements of previous sets, such as ${}^*\mathbb{N} \setminus {}^\sigma\mathbb{N}$ and ${}^*\mathbb{R} \setminus {}^\sigma\mathbb{R}$;
- the *halo* or *monad* of any real number, r: $hal(r) = \{R \in {}^*\mathbb{R} \mid |r - R| \text{ is infinitesimal}\}$—in particular $hal(0)$, which is the set of all infinitesimals;
- the standard part function *st* (also known as the shadow), which maps a (bounded) hyperreal number to the unique real number that is infinitesimally close to it (Goldblatt, 1998, section 5.6);
- the full powerset of the hyperextension of a standard, infinite set, A: $\mathcal{P}({}^*A)$, which is the collection of *all* subsets of ${}^*A$, both internal and external.

4.4 *Transfer Principle*

Consider some standard objects $A_1, \ldots, A_n$ and consider a property of these objects expressed as an *elementary sentence* (a bounded quantifier formula in first-order logic): $P(A_1, \ldots, A_n)$. Then, the *Transfer* principle says:

$$P(A_1, \ldots, A_n) \text{ is true } \Leftrightarrow P({}^*A_1, \ldots, {}^*A_n) \text{ is true.}$$

Observe: this is an implementation of Leibniz's "law of continuity" (or "*souverain principe*") in NSA (see Katz & Sherry, 2012, section 4.3). It may be helpful to consider two examples.

EXAMPLE 1: WELL-ORDERING OF $\mathbb{N}$ Consider the following sentence: "Every non-empty subset of $\mathbb{N}$ has a least element." Transfer does *not* apply to this, because the sentence is not elementary. Indeed, we can find a counterexample in ${}^*\mathbb{N}$: the set of infinite hypernatural numbers, ${}^*\mathbb{N} \setminus \mathbb{N}$, does not have a least element. (Of course, this is an external object.)

If we rephrase the well-ordering of $\mathbb{N}$ as follows: "Every non-empty element of $\mathcal{P}(\mathbb{N})$ has a least element," then we *can* apply Transfer to this. The crucial observation to make here is that ${}^*\mathcal{P}(\mathbb{N}) \subsetneq \mathcal{P}({}^*\mathbb{N})$.

EXAMPLE 2: COMPLETENESS OF $\mathbb{R}$ Consider the following sentence: "Every non-empty subset of $\mathbb{R}$ which is bounded above has a least upper bound." Again, Transfer does not apply to this, for the same reason as in Example 1. A counterexample in ${}^*\mathbb{R}$ is $hal(0)$, the set of infinitesimals. (Again, an external object.)

If we rephrase the completeness property of $\mathbb{R}$ as follows: "Every non-empty element of $\mathcal{P}(\mathbb{R})$ which is bounded above has a least upper bound," then we can apply Transfer to it. Similarly as before, the crucial remark is that ${}^*\mathcal{P}(\mathbb{R}) \subsetneq \mathcal{P}({}^*\mathbb{R})$.

5 FILTERS AND IDEALS

The introduction mentioned two ingredients for a new number system: the second one is a selection rule. This idea can be formalized using either filters or ideals. These are dual notions, and both are collections of subsets from an index set that fulfil additional criteria.

Intuitively, a filter on a set is a collection of its subsets that are "large enough," whereas an ideal is a collection of its subsets that are "small enough" or "negligible." The meanings of 'large enough' and 'small enough' are given by the formal definitions. The ultrapower construction of the hyperreal numbers crucially relies on the application of a particular kind of filter: a free ultrafilter. We review the relevant definitions here.[13]

$\mathcal{F}$ is a *proper, non-empty filter on* X if

$$\mathcal{F} \subseteq \mathcal{P}(X), \qquad \text{(collection of subsets)}$$

$$\emptyset \notin \mathcal{F}, \qquad \text{(proper)}$$

$$X \in \mathcal{F}, \qquad \text{(non-empty)}$$

$$A, B \in \mathcal{F} \Rightarrow A \cap B \in \mathcal{F}, \qquad \text{(closure under finite meets)}$$

13 Definitions are given, *e.g.*, in Schechter (1997, Ch. 5). For a further discussion of filters, including free ultrafilters, see, *e.g.*, Goldblatt (1998, p. 18–21) and Cutland (1983, section 1.1). For an introduction to the meaning and application of ultrafilters, see Komjáth and Totik (2008).

$$(A \in \mathcal{F} \wedge B \supseteq A) \Rightarrow B \in \mathcal{F}. \qquad \text{(upper set property)}$$

The smallest non-empty proper filter is simply $\{X\}$. A filter $\mathcal{F}$ is *principal* (or *fixed*) if $\exists x_0 \in X : \forall A \in \mathcal{F},\ x_0 \in A$.

A filter $\mathcal{F}$ is *free* if it is *not* principal, or equivalently: if the intersection of all the sets in $\mathcal{F}$ is empty. For an infinite set X, its *Fréchet filter* is the filter that consists of all the cofinite subsets of X. Such a filter is free, but it is not an ultrafilter. (For a finite set X, the Fréchet filter is not proper.)

$\mathcal{F}$ is an *ultrafilter* on X if $\mathcal{F}$ is a filter on X and

$$\forall A \subseteq X (A \notin \mathcal{F} \Rightarrow X \setminus A \in \mathcal{F}).$$

$\mathcal{F}$ is a *free ultrafilter* on X if $\mathcal{F}$ is an ultrafilter on X and $\mathcal{F}$ is free. This definition implies that a free ultrafilter contains no finite sets. Given the ultrafilter condition, it is equivalent to say that it does contain all cofinite sets. In other words: an ultrafilter is free if and only if it contains the Fréchet filter. Hence, free ultrafilters do not exist for finite X.

Given a (proper) filter on X, $\mathcal{F}$, the corresponding (proper) *ideal* in the Boolean algebra $\mathcal{P}(X)$, $\mathcal{I}$, is obtained as follows:

$$\mathcal{I} = \{X \setminus F \mid F \in \mathcal{F}\}.$$

The smallest proper ideal is simply $\{\varnothing\}$. The ideal corresponding to a free ultrafilter is called a *Boolean prime ideal*.

6 APPLICATION OF FREE ULTRAFILTERS: HYPERREAL NUMBERS

6.1 *Constructing the Real and Hyperreal Numbers*

In the introduction, we indicated that both the standard real numbers and the hyperreal numbers can be defined as equivalence classes of sequences.[14] They differ in the collection of sequences on which they operate and in the equivalence relation that they impose.

The real numbers can be constructed based on rational-valued Cauchy sequences. The set of such functions is defined as follows:

$$\mathcal{C} = \{(q_n) \in \mathbb{Q}^{\mathbb{N}} \mid \forall \epsilon \in \mathbb{Q}_{>0},\ \exists N \in \mathbb{N} :\ \forall n, m > N \left(|q_n - q_m| < \epsilon\right)\}.$$

Two sequences in this space are considered to be equivalent to each other if their difference (which is defined member-wise) is a sequence that gets *arbitrarily close to* (the rational number) zero, *eventually*. This means that for each target, from some position in the sequences onwards (*i.e.*, eventually

14 We will not consider Dedekind cuts or other constructions.

or cofinally), their member-wise difference is strictly smaller than the target. Symbolically, where $(q_n), (s_n) \in \mathcal{C}$:

$$(q_n) \sim (s_n) \Leftrightarrow \forall \epsilon \in \mathbb{Q}_{>0},\ \exists N \in \mathbb{N} :\ \forall n > N \left(|q_n - s_n| < \epsilon\right).$$

The hyperreal numbers can be constructed based on real-valued sequences (all of $\mathbb{R}^{\mathbb{N}}$)—this is called the ultrapower construction of ${}^*\mathbb{R}$.[15] Two sequences in $\mathbb{R}^{\mathbb{N}}$ are considered to be equivalent to each other if their member-wise difference is *exactly equal to* (the real number) zero, *except for a small set of indices*. In this case, the first part of the condition is clear and all we are left to specify is what counts as a "small" set. If we choose to define small sets as finite sets, and thus large sets as cofinite ones, this coincides with the "eventuality" condition used in the construction of the real numbers. This is equivalent to imposing the Fréchet filter, consisting of the cofinite subsets of $\mathbb{N}$ (the complements of "small" sets, these are "large" sets), to the indices of the sequences. This setup does allow us to construct a non-standard model of the real numbers; in fact, it was the first one that was ever constructed and it is still of interest because it yields a constructive non-standard model.[16] However, such a system is rather weak (too weak for some of the questions we are interested in). According to the Fréchet filter, many sets (such as arithmetic progressions) are neither small nor large. Usually, small and large sets are defined by fixing a free ultrafilter on $\mathbb{N}$: a set is large if it is in the ultrafilter and small if it is not, and the ultra-condition guarantees that for each set either it is in the ultrafilter, or its complement is.

Informally, the sequence-based construction of the hyperreals can be thought of as follows. Consider the old equivalence class of the sequences that we have come to regard as the real number zero and define new equivalence classes on it, making distinctions among the infinitesimal sequences depending on their rate of convergence. As such, we dissect the single infinitesimal real number into infinitely many infinitesimal hyperreal numbers. In fact, we perform a similar dissection for each of the real numbers simultaneously. Does this give us old wine in new packages? Not quite: it is more like breaking the chemical bonds in the molecules

15 The ultraproduct construction is a general method in model theory: see Keisler (2010) (including the references in the introduction) for more information. To see how the ultrapower construction is related to the existence proof of non-standard models using the Compactness theorem (see appendix section 16.2), observe that one way to prove the Compactness theorem is based on the notion of an ultraproduct (*cf.* Goldblatt, 1998, p. 11).

16 Schmieden and Laugwitz (1958) were the first to give a construction in this style and they used a Fréchet filter on $\mathbb{N}$ rather than a free ultrafilter. Unlike a free ultrafilter, the existence of a Fréchet filter does not require any choice axiom. However, in strictly constructivist approaches, the framework of classical logic as used by Schmieden and Laugwitz (1958) also has to be replaced by intuitionist logic (Martin-Löf, 1990). More recently, Palmgren (1998) has investigated constructive approaches to NSA. For an accessible introduction to a weak system of NSA based on Fréchet filters, see also Tao (2012).

of the wine, and maybe even breaking the atoms—tearing apart the very fabric of what the original numbers are made of, and recombining the fragments in a novel way (with a completely different order structure): we get an entirely new set of numbers out of the operation. Observe that we still have infinitely many real-valued sequences in the equivalence class of the hyperreal number zero (those that differ from zero at only finitely many positions), but—in as far as they converge in the standard sense at all—only a strict subset of them converge to the real number zero.

6.2 *Remarks on the Ultrapower Construction*

When a free ultrafilter is applied in the ultrapower construction of the hyperreal numbers, its various properties affect the properties of the hyperreals in the following ways (see section 8.5):

- the upper set property of a filter is required to obtain an equivalence relation on $\mathbb{R}^{\mathbb{N}}$;
- the property of an ultrafilter, which ensures that each set is either large (in the filter) or small (in the corresponding ideal), is required to obtained trichotomy on ${}^*\mathbb{R}$ (*i.e.*, for each $r, s \in {}^*\mathbb{R}$ either $r < s$ or $r = s$ or $r > s$);
- the property of being free in combination with being ultra, which ensures that every finite set is small, is required to ensure that $\mathbb{R} \subsetneq {}^*\mathbb{R}$.

Although free ultrafilters can be proven to exist (given the usual set-theoretic assumptions), it can also be proven that no explicit example of them can be given; they are inherently non-constructible objects or "intangibles" (Schechter, 1997).

If we drop the condition of being free, and apply the Fréchet filter instead, we obtain a weaker but constructive model of the hypernatural numbers. Let us consider the implication for probability by considering the example of a fair lottery on $\mathbb{N}$. On the one hand, using a Fréchet filter would still allow us to obtain probability functions that take infinitesimal values for finite events. On the other hand, the system is too weak to obtain probability functions that are defined on all of $\mathcal{P}(\mathbb{N})$. For instance, the subset of odd numbers and the subset of even numbers are neither in the Fréchet filter nor in the corresponding ideal, so according to this filter and ideal they are neither large nor small, such that these events would not receive any probability value.

7 KOLMOGOROV'S AXIOMS FOR PROBABILITY THEORY

Since this theory does not contain actual infinitesimals, it may seem of less importance for the topic of this chapter. However, Kolmogorov's approach was very successful and influential: it lies at the basis of the contemporary presentation of probability theory as a special case of measure theory, which itself is a branch of real analysis (calculus).[17] Hence, any later proposal for a new theory of probability, possibly including infinitesimals, has to compete with it. Therefore, we do include Kolmogorov's axioms here, or at least an equivalent formulation thereof (taken from Benci, Horsten, & Wenmackers, 2013). P is the probability function and Ω is the sample space, a set whose elements represent elementary events:

(K0) DOMAIN AND RANGE. The events are the elements of $\mathfrak{A}$, a σ-algebra over Ω,[18] and P is a function $P : \mathfrak{A} \to \mathbb{R}$.

(K1) NON-NEGATIVITY. $\forall A \in \mathfrak{A}, P(A) \geq 0$.

(K2) NORMALIZATION. $P(\Omega) = 1$.

(K3) ADDITIVITY. $\forall A, B \in \mathfrak{A}$ such that $A \cap B = \varnothing$,

$$P(A \cup B) = P(A) + P(B).$$

(K4) CONTINUITY. Let $A = \bigcup_{n \in \mathbb{N}} A_n$, where $\forall n \in \mathbb{N}, A_n \subseteq A_{n+1} \subseteq \mathfrak{A}$. Then

$$P(A) = \sup_{n \in \mathbb{N}} P(A_n).$$

The triple $(\Omega, \mathfrak{A}, P)$ is called a *probability space*.

17 For the incorporation of probability theory into measure theory, Kolmogorov's assumption of Countable Additivity was crucial. This move was motivated by mathematical convenience, rather than by philosophical reflection on the meaning of probability. Kolmogorov stated (with original italics):

> Infinite fields of probability occur only as idealized models of real random processes. *We limit ourselves, arbitrarily, to only those models which satisfy Axiom VI.* (Kolmogorov, 1933, p. 15)

Later, de Finetti (1974, Vol. I, p. 119) would write about Countable Additivity:

> it had, if not its origin, its systematization in Kolmogorov's axioms (1933). Its success owes much to the mathematical convenience of making the calculus of probability merely a translation of modern measure theory [...]. No-one has given a real justification of countable additivity (other than just taking it as a "natural extension" of finite additivity) [...].

Compare to Schoenflies' reaction to Countable Additivity in Borel measure (footnote 57).

18 $\mathfrak{A}$ is a *σ-algebra* over Ω if $\mathfrak{A} \subseteq \mathcal{P}(\Omega)$ such that $\mathfrak{A}$ is closed under complementation, intersection, and countable unions. $\mathfrak{A}$ is called the *event algebra* or *event space*.

For our present purposes, the continuity axiom is the most important one, so let me briefly mention two aspects of it. First, (K4) uses a supremum, which is defined in terms of a standard limit; this limit is guaranteed to exist for real-valued functions, but not on the hyperreal numbers. Still, the general idea of this axiom can be phrased without reference to the specific limit operation. It can be regarded as a specific form of a more general idea: that is, to define the absolute probability of any event from an infinite domain as the limit (in some sense) of a sequence of conditional probabilities associated with that event, conditional on a suitable family of finite events. This more general principle was called the "Conditional probability principle" in Benci et al. (2013, section 3.2) and Benci, Horsten, and Wenmackers (2018, section 3.2), where it was further shown how the same idea can be applied to hyperreal-valued probability functions (using a different kind of limit operation). Second, assuming the other axioms, (K4) is equivalent to requiring countable additivity, which is not compatible with hyperreal-valued probability functions (except in the trivial case of a finite domain).

PART III
AXIOMATIZATION OF INFINITESIMAL PROBABILITIES

In the historical overview, we have already encountered two approaches to probability theory that allow infinitesimal probabilities: the axiomatization of Nelson (1987) and the work of Loeb (1975). What is missing so far is an axiomatization of a theory that assigns probabilities to standard infinite sets (such as $\mathbb{N}$, on which Nelson's approach is silent) and that allows infinitesimal or other hyperreal values in the final result (unlike Loeb's approach, which is geared toward obtaining results in the standard domain). This is the purpose of the current part.

8 INFINITESIMAL PROBABILITIES AND COUNTABLE LOTTERIES

Within philosophy, infinitesimal probabilities have often been discussed in the context of the following example: a lottery on the natural numbers, $\mathbb{N}$, in particular a fair one (*i.e.*, a lottery in which each individual ticket receives the same probability as any other one). Since this example is so common, we discuss it first, before setting up a more general framework in the next section.[19] We start from a real-valued approach (in which zero is

19 In order to describe probability functions on infinite sample spaces, focusing on $\mathbb{N}$ as the sample space may seem like a very natural starting point, because $\mathbb{N}$ is the canonical example of a set with the smallest infinite cardinality. It will turn out that in some sense this problem is not the easiest one to describe, because it is in lockstep with other (less

the only infinitesimal) and investigate which modifications are required in order to allow for the assignment of non-zero infinitesimal probabilities.[20]

8.1 *Lotteries on Initial Segments of* $\mathbb{N}$

Ultimately, we want to describe a lottery, fair or weighted, on $\mathbb{N}$, but we start by considering a lottery, fair or weighted, on an arbitrary initial segment of $\mathbb{N}$: the sample space (set of atomic possible outcomes) is $\Omega_n = \{1, \ldots, n\}$. First, we introduce weights: a real number w_i for each of the elements i of Ω_n. Without loss of generality, we may assume these weights to be normalized, such that $\sum_{i=1}^{n} w_i = 1$ (*e.g.*, in a fair lottery $w_i = 1/n$ for all i). Then, we define the probability on Ω_n, P_n, of an arbitrary subset of $\mathbb{N}$, A, as follows:

$$P_n(A) = \sum_{i=1}^{n} w_i \times \#(A \cap \{i\}),$$

where # is the counting measure for finite sets. (This suffices: although A can be an infinite set, $A \cap \{i\}$ is empty or singleton.) In the case of a fair lottery, the probability $P_n(A)$ is just the relative frequency of A: the fraction of elements of A within Ω_n. That P_n is finitely additive follows directly from the counting measure being finitely additive.[21]

8.2 *Taking the Limit*

Now, we want to consider a lottery on $\Omega = \mathbb{N}$, rather than on $\Omega_n = \{1, \ldots, n\}$. The idea is to consider the lottery on $\mathbb{N}$ as the limiting case of a sequence of finite lotteries. This idea seems apt, since we have $\Omega =$

---

obvious) occurrences of $\mathbb{N}$. Among the infinite sets, $\mathbb{N}$ is our usual benchmark, so we use it in and out of season. As a result, there are hidden symmetries in the problem of a (fair) lottery on $\mathbb{N}$, which make it harder to analyze it. To understand this statement, we first need to encounter the problems alluded to, so we will progress as planned, but I will return to this observation in the middle of section 8.3.

20 The current section presents some of the ideas originally developed in Wenmackers and Horsten (2013) in a more straightforward way.

21 For, consider a finite family of mutually disjoint subsets of $\mathbb{N}$, $\{A_k \mid k \in \{1, \ldots, m\}, A_k \subseteq \mathbb{N}\}$ (for some $m \in \mathbb{N}$) such that for each $i \neq j$, $A_i \cap A_j = \varnothing$. Defining the union of members of the family $A = \bigcup_{k=1}^{m} A_k$, we obtain for the probability of A:

$$\begin{aligned} P_n(A) &= \textstyle\sum_{i=1}^{n} w_i \times \#(\bigcup_{k=1}^{m} A_k \cap \{i\}) \\ &= \textstyle\sum_{i=1}^{n} w_i \times \sum_{k=1}^{m} \#(A_k \cap \{i\}) \\ &= \textstyle\sum_{k=1}^{m} \sum_{i=1}^{n} w_i \times \#(A_k \cap \{i\}) \\ &= \textstyle\sum_{k=1}^{m} P_n(A_k). \end{aligned}$$

$\lim_{n\to\infty} \cup_{i=1}^{n} \Omega_i$.[22] We will define the probability, P, for an arbitrary subset of $\mathbb{N}$, A, analogously to the limiting relative frequency:

$$P(A) = \lim_{n\to\infty} P_n(A).$$

Remarks:

- P is not defined for all subsets of $\mathbb{N}$.[23]
- Taking the limit of fair lotteries on Ω_n (where $P(\{i\}) = 1/n$ for any $i \in \Omega_n$) results in a fair lottery on $\mathbb{N}$, with $P(\{i\}) = 0$ for all $i \in \mathbb{N}$.
- For a fair lottery on $\mathbb{N}$, P is the *natural density* (also known as the *arithmetic density* or the *asymptotic density*).
- In a fair lottery, P is zero for all finite subsets as well as for some infinite ones (such as the set of squares and the set of primes),[24] unity for cofinite sets as well as for some infinite ones (such as the complements of the previous examples), and intermediate values for other infinite sets (such as arithmetic progressions[25] that receive probability $1/n$ for some n; *e.g.*, $1/2$ for the set of even numbers and for the set of odd numbers).

For those who have the intuition that the probability of a particular outcome in a fair lottery on the natural numbers ought to be *infinitesimal*, the above real-valued function P that assigns probability zero to such outcomes does fine: zero is *the* infinitesimal probability, the only one in the $[0, 1]$ interval of $\mathbb{R}$. Nevertheless, it may bother some that this function does not allow us to distinguish between the impossible event (represented by $A = \varnothing$) and some infinitely unlikely but possible events. The worry is that

22 On the other hand, the ordered set $(\mathbb{N}, <)$ is qualitatively different from any $(\Omega_n, <)$: unlike all of its initial segments, $\mathbb{N}$ does not have a last element. This observation is suggestive of taking a different kind of limit, which involves a hyperfinite set (which does have a last element) rather than a standard infinite one.

23 The collection of subsets for which P is defined does not form a σ-algebra. P can be extended to all of $\mathcal{P}(\mathbb{N})$ but the extension relies on Banach limits and is not unique. Whereas the usual limit relies on the notion of "eventuality" that can be captured by the Fréchet filter, which is a free filter that is constructively available, the Banach limit depends on a free ultrafilter on $\mathbb{N}$, which relies crucially on a non-constructive axiom (the ultrafilter principle, UF). See section 8.5 below for more details.

24 As such, this probability function can help us to make sense of Galileo's paradox, which revolves around the question of whether or not the set of perfect squares is smaller than the set of natural numbers (see Mancosu, 2009). As measured by the natural density, the answer to that question is affirmative: it assigns probability unity to the set of natural numbers and probability zero to the set of perfect squares. On the other hand, the function does not discriminate between a finite set, the set of perfect squares, and the set of primes.

25 Arithmetic progressions are sets of the form $a\mathbb{N} + b = \{n \in \mathbb{N} \mid n \mod a = b\}$ for some $a \in \mathbb{N}$ and some $b \in \{0, 1, \ldots, a-1\}$.

the probabilities of these events are represented by the same infinitesimal, and since there can only be one zero (*i.e.*, neutral element under addition), this observation may motivate a search for *non-zero infinitesimals*. However, this worry may be partially addressed by considering a non-Archimedean ordering of the events, which is a question for qualitative probability theory[26] rather than for quantitative probability theory. Despite this, there is an underlying issue that cannot be addressed without considering numerical probabilities: it is that of additivity. We consider this in the next section.

8.3 *Additivity of P: Finite, Countable, or Ultra*

We briefly mentioned (section 5) that Leibniz's approach to the calculus was based on infinite sums (integrals), unlike Newton's, for whom the notion of "fluxions" (derivatives) was more basic. Since infinitesimals were most prominent in Leibniz's approach, it should come as no surprise that the concept of infinitesimal probabilities is closely connected to foundational discussions concerning the additivity of probability values.

Skyrms (1983b) interprets the intuition that measures should be regular (that only the null set should receive measure zero) as a Zenonian intuition (*cf.* section 16.3 of the appendix): a whole of positive magnitude should not be made up of parts of measure zero. He argues that a principle of "ultra-additivity"[27] has been present, albeit often implicitly, in discussions concerning measures at least since the times of Zeno and Aristotle. Since the belief in ultra-additivity appears to be so deeply rooted in Western thinking about measures, it should not surprise us if it is present, whether presented as an explicit assumption or a tacit one, in many discussions about probability measures, too.

In fact, it was exactly such a principle that motivated my own search for a fair probability function on $\mathbb{N}$. My main motivation for wanting to assign non-zero probability to non-empty sets is that it should allow us to make arbitrary unions of events and obtain their probability by an addition rule for the individual probabilities (in the case of disjoint events, by taking the analogous arbitrary sum).[28]

---

26 Recall the work by de Finetti (1931) as discussed in section 1. See also Pedersen (2014), Easwaran (2014, p. 17), and Konek (this volume).

27 Ultra-additivity means additivity for arbitrary collections of disjoint events; it is sometimes called perfect additivity (see, *e.g.*, de Finetti, 1974, Vol. II, p. 118) or arbitrary additivity (Hofweber, 2014).

28 Wenmackers (2011, p. 36): "Intuitively, one could expect probabilities to exhibit perfect rather than countable additivity. However, this is clearly not possible with real-valued probability functions. Even the weaker requirement of countable additivity may be problematic, as we have seen in the example of the infinite lottery. Yet, the property of perfect additivity may be attainable by non-Archimedean probabilities." Unaware of the work

Let us return to the probability functions of the previous sections. Finite additivity obtains for such a P, like it does for all the functions P_n. Since the function P is the limit of the sequence of functions (P_n), each member of which has the property of finite additivity (FA), one might suspect P to have the limiting property of FA: countable additivity (CA). However, this is not the case: limiting relative frequencies are not CA, because the relevant limiting operations (from the construction of P and from the condition of CA) do not commute. To illustrate this, consider a countably infinite family of mutually disjoint subsets of $\mathbb{N}$, $\{A_k \mid k \in \mathbb{N}, A_k \subseteq \mathbb{N}\}$ such that for each $i \neq j$, $A_i \cap A_j = \varnothing$, and define the union of members of the family, $A = \bigcup_{k\in\mathbb{N}} A_k$. We say that CA holds for a function p if the following equality holds:

$$p(A) = \lim_{n\to\infty} \sum_{i=1}^{n} p(A_i). \tag{2}$$

In the case of P, we find for the lefthand-side of equation (2):

$$\begin{aligned} P(A) &= \lim_{n\to\infty} P_n(A) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} w_i \times \lim_{m\to\infty} \sum_{k=1}^{m} \#(A_k \cap \{i\}). \end{aligned}$$

Let us now consider a fair lottery (substituting $w_i = 1/n$) with $A_k = \{k\}$ such that $A = \mathbb{N}$; we find:

$$\begin{aligned} P(A) &= \lim_{n\to\infty} (n \times 1/n) \\ &= 1. \end{aligned}$$

Then, we consider the righthand-side of equation (2), applying it to P in the fair case, where $P(A_i) = 0$ for all i:

$$\begin{aligned} \lim_{n\to\infty} \sum_{i=1}^{n} P(A_i) &= \lim_{n\to\infty} \sum_{i=1}^{n} 0 \\ &= 0. \end{aligned}$$

Clearly, 0 is not equal to 1, so CA does not obtain for P, the real-valued probability function for a fair lottery on the natural numbers.

by Skyrms (1983b), Wenmackers and Horsten (2013, p. 40) clumsily referred to a "SUM" intuition: "SUM [is the intuition that] [t]he probability of a combination of tickets can be found by summing the individual probabilities. [...] The assumption SUM is motivated by the intuition that the probability of a set containing the winning number supervenes on the chances of winning that accrue to the individual tickets. The usual assumption of countable additivity (CA, sometimes also called σ-additivity) is one attempt of making the intuition that is encapsulated by SUM precise. We will argue, however, that this is not the right way to do it in this case. In other words, we will argue that the implementation of SUM is not as straightforward an affair as is commonly thought."

The righthand-side requires us to consider the function P and thus to take the limit of n to infinity of $P_n(\{i\}) = 1/n$ first, which is zero; taking the limit of a sum of zeros is zero. The lefthand-side requires us to consider P_n. Sure, as n increases, $P_n(\{i\})$ tends to zero for any $i \in \Omega_n$ (like $1/n$), but the sum of all singleton probabilities is in lock-step with this decrease: $n \times 1/n = 1$, such that the sum of probabilities of all singletons equals the probability of the entire sample space (total number of tickets times probability of each ticket), which is unity. This is just FA and it holds for any n, no matter how large. It also holds that $\lim_{n\to\infty}(n \times 1/n) = 1$, but this cannot be read as "the number of tickets times the probability of each ticket." It is no additivity principle and it does not suggest an alternative way of obtaining a real-valued probability function either.[29] Yet, it does suggest the following: that the singleton probability in a fair lottery on the natural numbers ought to be a non-zero infinitesimal, such that some sort of infinite sum over them can result in a non-zero (and non-infinitesimal) value corresponding to the probability of the corresponding union of events. In particular, the sum can be unity if we sum the probabilities of all point events.[30]

There is another strange aspect to setting $P(\{n\}) = 0$ for all $n \in \mathbb{N}$: it is not so much that it can be used to represent a fair lottery on $\mathbb{N}$, but rather that it can also represent the limit of many kinds of non-fair probability distributions. Consider, for instance, finite lotteries in which (i) the set of even numbers is double as likely as the set of odd numbers, (ii) all even numbers are equally likely and (iii) all odd numbers are equally likely. For the limit of such weighted lotteries, too, we would have to assign probability zero to all singleton events (and thus obtain a fair distribution in the limit).[31]

8.4 *Diagnosis*

Within the context of standard probability theory, we have a single infinitesimal probability at our disposal: zero. Even for a lottery on a sample space that is countably infinite, the lowest infinite cardinality, this turns out to be too little for three reasons.

1. Across lotteries, it does not allow us to obtain different singleton probabilities for limits of sequences of qualitatively different finite

29 Although this idea is suggestive of a *procedure* for assigning probabilities in such a way that we can make sense of infinite sums, it does not allow us to define a probability *function*.

30 Recall the quote on p. 199 by de Finetti (1974, p. 347) concerning the absurdity of $0 + 0 + 0 + ... + 0 + ... = 1$. It turns out that this idea is false if the sum represents the usual, countably infinite sum: such a sum is not defined for infinitesimal terms.

31 As far as I know, this worry has not yet appeared in the literature.

lotteries (*e.g.*, finite lotteries that assign equal probability to even and odd versus finite lotteries that do not).

2. Within a fair lottery, it does not allow us to discriminate between the probability of many events that are strict subsets of each other (*e.g.*, all perfect squares versus a single perfect square).

3. Within a fair lottery, it does not allow us to define an adequate infinite additivity principle; alternatively, if we insist on countable additivity, it does not allow us to describe a fair lottery on the natural numbers.

The first reason is related to a more general observation: like any real number, zero is the limit of qualitatively different sequences (of rational or real numbers). In particular, sequences may differ in their speed of convergence. The rate at which the corresponding sequence of relative frequencies tends to zero is smaller for a singleton event (the convergence of $1/n$ is linear) than it is for the set of perfect squares (the convergence of $1/n^2$ is quadratic) or for the powers of two (the convergence of $1/2^n$ is exponential). This suggests that within the collection of sequences that are considered to be infinitesimal, and thus to converge to zero, some are smaller than others (even though their limits are all defined to be zero when working within the real numbers). This brings us to reconsider what the real number zero is, continuing along the lines set out in the introduction, and to define an alternative limit operation on sequences. One way to achieve this is found in the construction of a non-standard model of a real closed field as was shown in section 6.

8.5 *Alternative Approach with Non-Zero Infinitesimal Probabilities*

We apply the equivalence relation discussed that is used to construct the hyperreals (section 6) to the sequence of relative frequencies belonging to initial segments of $\mathbb{N}$. This results in a different kind of probability function, which takes its values in the $[0, 1]$ interval of the hyperreal numbers.[32]

Wenmackers and Horsten (2013) assumed all of NSA as given, whereas we mainly needed this alternative equivalence relation on the sequences of relative frequencies in order to obtain a hyperreal-valued probability value on $\mathbb{N}$ that allows for an infinite additivity principle.

Now that we know the outlines of our labyrinth, we can drastically reduce the length of our escape route. With the benefit of hindsight, we

32 Actually, it is more accurate to say: *a* set of hyperreal numbers (*cf.* footnote 1), because the result of the construction depends on the free ultrafilter and there are uncountably many. We do not dwell on the issue of non-uniqueness now, but we will come back to it in section 14.

see ways to obtain our results with much less baggage. One way, which is suitable only for fair lotteries and which is alluded to in the 2013 paper, is to assume a numerosity function on $\mathbb{N}$ and to normalize it. Numerosity theory has been developed to address some of the very same problems that are also discussed in the literature on a fair lottery on $\mathbb{N}$ (Benci & Di Nasso, 2003; Mancosu, 2009). The main difference is that it is not a probability function but a measure of set size that should coincide with the usual counting measure for finite sets, so it is not normalized and assigns unity to singletons rather than to $\mathbb{N}$. However, because of its nice algebraic properties, normalizing the numerosity function, in order to obtain a fair probability measure, does not cause any complications at all.

Alternatively and more elegantly, one could set up an axiomatic system that states the existence of probability functions on $\mathbb{N}$ that may assign non-zero values to singleton outcomes (possibly all equal) and repurpose the previous results in order to prove its consistency.

For instance, consider this proposal for the axioms governing P.

> Everywhere defined. P is defined on all subsets of $\mathbb{N}$: its domain is the powerset of $\mathbb{N}$, $\mathcal{P}(\mathbb{N})$.
>
> Hyperreal-valued. The range of P is the unit interval of some suitable field $\mathcal{R}$.
>
> Regular. $P(A) = 0$ iff $A = \varnothing$.
>
> Normalized. $P(\mathbb{N}) = 1$.
>
> Finitely additive. $\forall A, B \in \mathcal{P}(\mathbb{N})$ if $A \cap B = \varnothing$, then $P(A \cup B) = P(A) + P(B)$.
>
> Ultra-additive. For any collection of mutually disjoint subsets of $\mathbb{N}$[33] an analogous additivity property holds.

We do not prove the joint consistency of the proposed axioms here: it is a consequence of what preceded and can be viewed as a special case of the proof in Benci et al. (2013).

8.6 *Examples*

Now that we have seen that there exists a hyperreal measure that captures the idea of a uniform probability distribution over the natural numbers, let's illustrate some consequences. In this section, P always refers to such a distribution. (For proofs, see Benci et al. 2013.)

33 The collection can have an arbitrary cardinality, although, of course, at most countably many of its members can be non-empty.

By assumption, P assigns the same infinitesimal probability to any singleton outcome of the lottery. If we regard P as a normalized numerosity function, we see that $\forall n \in \mathbb{N},\ P(\{n\}) = 1/\alpha$, where $\alpha \in {}^*\mathbb{N} \setminus \mathbb{N}$ is the numerosity of $\mathbb{N}$.

For any finite set $A \subset \mathbb{N}$, the numerosity equals the finite cardinality (#), so: $P(A) = \#(A)/\alpha$, which is an infinitesimal. For example, $P(\{1, 2, 4, 8, 16, 32\}) = 6/\alpha$.

For an infinite subset B, $P(B)$ differs by at most an infinitesimal from the natural density of B (if the latter exists). For example, if B is the set of even numbers, the natural density is $1/2$ and either $P(B) = 1/2$ (if the even numbers are in the free ultrafilter used to construct P) or $P(B) = (1 - 1/\alpha)/2$.

For a set that lacks a natural density, P is infinitesimally close to some Banach limit. Different Banach limits of the same set and Ps constructed by a different free ultrafilter can differ by more than an infinitesimal amount. (See Kerkvliet and Meester, 2016, for an example.) In particular, there are subsets of $\mathbb{N}$ for which the possible P-values range from an infinitesimal to one minus an infinitesimal. This range can be regarded as a measure of how pathological a set is.

9 MORE SCENARIOS INVOLVING INFINITESIMAL PROBABILITIES

In the previous section, we discussed one particular scenario that involves infinitesimal probabilities: a lottery on the set of natural numbers. In this section, we give a more comprehensive overview of common examples that feature in discussions of infinitesimal probabilities. Then we show how we can generalize the approach of the previous section to an all encompassing theory that is able to assign infinitesimal probabilities to all of these scenarios.

9.1 *Common Examples*

We list the examples involving infinitesimal probabilities below, sorted by increasing cardinality of the sample space: finite, countably infinite, or uncountably infinite.

First, there are some examples with finite sample spaces that allow for infinitely small differences in probability among the possible outcomes. The simplest such case is that of an *almost* fair coin toss, in which there is an infinitesimal advantage to one of the sides.

Second, there are examples with countably infinite sample spaces, in particular with uniform probability distributions. We already discussed the most common example of this kind: a lottery on the set of natural numbers,

in particular a fair one. A fair lottery on $\mathbb{N}$ is also known as the de Finetti lottery (Bartha, 2004) or God's lottery (McCall & Armstrong, 1989). In this category, there are also fair lotteries on other countable sets, such as $\mathbb{Z}$, $\mathbb{Q}$, and the unit interval of the rational numbers: $[0,1]_{\mathbb{Q}}$. Discussions of non-uniform probability distributions on countable domains are less common, but they do exist, especially in the context of discussions of the incompatibility between CA and uniform probability distributions on countable domains.[34]

Third, there are examples with uncountable sample spaces, with uniform and arbitrary probability distributions. Two popular ways of presenting this is as throwing darts uniformly at the unit interval of the real numbers, $[0,1]_{\mathbb{R}}$ (*e.g.*, Bernstein & Wattenberg, 1969) or as a fair spinner with unit circumference (*e.g.*, Skyrms, 1995; Barrett, 2010).[35] Three-dimensional variations on this theme include the uniform probability on a unit sphere and the associated Borel–Kolmogorov paradox of a meridian versus the equator. A different way of obtaining an uncountable domain is by considering a countably infinite sequence of stochastic processes, each with a countable number of possible outcomes. The most common example of this kind is an infinite sequence of tosses with a fair coin (in which the outcomes of the tosses are taken to be statistically independent: an infinite Bernoulli process; *e.g.*, Skyrms, 1980; Williamson, 2007; Weintraub, 2008).[36]

Categorizing a probabilistic problem by one of these three labels need not be final. Once we have a method of representing probability distributions on uncountable domains, we may arrive back at the finite and countably infinite case by conditionalization (assuming the relevant events are measurable; *cf.* Skyrms, 1983b). It may also happen that we want to replace a finite sample space by an infinite refinement of it (for instance, a suitable product space of the initial sample space). For instance, Pedersen (2014, p. 827) mentions a case in which "an agent's state of belief cannot rule out arbitrarily deep[ly] nested subdecompositions of a finite decomposition of a dartboard."

34 For instance, Kelly (1996) has reflected on the consequences of denying the existence of a fair infinite lottery: this would have the strange implication that when one wants to test a universal hypothesis by repeated experiments, one would—in the case in which the hypothesis is false—encounter a counterexample sooner rather than later.

35 This example was also mentioned in Lewis (1980), as well as many others.

36 It should be noted that Skyrms (1980) refers to the work of Bernstein and Wattenberg (1969), but they only described a hyperreal-valued probability measure on subsets of $[0,1]$. However, for assigning infinitesimal probabilities to infinite sequences of coin tosses, a hyperreal-valued probability measure on subsets of $\{0,1\}^{\mathbb{N}}$ would be needed instead. Yet, the informal account given by Skyrms (1980, pp. 30–31) is consistent with later developments of hyperreal probability functions on $\{0,1\}^{\mathbb{N}}$ (see, *e.g.*, Benci et al., 2013).

Some of these scenarios cannot be described by standard probability theory, whereas others—it has been argued—cannot be described adequately by it, or would benefit from an alternative treatment involving infinitesimal probabilities. So far, we have seen isolated recipes for hyperreal-valued probability functions: Bernstein and Wattenberg (1969) gave a recipe to assign uniform probabilities to subsets of the unit interval of the real numbers. And, in the previous section, we discussed a recipe for assigning regular probabilities to the canonical countably infinite sample space, $\mathbb{N}$. In the end, we would like to have a method that is fully general, which can be applied to all the examples above, and more. We describe such a method below.

9.2 *Non-Archimedean Probability (NAP) Theory*

In this section, we will review some crucial elements that allow us to generalize the approach from section 8.[37] In section 8.5, we replaced the standard limit operation that associates at most one real number with a sequence of (possibly weighted) relative frequencies by a non-standard limit that associates a hyperreal number with each of these sequences. Sequences can be thought of as functions from $\mathbb{N}$ (the index set) to some set, X. In the case of relative frequencies $X = \mathbb{Q}$, but in general we allow real-valued weights, so then $X = \mathbb{R}$. Both the standard and the non-standard limit operation can be understood such as to involve a filter on the index set (the Fréchet filter on $\mathbb{N}$ and a free ultrafilter on $\mathbb{N}$, respectively).

A probability function has to assign values to sets in $\mathcal{P}(\mathbb{N})$, not to $\mathbb{N}$ itself, so the appropriateness of using countable sequences and filters on $\mathbb{N}$ to set up such a function is not immediate, even in cases in which the sample space is countable. Observe that we used the countable indices to correspond to the relative frequencies of initial segments of $\mathbb{N}$. Since the usual ordering of the natural numbers induces a natural ordering on this collection of initial segments, we are able to work with sequences of the corresponding relative frequencies and with filters on $\mathbb{N}$.

Our choice for the collection of initial segments may seem self-evident, because we are familiar with it from the context of natural density, but it is not canonical: we could have considered $\mathcal{P}_{\text{fin}}(\mathbb{N})$, the collection of all finite subsets of $\mathbb{N}$ (or those except the empty set, $\mathcal{P}_{\text{fin}}(\mathbb{N}) \setminus \varnothing$). In that case, we can slightly generalize the approach: $\mathcal{P}_{\text{fin}}(\mathbb{N})$ with the subset

37 The information given here suffices to get a rough idea of the approach. Further details (for instance, restrictions on the free ultrafilter to secure certain properties of the resulting probability functions) can be found in Benci et al. (2013).

ordering forms a directed set.[38] We can use this directed set as an index set, instead of $\mathbb{N}$, obtaining a generalized sequence, also called a *net* (see, *e.g.*, Schechter, 1997, pp. 157–158): a function from a directed set, which serves as the index set, to a set, X. Filters on $\mathbb{N}$ are a special case of this more general setup, since they are collections of subsets of $\mathbb{N}$ that can be directed by the subset relation.

If we want to assign probability functions to subsets of some sample space Ω other than $\mathbb{N}$, we can follow a similar approach: change the relevant index set to $\mathcal{P}_{\text{fin}}(\Omega) \setminus \varnothing$. In this case, we also have to consider free ultrafilters on Ω.

These are the axioms for Non-Archimedean Probability (NAP) theory from Benci et al. (2013), where the triple (Ω, P, J) is called a *NAP space*:

(No) Domain and range. The events are all the elements of $\mathcal{P}(\Omega)$ and P is a function

$$P : \mathcal{P}(\Omega) \to \mathcal{R}$$

where $\mathcal{R}$ is a superreal field.

(N1) Non-negativity. $\forall A \in \mathcal{P}(\Omega)$, $P(A) \geq 0$.

(N2) Normalization. $\forall A \in \mathcal{P}(\Omega)$, $P(A) = 1 \Leftrightarrow A = \Omega$.

(N3) Additivity. $\forall A, B \in \mathcal{P}(\Omega)$ such that $A \cap B = \varnothing$,

$$P(A \cup B) = P(A) + P(B).$$

(N4) Non-Archimedean Continuity. $\forall A, B \in \mathcal{P}(\Omega)$, with $B \neq \varnothing$, let $P(A|B)$ denote the conditional probability, namely

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Then

- $\forall \lambda \in \mathcal{P}^0_{\text{fin}}(\Omega)$, $P(A|\lambda) \in \mathbb{R}^+$, and
- there exists an algebra homomorphism

$$J : \mathfrak{F}\left(\mathcal{P}^0_{\text{fin}}(\Omega), \mathbb{R}\right) \to \mathcal{R}$$

such that $\forall A \in \mathcal{P}(\Omega)$, $P(A) = J\left(\varphi_A\right)$, where $\varphi_A(\lambda) = P(A|\lambda)$ for any $\lambda \in \mathcal{P}^0_{\text{fin}}(\Omega)$.

38 A directed set $(X, \preccurlyeq)$ is a special case of a preordered set (see, *e.g.*, Schechter, 1997, p. 52). A preordered set is a pair $(X, \preccurlyeq)$ consisting of a set X and a preorder $\preccurlyeq$ on X, *i.e.*, a relation on X that is transitive (for all $x, y, z \in X$, if $x \preccurlyeq y$ and $y \preccurlyeq z$ then $x \preccurlyeq z$) and reflexive (for all $x \in X$, $x \preccurlyeq x$). For a directed set, there is an additional condition on the preorder:

$$\forall x_1, x_2 \in X,\ \exists y \in X : (x_1 \preccurlyeq y \wedge x_2 \preccurlyeq y).$$

Axiom (N4) specifies P for an infinite sample space Ω as a non-standard limit of probability functions restricted to (or conditionalized on) finite subsets of Ω.

Some properties of NAP theory:

- NAP theory produces regular probability functions. Hence, they allow us to conditionalize on any possible event by a ratio formula (*i.e.*, any subset of the sample space, except the empty set).
- Within NAP theory, the domain of the probability function can be the full powerset of any standard set from applied mathematics (*i.e.*, of any cardinality), whereas the general range is a non-Archimedean field. Hence, there are no non-measurable sets.
- Kolmogorov's countable additivity (which is a consequence of the use of standard limits) is replaced by a different type of infinite additivity (due to the use of a non-Archimedean limit concept).
- For fair lotteries, the probability assigned to an event by NAP theory is directly proportional to the numerosity of the subset representing that event.
- NAP functions are external objects: they cannot be obtained by taking a standard object (such as a family of standard sets) and applying the star-map to it.

A price one has to pay for all this is that certain symmetries, which hold for standard measures, do not hold for NAP theory. This theory is closely related to numerosity and has a similar Euclidean property: a strict subset has a smaller probability, as is necessary by regularity. Hence, for infinite sample spaces, NAP is bound to violate the Humean principle of one-to-one correspondence. This principle requires that if the elements of a given set can be put in a one-to-one correspondence with the elements of another set, then their "sizes"—or in this case, probabilities—will be equal. Translation symmetries require that $P(A) = P(A + t)$ (with $A, A + t \subseteq \Omega$ and $A + t = \{a + t \mid a \in A\}$). Since this amounts to a particular type of one-to-one correspondence, these symmetries are not guaranteed to hold in NAP (*cf.* Williamson 2007; Parker 2013; and section 14.1), although they can hold up to an infinitesimal (Bernstein & Wattenberg, 1969). Bartha (2004) and Weintraub (2008) have pointed out before that these measures are strongly label-dependent, but it is probably more accurate to say that once events have been embedded in a sample space (*i.e.*, each event is described as a particular subset of a particular sample space Ω), this embedding needs to be applied in a consistent way henceforth (Hofweber, 2014; Benci et al., 2018).

For more details and proofs, see Benci et al. (2013). In the next part, we elaborate on the motivation for and the philosophical discussion of these results.

PART IV
PHILOSOPHICAL DISCUSSION

10 MOTIVATIONS FOR INFINITESIMAL PROBABILITIES

In the foregoing parts, we have encountered motivations for introducing infinitesimal probabilities as given by various authors. Most of these motivations occurred in the context of a particular interpretation of probability, with some arguing for the relevance of infinitesimal chances and others advocating for the introduction of infinitesimal credences. In this section, we search for the leitmotifs that arise from this polyphony.

Let us first revisit Bernstein and Wattenberg (1969): although they gave a probabilistic scenario as the motivation of their paper, the technical details of their results do not depend on the interpretation in terms of probability. If we want a measure of length that allows us to represent the length of countable collections of points as a non-zero infinitesimal, we can use the result of Bernstein and Wattenberg (1969) without modification. On the one and, it may fit even better in such a context, since the Lebesgue measure was originally motivated as an idealization of length measurements. Hence, obtaining a non-standard measure that is infinitely close to Lebesgue measure (at least, where the latter is defined) can be regarded as an alternative idealization of length measurements. On the other hand, the request for representing the measure of non-null countable sets as an infinitesimal may seem especially pressing when this measure is a measure of probability (rather than length). This motivation may be formulated as follows: probability measure should be maximally sensitive to distinguish possibility from impossibility. Indeed, we have encountered this motivation for infinitesimal probabilities via regularity at various instances throughout this chapter.

Depending on the context, this motivation is related to a different kind of modality:

- objective probability: some chance (quantifying an ontic possibility);
- subjective probability: open-mindedness (quantifying an epistemic possibility).

We have encountered the epistemic motivation under the names 'strict coherence' and 'regularity'. Hájek (2012b, p. 1 of draft) "canvass[es] the fluctuating fortunes of a much-touted constraint, so-called *regularity*," which "starts out as an intuitive and seemingly innocuous constraint that bridges modality and probability, although it quickly runs into difficulties in its exact formulation." He takes "to be its most compelling version: a constraint that bridges doxastic modality and doxastic (subjective) probability." Easwaran (2014) presents regularity as a normative constraint on rational credences, which are related to doxastic modality, but he adds that other authors allow for various transmodal connections. Dennis Lindley called this demand, that prior probabilities of zero or one should only be assigned to logical truths or falsehoods, "Cromwell's rule."[39] Regarding the ontic motivation, Hofweber (2014) introduces a minimal constraint (MC) on the proper measurement of chances, which is akin to but not quite the same as regularity, which can be phrased in relation to various modalities. He concludes that: "In the regularity principle, modality is best understood as epistemic, and chance is best understood as credence. In (MC) chance should be understood as objective chance" (p. 6).

At the root of this common motivation for infinitesimal chances and infinitesimal credences, there may be an even more basic motivation or implicit assumption, which Skyrms (1983b) calls the principle of "ultra-additivity" (and which also constituted my main motivation for starting a research project on infinitesimal probabilities). We discussed this in section 8.3 (see also appendix section 16.3). Thus, the motivation for introducing infinitesimal probabilities can be summarized by the following slogan:[40]

Without infinitesimals, probabilities just don't add up.

39 This is a reference to the following phrase from a 1650 letter by Oliver Cromwell: "I beseech you, in the bowel of Christ, think it possible you may be mistaken" reprinted in, Carlyle (1845). Like strict coherence, Cromwell's rule is clearly intended as a criterion for open-mindedness: even a well-confirmed theory like Einstein's is not as certain as a logical truth. Lindley (1991, p. 104) asks us to "leave a little probability for the moon being made of green cheese; it can be as small as 1 in a million, but have it there since otherwise an army of astronauts returning with samples of the said cheese will leave you unmoved." And Lindley (2006, p. 91) links this open-mindedness criterion also to the Jain maxim "It is wrong to assert absolutely." (This was probably influenced by statistician Kantilal Mardia, who practised Jainism.)

40 Benci et al. (2018) list perfect additivity as one among four desiderata for their theory, the others being: regularity, totality, and weak Laplacianism.

11 ALTERNATIVES TO HYPERREAL PROBABILITIES

11.1 *Other Ways to Introduce Infinitesimal Probabilities*

There do exist ways to formalize infinitesimals other than Robinson's hyperreal numbers. One of them is smooth-infinitesimal analysis (SIA), which describes nilpotent infinitesimals: non-zero numbers whose square is zero. This system relies on intuitionistic logic. However, I am not aware of any proposals for smooth-infinitesimal probabilities.

Then there is the class of Conway numbers, which includes the infinitesimals from any non-standard field. This option has been suggested for application to probability theory, for instance, by Hájek (2003; see section 12 below) and by Easwaran (2014). I, too, believe this can be a fertile approach. A first proposal has been offered by Chen and Rubio (2018), but it is too early to evaluate it here.

11.2 *Related Approaches Without Infinitesimals*

Besides the possibility of introducing infinitesimals within a different framework, there are also relations between hyperreal infinitesimals and systems that do not include any infinitesimal numbers at all. For instance, one may combine an Archimedean quantitative probability theory (in particular, the orthodox approach with real-valued probability functions), with a non-Archimedean qualitative probability theory.[41] Moreover, Halpern (2010) reveals some deep connections between hyperreal-valued probability functions, conditional probabilities (including Popper functions; see also Vann McGee, 1994), and lexicographic probabilities. Recently, Brickhill and Horsten (2018) have given a representation theorem that relates NAP functions and Popper functions; they also give a lexicographic representation.

Skyrms (1983a) considers three ways of giving probability assignments a memory. One of his proposals was to "utilize orders of infinitesimals to implement long term-memory," such that "[s]uccessive updatings do not destroy information, but instead push it down to smaller orders of infinitesimals" (p. 158). He evaluates this proposal as having a certain theoretical simplicity, but lacking practical feasibility. However, given that the proposal essentially boils down to introducing lexicographical probabilities, it may turn out that this judgment was too harsh.

41 This was suggested by de Finetti, *cf*. section 1. See also the discussion of the "numerical fallacy" by Easwaran (2014).

11.3 *Yet Another Point of View*

Introducing non-standard probabilities amounts to changing the range of the probability function. Skyrms (1995) considers an alternative way to achieve strict coherence, which involves changing the domain, such that the events to which infinitesimal probabilities are assigned in the previous approach are no longer in the event space at all. In this context, he cites (Jeffrey's translation of) Kolmogorov (1948):

> The notion of an elementary event is an artificial superstructure imposed on the concrete notion of an event. In reality, events are not composed of elementary events, but elementary events originate in the dismemberment of composite events.

Let me unpack this. In Kolmogorov's (1933) approach, the sample space was assumed to contain all fully specific possible outcomes: the elements of the sample space are called "elementary events." On the other hand, we have the informal notion of concrete events or possible outcomes, which does not presuppose infinite precision. Here we see that Kolmogorov (1948) rejected his former approach in favour of a more realistic one: if we take into account the limited precision of any physical measurement, we can distinguish outcomes only with limited precision, too. With increasing precision, we can decompose events into more fine-grained ones, but not up to elementary precision.

Although no infinitesimal probabilities occur in the second approach, it is still relevant in the context of the current chapter, because of an interesting analogy: in both cases, starting from the orthodox approach, a symmetry is quotiented out to arrive at the new structure (*cf.* the reference to quotient spaces in the introduction).

12 INTERPLAY BETWEEN INFINITESIMAL PROBABILITIES AND INFINITE UTILITIES: PASCAL'S WAGER

We have seen in section 3, that discussions of rational degrees of belief often proceed via a betting interpretation (*e.g.*, motivating adherence to the axioms of probability theory by the avoidance of a sure loss). As such, they involve considerations of monetary loss or gain. However, the subjective value of money need not be linear. Therefore, it is useful to introduce *utility* as a more abstract measure that represents subjective worth directly. Utility is usually taken to be a real-valued (interval scale) measure.

However, non-Archimedean probabilities do not mix well with real-valued utilities. Hence, to deal adequately with infinitesimal probabilities in the context of decision theory, a non-Archimedean utility theory is needed, such as the one developed by Pivato (2014).

We consider the famous example of Pascal's wager. With this argument, found in his *Pensées*, Pascal purported to show that it is rational to wager for God's existence. In modern terminology, we have to consider all combinations of the existence or non-existence of God, on the one hand, and an agent's belief or disbelief in God, on the other hand. This leads to four cases each with its own expected utility. In the case that God exists, it is assumed that there are everlasting heavenly rewards for those who believe (positive infinite expected utility) and everlasting infernal punishments for those who disbelieve (negative infinite expected utility). In the case that God does not exist, there are a lifetime of earthly burdens for those who believe (negative finite expected utility) and a lifetime of earthly pleasures for those who disbelieve (positive finite expected utility). If the agent is maximally uncertain about the existence of God (assigning 50% probability to the possibility of existence and 50% probability to the possibility of non-existence), the expected utility of believing is infinitely better than that of disbelieving. So, according to this argument, if one has to wager, it is better to wager for God's existence.

In the context of a discussion of Pascal's wager, Oppy (1990, p. 163) considers the epistemic possibility "that the probability that God exists is infinitesimal," in which case "the calculation of the expected return of a bet on [the existence of] God is no longer as straightforward as the initial argument suggested."

Following up on this suggestion, Hájek (2003) considers whether salvation has an infinite utility. He mentions two formal approaches that allow us to tell apart various infinite expectation values that occur in Pascal's wager and related problems. Hájek mentions NSA as one possibility of dealing with infinitesimal probabilities and infinite utilities, but he favours Conway's numbers, citing their ingenuity and user-friendliness. He speculates that such a formal approach can illuminate a whole range of problems involving infinitesimal probabilities (such as the two envelope paradox).

On p. 38, Hájek writes that "the infinitesimal probability can 'cancel' the infinite utility so as to yield a finite expectation for wagering for God." The idea of cancelling is indeed what NSA allows us to formalize: each infinitesimal is the reciprocal of an infinite number and vice versa. Multiplying an infinite hyperreal number and its multiplicative inverse, a particular infinitesimal, yields unity. So, on the one hand, we may obtain finite (non-infinite) and non-infinitesimal values by multiplying infinite and infinitesimal numbers. On the other hand, there are also combinations of infinite and infinitesimal numbers whose product is an infinitesimal or an infinite number. More details can be found in Wenmackers (2018). For a treatment with surreal probabilities and utilities, see Chen and Rubio (2018): their approach also allows them to treat the St. Petersburg paradox.

13 THE LOCKEAN THESIS AND RELATIVE INFINITESIMALS

Whereas standard probability measures may seem too coarse-grained for some applications, where we would like to distinguish between possible and impossible events, they may not seem coarse-grained enough for other applications, as we will see in this section.

Suppose that you have detailed knowledge of the probabilities in a given situation. It has been argued that it may still be beneficial to hold some full (dis-)beliefs (Foley, 2009). But when is it rational to believe something in this case? The Lockean thesis suggests that it is rational to believe a statement if the probability of that statement is sufficiently close to unity.[42] This is usually modelled by means of a probability threshold. As is demonstrated by the Lottery Paradox (Kyburg, 1961), the threshold-based model is incompatible with the Conjunction Principle. Moreover, it can be objected that the actual probabilities are too vague to put a sharp threshold on them, and that a threshold should be context-dependent.

Based on certain analogies between large and infinite lotteries, Wenmackers (2012) suggests the use of NSA to introduce a form of vagueness or coarse-graining and context-dependence in the formal model of the Lockean thesis.[43] Hrbáček (2007) develops relative or stratified analysis, an alterative approach to NSA that contains "levels" as a formalization of the intuitive scales-of-magnitude concept. Applying Hrbáček's framework, Wenmackers (2013) introduces "Stratified Belief" as an alternative formalization of the Lockean Thesis.[44]

The basic idea is to interpret the Lockean thesis as follows: it is rational to believe a statement if the probability of that statement is indistinguishable from unity (in a given context). The context-dependent indistinguishability relation is then modelled using the notion of differences up to a level-dependent, ultrasmall number. These ultrasmall numbers, also called "relative infinitesimals," are ordinary real numbers, which are merely unobservable, or do not have a unique name, in a given context. The aggregation rule for this model is the "Stratified conjunction principle," which entails that the conjunction of a standard number of rational beliefs is rational, whereas the conjunction of an ultralarge number of rational beliefs is not necessarily rational.[45]

42 This is reminiscent of the concept of "moral certainty"; see also footnote 78.

43 An earlier version can be found in Wenmackers (2011, Ch. 4).

44 An earlier version can be found in Wenmackers (2011, Ch. 3).

45 Although this model is intended to describe beliefs that are almost certain, it can be used for weaker forms of belief by substituting a lower number instead of unity.

14 RECENT OBJECTIONS AND OPEN QUESTIONS

In this section, we give a brief overview of developments from the two last decades in which new objections against and defences for infinitesimal probabilities have been added to the literature. It may be too early to evaluate the most recent collection of attempted refutations and acclaims for infinitesimal probabilities. Still, we briefly mention some here. More discussion can be found in Benci et al. (2018).

14.1 *Symmetry Constraints and Label Invariance*

In a number of publications, Paul Bartha applies ideas from non-standard measure theory to problems in the philosophy of probability. Bartha and Hitchcock (1999) use NSA in the usual way, *i.e.*, in order to obtain a real-valued probability function. Bartha and Johns (2001) also consider the application of NSA to a probabilistic setting, but they favour a simpler appeal to symmetry in order to obtain the conditional probabilities relevant to their problem. (Later, Bartha, 2004, discusses de Finetti's lottery and uses infinitesimal probabilities as one way to escape the conclusion that CA is mandatory, since they exhibit hyperfinite additivity instead.)

Considering the case of an ω-sequence of coin tosses, Williamson (2007) demonstrates the incompatibility between infinitesimal probabilities and requiring the equiprobability of what he calls "isomorphic events," which are "events of exactly the same qualitative type" (p. 175). In particular, for ω-sequences of coin tosses, he argues that the probability assigned to the event should not depend on when exactly the tossing started. Williamson contrasts his finding with that of Elga: whereas Elga (2004) finds regularity to lead to too many eligible non-standard distributions, Williamson finds regularity in combination with what he calls "non-arbitrary constraints" to rule out all candidate distributions.

Weintraub (2008) attempts to demonstrate that Williamson's argument depends on the assumption of label-independence, which is itself incompatible with infinitesimal probabilities. More recently, Benci et al. (2018) analyze Williamson's argument in the light of NAP theory. They, too, conclude that isomorphic events cannot be assigned equal hyperreal-valued probabilities without contradicting the assumptions on which this theory relies. Simultaneously, Howson (2017) argues—without using any details of NAP theory—that "it is not regularity which fails in the non-standard setting but a fundamental property of shifts in Bernoulli processes." Parker (2018) argues that these objections to the argument of Williamson (2007) fail.

14.2 *Non-uniqueness of Hyperreal Probabilities*

Elga (2004) considers the zero-fit problem of the "best system" analysis of laws: if all systems of laws assign probability zero to the actual history up to now, then one cannot identify the best system based on a measure of goodness-of-fit. He entertains the option of applying non-standard probability functions and thus to assign a non-zero infinitesimal probability to the actual history, thereby escaping a zero fit. Ultimately, however, he rejects this proposal:

> We have required our nonstandard probability function to be regular, and to approximate given standard probability functions. But those requirements only very weakly constrain the probabilities those functions assign to any individual outcome. [...] And the fit of a system associated with such a function is just the chance it assigns to actual history. So the fit of such a system indicates nothing about how well its chances accord with actual history.

The relevant construction of a non-standard probability function is given in an appendix, where Elga phrases the conclusion as follows: "[T]he probabilities that these approximating functions ascribe to actual history span the entire range of infinitesimals [...]. So by picking an appropriate approximating function, we can get any such system to have any (infinitesimal) fit we'd like." In other words, Elga concludes that there are too many ways of assigning different infinitesimal probabilities to the same history and that there is no principled way to prefer one over the others.

Herzberg (2007) contrasts Elga's viewpoint, in which all hyperreal-valued functions that differ from a particular real-valued function by at most an infinitesimal (where the latter is defined) are to be treated on a par, with the praxis of NSA. As Herzberg points out, applications of NSA typically involve the construction of a *particular* non-standard object, usually some hyperfinite combinatorial object, leading to a particular internal probability measure. In order to appreciate how Herzberg's viewpoint differs from Elga's, it is helpful to consider an example.[46] Anderson (1976) presents an internal representation of Brownian motion, which makes it possible to treat Brownian motion in terms of (infinite) combinatorics.[47] In order to be scientifically relevant, however, such an alternative description has to fulfil two criteria: (1) it has to approximate the standard probability function associated with the process (in this case, the Wiener measure)[48]

46 I am grateful to Frederik Herzberg for this suggestion.

47 See also Albeverio, Fenstad, Hoegh-Krøhn, and Lindstrøm (1986, section 3.3).

48 Since internal probability functions differ from standard ones both in terms of domain and of range, this approximation can be thought of as a two-step procedure, the second of which involves the standard part function.

and (2) it has to promote further research (as is indeed the case for Anderson's work; consider, for instance, Perkins', 1981, work on Brownian local time). Although many non-standard measures fulfil the first condition, the vast majority of them do not fulfil the second one.

Many worries and some open questions about infinitesimal probabilities arise due to the non-uniqueness and associated arbitrariness of hyperreal-valued probability measures (also discussed, *e.g.*, by Hofweber, 2014).[49] When comparing the situation to that of real-valued probability functions that are CA, there is a trade-off between definiteness of the domain and definiteness of the range. In the case of an infinite sample space, CA functions have many non-measurable events in the powerset of that sample space. Which subsets of the sample space are measurable and which are not is to a certain extent arbitrary. If we settle for FA, we can extend the real-valued function to the entire powerset (by considering Banach limits; see for instance Schurz & Leitgeb, 2008), but then we introduce a lot of arbitrariness. Again in the case of an infinite sample space, NAP functions allow for the same kind of variation in their standard part as the FA functions do, and more given that also the infinitesimal part may vary (see for instance Kremer, 2014). Given that it reappears in slightly different guises across different frameworks, we cannot set aside this arbitrariness as a flaw of one particular theory. Rather, it reminds us that the powerset of an infinite sample space contains a lot of uncharted territory.[50]

At least some of the worries related to arbitrariness are alleviated if we take into account the distinction between the ontology of infinitesimal probabilities and the deductive procedures they encourage: very similar modes of reasoning can be applied in related frameworks that suggest a different ontology (recall section 11).[51]

More generally, various authors argue that hyperreal numbers are not quite right for the task at hand (*e.g.*, that the infinitesimals are too small; Easwaran, 2014; Pruss, 2014). Easwaran (2014, pp. 34–35) argues that "the structure of the hyperreals goes beyond the physical structure of credences" and that they "can't provide a faithful model of credences of the sort wanted by defenders of Regularity." On the other hand, Hofweber

49 As mentioned in section 5, free ultrafilters are intangible objects. As a result, non-standard probability functions that rely on these filters are intangibles, too.

50 In particular, even if the sample space is just countably infinite, its powerset (which contains the events to which we want to assign probabilities) is uncountably large. Among the uncountably many sets that are neither finite nor co-finite, there is a wild variety (for instance, in terms of Turing degrees or other complexity measures) and it is here that we should take heed of Feferman's reservations about considering the totality of all arbitrary subsets of $\mathbb{N}$, $\mathcal{P}(\mathbb{N})$, as a well-defined notion; see, *e.g.*, Feferman (1979, p. 166) and Feferman (1999). I am grateful to Paolo Mancosu for suggesting this connection.

51 Following a distinction introduced by Benacerraf (1965), a similar remark has been made by Katz (2014, section 2.3) regarding interpretations of the work of Euler (and also that of Leibniz) in the context of standard or non-standard analysis.

(2014) tries to defend infinitesimal chances and outlines some additional principles (non-locality, flexibility, and arbitrary additivity) that are required for a theory to capture our concept of chance. Also Benci et al. (2018) are optimistic that NAP theory can be defended against many of the previously raised objections.

14.3 *Cardinality Considerations*

Hájek (2012b) argues that regularity is an untenable constraint on credences, even if we allow probability functions to take hyperreal values. He invites us to "imagine a spinner whose possible landing points are randomly selected from the $[0,1)$ interval of the hyperreals," concluding that regularity fails if we apply the same interval of hyperreals as the range of a function that assigns probabilities to events associated with this hyperreal spinner. He envisages

> a kind of arms race: we scotched regularity for real-valued probability functions by canvassing sufficiently large domains: making them uncountable. The friends of regularity fought back, enriching their ranges: making them hyperreal-valued. I counter with a still larger domain: making its values hyperreal-valued

and so on. Following up on Hájek's informal suggestion of an arms race, Alexander Pruss (2013) proves that for each set of probability values, possibly including hyperreal values, there exists a domain on which regularity fails.

However, as NAP theory illustrates, the defender of regularity need not participate in this race at all and Hájek considers this option, too: "Perhaps we could tailor the range of the probability function to the domain, for each particular application?" However, he worries "that in a Kolmogorov-style axiomatization the commitment to the range of P comes *first*." He continues by saying that "[i]t is not enough to say something unspecific, like 'some non-Archimedean closed ordered field…' Among other things, we need to know what the additivity axiom is supposed to be." Of course, NAP theory does exactly this: by requiring ultra-additivity, for any sample space a range can be constructed that ensures regularity. However, one cannot switch the quantifiers: in agreement with Pruss (2013), there is no universal range that can ensure regularity for all sample spaces.[52]

52 Hájek (2012b) also states that "[i]f we don't know exactly what the range is, we don't know what its notion of additivity will look like." Maybe prolonged exposure to real-valued measures, in which ultra-additivity is clearly unattainable, makes us overlook this very natural notion of additivity that does not depend on any further parameters?

14.4 *Non-conglomerability*

Before we can address this worry, we first have to introduce the notion of conglomerability.

We will call a (hyper-)real-valued probability function P finitely, countably, or uncountably conglomerable if and only if for any finite, countable, or uncountable (resp.) partition $\{A_1, A_2, \ldots\}$ of the sample space (whose members are measurable according to P) and for any event A that is measurable according to P, the following conditional statement holds. If a and b are (hyper-)real numbers such that $\forall A_n \in \{A_1, A_2, \ldots\}, a \leq P(A|A_n) \leq b$, then $a \leq P(A) \leq b$.

In standard probability theory, both finite and countable conglomerability are guaranteed to hold. The proof of this relies crucially on the axiom of normalization and on the axiom of finite or countable additivity (resp.). Even in the standard approach, uncountable conglomerability does not hold in general.

Theories that lack normalization or countable additivity, are not guaranteed to be countably conglomerable. In particular, both de Finetti's proposal for FA probability theory and NAP theory are finitely but not countably conglomerable.[53]

Pruss (2012, 2014) raises this as an objection to theories that allow infinitesimal probabilities. In recent work, DiBella (2018, p. 1200) shows that the failure of countable conglomerability already arises in qualitative probability theories that are non-Archimedean and that this carries over to any quantitative theory that is non-Archimedean (of which NAP theory is an example). Since it is such a general feature of the underlying probability ordering, he suggests that non-conglomerability is not suitable as a criticism of non-Archimedean theories.

15 EPILOGUE: ON THE VALUE OF METHODOLOGICAL PLURALISM

I would like to end this chapter with some remarks that may apply to formal epistemology (and related endeavours) more generally. Only by comparing different methodologies may one obtain some indication of their strengths and limitations and how they distort the results.

We tend not to notice what is always present. An atmosphere was present before our ancestors developed eyes and to this day the air between us remains invisible to us. By experimenting with other gas mixtures, we learn, not only about those new substances, but also about the air that

53 The failure of countable conglomerability can be seen by considering a uniform distribution over the sample space $\mathbb{N} \times \mathbb{N}$ and two countable partitions: $A_i = \{(i, n) | n \in \mathbb{N}\}$ and $B_i = \{(n, i) | n \in \mathbb{N}\}$. For the demonstration in the case of FA probability, see de Finetti (1972, Ch. 5).

surrounds us. We become aware of its weight, its oxygen content, and its capacity to carry our voice. And although we keep living in air for most of the time, for particular purposes, we may prefer other mixtures over air (*e.g.*, increasing the oxygen content to help someone breathe or decreasing the oxygen content to avoid oxidation).

Like air in our biosphere, the real numbers are equally pervasive in our current mathematical practice. It appears to me that we are subjected to methodological adaptation to an extent no less than we are to sensory adaptation. The study of infinitesimal probabilities involves a departure from the standard formalism of real-valued probability functions. By changing our methodological environment, we may start to notice certain assumptions in the usual approach. Dealing with a familiar problem in an unfamiliar way thus presents a unique opportunity: it allows us to distinguish elements that are essential to its solution from aspects that are merely artifacts due to the method that has been applied.

Investigating in a formal way a rich concept such as probability cannot be carried out within the bounds of any single formalization, but challenges us to combine perspectives from an equally rich selection of frameworks. In particular, I believe that methods involving hyperreal probability values, while detracting nothing from the merits of the monometric standard approach, have much to add to this polymetric selection.

16 APPENDIX: HISTORICAL SOURCES CONCERNING INFINITESIMAL PROBABILITIES (1870–1989)

This part does not contain an overarching story arc, but it can be used as an annotated bibliography or to look up specific details.

Despite its length, this appendix does not pretend to be exhaustive; some developments—especially the early ones—are merely sketched. The subdivision into decades is indicative rather than strict. Usually, the publication date is taken as the decisive factor for the chronology, except for Carnap's work from 1960: this work was only published in 1980, but it is included in an earlier section, for thematic reasons.

16.1 *Before 1960: pre-Robinsonian era*

The 1870s: The Real Numbers and the Standard Limit

The modern approach to standard analysis was developed by "the great triumvirate" (Boyer, 1949, p. 298): Georg Cantor, Richard Dedekind, and Karl Weierstrass. First, Cantor gave a construction of the real numbers via Cauchy sequences (recall section 8.5). Then, Dedekind gave an alternative construction of the real numbers via Dedekind cuts (which we will not

discuss). Weierstrass introduced the modern epsilon-delta definition of the limit (which builds on earlier work by Bernard Bolzano in the 1810s and by Augustin-Louis Cauchy in the 1820s).

As an example, we consider the derivative as a limit of the quotient of differences and express this limit in terms of an epsilon-delta definition:

$$\begin{aligned}\frac{dy}{dx} &= \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{y(x+\Delta x) - y(x)}{\Delta x},\end{aligned}$$

where

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = L$$

if and only if

$$\forall \epsilon > 0 \in \mathbb{R},\ \exists \delta > 0 \in \mathbb{R} : \forall \Delta x \in \mathbb{R} \left(0 < |\Delta x| < \delta \Rightarrow |\frac{\Delta y}{\Delta x} - L| < \epsilon \right).$$

The 1880s: The Archimedean Axiom

In the introduction, we encountered the criterion to decide whether a number system is Archimedean or non-Archimedean (condition 1). In particular, hyperreal fields are non-Archimedean and those can be employed to represent infinitesimal probabilities. Here, we investigate the origins of this sense of the word 'Archimedean'.

Around 225 BC, Archimedes of Syracuse published two volumes known in English as "On the Sphere and Cylinder". At the beginning of the first book, Archimedes stated five assumptions. The fifth assumption is that,[54] starting from any quantity, one may exceed any larger quantity by adding the former quantity to itself sufficiently many times.[55] In a paper on ancient Greek geometry, Otto Stolz (1883) discussed this postulate, which he calls "*das Axiom des Archimedes*" for ease of reference. Although Stolz was well aware that Archimedes himself attributed an application of this axiom to earlier geometers, apparently he did not notice that the axiom also appeared in Euclid's Elements (Bair et al., 2013, p. 888). In his textbook on arithmetic, which was very influential according to Ehrlich (2006, p. 5), Stolz (1885) presented examples of *Grössensysteme* (systems of

54 Heath (1897, p. 4) translates the assumption as follows: "Further, of unequal lines, unequal surfaces, and unequal solids, the greater exceeds the less by such a magnitude as, when added to itself, can be made to exceed any assigned magnitude among those which are comparable with [it and with] one another."

55 This formulation suggests a strong relation between Archimedean quantities and addition. Additivity also plays an important role in intuitions concerning infinitesimal quantities, including infinitesimal probabilities, even though these are non-Archimedean probabilities: recall the discussion on ultra-additivity (section 8.3 and appendix section 16.3).

magnitudes) that fail to satisfy this Archimedean axiom, whereas systems that are continuous in the sense of Dedekind do satisfy it.

The 1890s: Infinitesimal Probabilities in a Geometric Context

In 1891, Giulio Vivanti and Rodolfo Bettazzi discussed infinitesimal line segments in the context of probability (see Ehrlich, 2006). In these early discussions, infinitesimal probabilities are considered in the context of a geometric interpretation of probability. As such, this provides an interesting contrast to the more recent literature, in which infinitesimal probabilities are usually introduced in the context of subjective interpretations of probability (related to a criterion of open-mindedness).

Later on, in the 1910s, Federigo Enriques discussed the (impossibility of) infinitesimal probabilities on two occasions, again in a geometric context.[56]

The 1900s: Measurability and Non-measurability

Building on émile Borel's countably additive measure from the 1890s, Henri Lebesgue introduced his translation invariant and countably additive measure in 1902. In 1905, Giuseppe Vitali gave the first example of a non-Lebesgue measurable set. See for instance Skyrms (1983b) for some discussion.[57]

The 1930s: Kolmogorov, Skolem, and de Finetti

KOLMOGOROV'S PROBABILITY MEASURES Andrey Kolmogorov (1933) introduced probability as a one-place function with as the domain a field of sets over a given sample space and as the range the unit interval of the real numbers. In the first chapter of his book, he laid out an elementary theory of probability "in which we have to deal with only a finite number of events." The axioms for the elementary case stipulate non-negativity, normalization, and the addition theorem (now called "finite additivity," FA). In the second chapter, dealing with the case of "an infinite number of random events," Kolmogorov introduced an additional axiom: the Axiom of Continuity. Together with the axioms and theorems for the finite case

56 Thanks to Philip Ehrlich for this addition. He is planning an article on the work of Enriques; meanwhile, Ehrlich (2006) contains the relevant references.

57 Skyrms (1983b) argues that the Peano-Jordan measure (which preceded the Borel measure) only employs ideas that were available in Plato's time, whereas Borel measure crucially relies on distinctions among infinite cardinalities only introduced by Cantor. Peano-Jordan measure is finitely additive, which follows from its definition, and it lacks the stronger property of countable additivity (CA). Borel measure is CA, but this has to be specified in the definition by hand. Skyrms observes that this approach was contested, for instance by Schoenflies in 1900, who objected that the matter of extending additivity into the infinite cannot be settled by positing it. Lebesgue measure is CA, too, and it is translation invariant, which is appealing to our intuitions.

(in particular, FA), this leads to the generalized addition theorem, called "σ-additivity" or "countable additivity" (CA) in the case where the event space is a Borel field (or σ-algebra, in modern terminology). We reviewed his axiomatization in section 7.

SKOLEM'S NON-STANDARD MODELS OF PEANO ARITHMETIC The second-order axioms for arithmetic are categoric: all models are isomorphic to the intended model $\langle \mathbb{N}, 0, +1 \rangle$, a triple consisting of the domain of discourse (infinite set of natural numbers), a constant element (zero), and the successor function (unary addition). Dedekind (1888) was the first to prove this. His "rules" for arithmetic were turned into axioms by Giuseppe Peano (1889), giving rise to what we now call "Peano Arithmetic" (PA).

The first-order axioms for arithmetic are non-categoric: there exist non-standard models $\langle {}^*\mathbb{N}, {}^*0, {}^*+1 \rangle$ that are not isomorphic to $\langle \mathbb{N}, 0, +1 \rangle$. Thoralf Skolem (1934) was the first who proved this.[58] With the Löwenheim-Skolem theorem, it can be proven that there exist models of any cardinality. ${}^*\mathbb{N}$ contains finite numbers as well as infinite numbers. We now call ${}^*\mathbb{N}$ a set of hypernatural numbers.[59]

DE FINETTI ON NON-ARCHIMEDEAN PROBABILITY RANKINGS In 1931, Bruno de Finetti addressed the relation between qualitative and quantitative probability. Qualitative probability deals with ordering or ranking events by a partial order relation, $\succeq$, interpreted as "at least as likely as." Quantitative probability deals with probability functions that assign numerical values—usually real numbers—to events.

On pp. 313–314, de Finetti (1931, section 13) presented four postulates for the probability ordering.[60] In particular, the second postulate states that every event that is merely possible (rather than impossible or certain) is strictly more likely than the impossible event and strictly less likely than

58 See Stillwell (1977, section 3) and Kanovei, Katz, and Mormann (2013, section 3.2) for some comments on the direct construction given by Skolem (1934). In contrast to Skolem's result, the proof given in modern presentations usually relies on the Compactness property of first-order logic. First, consider a first-order language for arithmetic, $\mathcal{L}_{PA}$, which has a name for each natural number. Call PA the set of sentences in $\mathcal{L}_{PA}$ that are true about arithmetic. Then, add a new constant, c, to the language and consider PA', which is the union of the PA and $\{c > 0, c, > 1, c > 2, \ldots\}$. Since each finite subset of PA' has a model (in which c is a natural number that is larger than any of the other natural numbers that are named in the the finite subset), it follows from the Compactness of first-order logic that PA' has a model (which contains a copy of the natural numbers and in which c is an infinite hypernatural number).

59 For a discussion of the order-type of countable non-standard models of arithmetic, see *e.g.* Boolos, Burgess, and Jeffrey (2007, Ch. 25, p. 302–318) and McGee (2002). More advanced topics can be found in the book by Kossak and Schmerl (2006).

60 Thanks to Paul Pedersen for some pointers to de Finetti's early work on non-Archimedean probability rankings.

the certain event. He considers the question whether such a ranking is compatible with the usual way of measuring probabilities by real numbers. De Finetti observed that such a probability ranking has a non-Archimedean structure, whereas real-valued probability functions are Archimedean. Related to this point, de Finetti (1931, p. 316) wrote:

> However, it is anyway possible to satisfactorily measure probabilities by numbers, that is by making such a structure Archimedean by neglecting the infinitely small probabilities [...]

Since this was written well before the development of NSA, we should be careful not to interpret "infinitely small probabilities" as the values of a hyperreal-valued probability function, which can subsequently be truncated by the standard part function. On the other hand, de Finetti was not merely referring to infinitesimal probabilities in an informal sense, either. In the continuation of the sentence quoted above, he stated, concerning infinitely small probabilities:

> [...] that, when multiplied [...] by a number n, however large, they never tend to certainty, that is in other words, they are always less than the probability $1/n$ of one among n incompatible, identically probable events forming a complete class.

As a result, the partial order on the probability of events (which is just the order relation on the real numbers, $\geq$) does not coincide with the partial order on events ($\succeq$): taking A and B to be events, $P(A) \geq P(B)$ implies $A \succeq B$, but not vice versa, and $A \succeq B$ together with $B \succeq A$ implies $P(A) = P(B)$, but not vice versa. (Counterexamples to the inverse implications can be obtained by considering A to be the impossible event, $\varnothing$, and B a possible event with $P(B) = 0$.) The non-Archimedean partial ordering of events can be said to be more fine-grained than the Archimedean partial ordering of probabilities of those events, since the former leads to more equivalence classes (sets of events $\{B \mid B \succeq A \wedge A \succeq B\}$ for some event A) than the latter (with equivalence classes of events of the form $\{B \mid P(A) = P(B)\}$ for some event A).

In 1936, de Finetti reflected on the meaning of possible events (*i.e.*, events represented by non-empty sets) that have probability zero. He agrees with Borel and Lévy[61] that these are merely theoretical constructs: they do not represent events that are practically observable, but are merely defined as limiting cases thereof. They would require information from infinitely many experiments or an experiment involving an absolutely exact measurement, both of which exceed what is practically achievable.[62] In this

61 See also footnote 78 for the relation to Cournot's principle.

62 This is the relevant quote in French (de Finetti, 1936, p. 577): "*Il n'y a pas de doute, ainsi que l'a remarqué M. Borel, et comme cela se trouve très clairement expliqué dans le traité de M. Lévy,*

context, and unlike the 1931 article, de Finetti did consider the option of infinitesimal probability values and even an infinite hierarchy thereof ("*chacune infiniment petite par rapport á la précédente*", p. 583). Ultimately, however, he advocated sticking to real numbers as probabilities and dropping the assumption of countable additivity (p. 584), which is a position he stood by throughout all of his later work (see section 16.3).

The 1950s: From Weak to Strict Coherence

In the context of Bayesianism and decision theory, infinitesimal probabilities have been discussed in relation to "strict coherence"[63] and "regularity." This discussion started in the 1950s, with the Ph.D. dissertation of Abner Shimony followed by the publication of Shimony (1955).

Earlier, both Frank P. Ramsey (1931) and de Finetti (1937) had combined a subjective interpretation of probability with an important rationality constraint, imposed on the set of an agent's degrees of belief: in order to be considered rational, a person's set of beliefs must meet the condition of "coherence." This condition can be regarded as a probabilistic extension of the consistency condition from classical logic. In particular, an agent's degrees of belief are coherent just in case no Dutch book can be made against the agent: no finite combination of bets, of which the prizes are set in accordance with the agent's degrees of belief, should lead to a sure loss. De Finetti (1937) showed that an agent's degrees of belief are coherent (and thus that no Dutch Book can be made against him) just in case his degrees of belief are such that they respect the axioms for finitely additive probability functions.

SHIMONY'S STRICT COHERENCE Shimony (1955) strengthened the earlier notion of coherence (now called "weak coherence") to that of coherence "in the strong sense" (now "strict coherence"): no finite combination of bets, of which the prizes are set according to the agent's degrees of belief, should lead to a sure loss (as before) or a possible net loss without the possibility of a net profit (stronger condition). To obtain strong coherence, Shimony had to strengthen one of the probability axioms accordingly. The original axiom says that the degree of confirmation (or conditional credence) of some hypothesis h given a piece of evidence e is 1 if e entails

que la notion d'événement possible et de probabilite nulle est purement théorique, car il s'agit en géneral d'événements définis comme des cas limites d'événements pratiquement observables, et leur vérification exigerait par conséquent une infinité *d'expériences ou une expérience comportant une mensuration absolument* exacte."

63 In the early literature, there circulated other names for this criterion as well: 'strict fairness' (Kemeny) and [strong] 'rationality' (Lehman, Adams). See Carnap (1971a, p. 114) for a helpful overview of the terminology in the early literature.

h, whereas the stronger version reads: the degree of confirmation of h given e is 1 if and only if e entails h.

Initially, Shimony (1955) only defined (strict) coherence for finite sets of beliefs, but in a later section he did discuss "[t]he difficulty of extending the notion of coherence so as to apply to infinite sets" (p. 11). In this context, he wrote (p. 20):

> An appropriate betting quotient would be an 'infinitesimal', which is neither 0 nor finite; but this is impossible because of the Archimedean property of the positive real numbers.

Shimony also remarked that strong coherence on infinite sets of belief cannot be used to justify CA (which he calls "the Principle of Complete Additivity" on p. 18).

STRICT COHERENCE WITHOUT INFINITESIMALS The work on strict coherence initiated by Shimony was soon picked up by others. Some of the ensuing publications were related to the notion of "regularity." In the context of finite sample spaces, Rudolf Carnap (1950, Ch. 5) had introduced regularity as the condition that a function should assign positive values to state descriptions that sum to unity. In particular, he applied this condition to credence functions (probability functions in the sense of rational degrees of belief) associated with a finite set of state descriptions (finite sample space).[64]

Combining the earlier result of Shimony (1955) on the one hand and that of John G. Kemeny (1981) and R. Sherman Lehman (1955) on the other hand, we have that a probability function on a finite sample space is strictly coherent if and only if it is "regular" (*cf*. Carnap, 1971b, p. 15).

Ernest W. Adams (1959, 1962–63, 1964) was interested in the case of infinite sample spaces: he focused on the issue of additivity. Walter Oberschelp (1962–63) wrote on a similar topic in German: he looked for a similar, but weaker constraint for the infinite case than Adams'.

So, none of these authors did follow up on Shimony's remark regarding infinitesimal probabilities. An important exception was Carnap: in 1960, he explicitly considered the option of non-real-valued degrees of belief that admit infinitesimal values. (Although this work was published posthumously, in 1980, we do discuss it already at this point.)

64 For infinite sample spaces, Carnap (1950) considers limits of unconditional and conditional probability functions; although those limit functions may assign zero to state descriptions, Carnap calls them "regular," too. This usage should be contrasted with that in contemporary writings on infinitesimal credences, where regularity is (equivalent to) the condition that a probability function should assign strictly positive values to singleton events, even for infinite sample spaces.

CARNAP'S QUEST FOR NON-ARCHIMEDEAN CREDENCES Inspired by Shimony's work on strict coherence, Carnap (1980) considered a language with real-valued functions, $\mathcal{L}$, and a credence function with non-Archimedean range, $\mathcal{C}$. He wrote (p. 146):

> we could regard these axioms as axioms of regularity for $\mathcal{L}$; and we would call $\mathcal{C}$ regular iff it fulfilled all these axioms. However, to carry out this program would be a task beset with great difficulties.

The first problem he considered is that of finding axioms for the binary relations *IS* (to be read as: 'is Infinitely Small compared to') and *SEq* (to be read as: "is Smaller or Equal in size to"), both defined on the class of all subsets of the set of real numbers.[65] Further on, Carnap considered the problem of constructing a measure function π that is defined on all subsets of the set of real numbers. He stated (p. 154, italics in the original): "The values of π are not real numbers but numbers of a *non-Archimedean number system* Ω to be constructed."

16.2 *The 1960s: Robinson's NSA and Bernstein & Wattenberg's Non-standard Probability*

The development of non-standard analysis by Abraham Robinson in the 1960s allowed for a formal and consistent treatment of infinitesimal numbers. Soon enough, this work was applied to measure theory in general and to probability theory in particular. Beyond this point, some technical notions from NSA appear: please consult sections 4 and 5 for the meaning of unfamiliar terms.

Non-standard Models of Real Closed Fields and Robinson's NSA

Robinson (1961, 1966) founded the field of NSA: he combined some earlier results from mathematical logic[66] in order to develop an alternative framework for differential and integral calculus based on infinitesimals and infinitely large numbers.

Robinson's hyperreal numbers are a special case of a real closed field (RCF). In general, a RCF is any field that has the same first-order properties as $\mathbb{R}$. The second-order axioms for the ordered field of real numbers are

65 Upon publication of these notes, Hoover (1980) remarked that one of the axioms Carnap had proposed for *SEq* was in contradiction with the others (axiom 3f on p. 147 amounted to countable additivity, which is incompatible with a non-Archimedean range); also one of the proposed axioms for *IS* was in contradiction with the others (axiom 7p on p. 148).

66 See Robinson (1966, p. 48) for some references. In particular, Hewitt (1948) had constructed hyperreal fields using an ultrapower construction and Łoś (1955) had proven a transfer theorem for these fields.

categoric: all models are isomorphic to the intended model $\langle \mathbb{R}, +, \times, \leq \rangle$, a quadruple consisting of the set of real numbers, the binary operations of addition and multiplication, and the order relation. Skolem's existence proof of non-standard models of arithmetic (section 16.1) can be applied to RCFs, too.[67] The axioms for RCFs (always in first-order logic) are non-categoric: there exist non-standard models $\langle {}^*\mathbb{R}, {}^*+, {}^*\times, {}^*\leq \rangle$ that are not isomorphic to $\langle \mathbb{R}, +, \times, \leq \rangle$.

Applying the Löwenheim-Skolem theorem, it can be proven that there exist models of any cardinality; in particular, there are countable models (*cf.* the "paradox" of Skolem, 1923). In the context of hyperreal numbers, however, only uncountable models are considered. First of all, in this context the uncountable set of real numbers is assumed to be embedded in the non-standard model. Moreover, in the context of NSA also functions are transferred, which requires uncountably many symbols, thereby blocking the construction of a countable model.

The standard real numbers are Archimedean, *i.e.*, they contain no non-zero infinitesimals in the sense of condition (1):

$$\forall a \in \mathbb{R} \setminus \{0\},\ \exists n \in \mathbb{N} :\ \frac{1}{n} < |a|.$$

In particular, $\langle \mathbb{R}, +, \times, \leq \rangle$ is the only complete Archimedean field.[68] In contrast, non-standard models do not have such a property: $\langle {}^*\mathbb{R}, {}^*+, {}^*\times, {}^* \leq \rangle$ is a non-Archimedean ordered field and it is not complete. Saying that ${}^*\mathbb{R}$ is non-Archimedean means that it does contain non-zero infinitesimals in the sense of condition (1):

$$\exists a \in {}^*\mathbb{R} \setminus \{0\},\ \forall n \in \mathbb{N} :\ \frac{1}{n} \geq |a|.$$

In other words: ${}^*\mathbb{R}$ contains infinitesimals. As a consequence, for any such a hyperreal infinitesimal a it holds that

$$\forall n \in \mathbb{N} :\ \sum_{i=1}^{n} |a| < 1.$$

${}^*\mathbb{R}$ contains finite, infinite and infinitesimal numbers; we call ${}^*\mathbb{R}$ a set of hyperreal numbers.

67 Applying the idea of footnote 58 to RCF instead of PA, c will represent an infinite hyperreal number and its multiplicative inverse will represent an infinitesimal number.

68 Here, 'complete' can refer both to Cauchy or limit completeness (meaning that each Cauchy sequence of real numbers is guaranteed to converge in the real numbers) and to Dedekind or order completeness (meaning that each non-empty set of real number that has an upper bound is guaranteed to have a least upper bound), because Cauchy completeness together with the Archimedean property implies Dedekind completeness.

Bernstein & Wattenberg's Non-standard Probability Function

The infinitesimal numbers contained in the unit interval of a non-standard model of a RCF can be used to represent infinitesimal probabilities. Allen R. Bernstein and Frank Wattenberg (1969) were the first to apply Robinson's NSA in a probabilistic setting and thus to describe infinitesimal probabilities in a mathematically rigorous framework. On p. 171, they stated the following goal: "Suppose that a dart is thrown, using the unit interval as a target; then what is the probability of hitting a point?" They followed up this question with an informal answer:

> Clearly this probability cannot be a positive real number, yet to say that it is zero violates the intuitive feeling that, after all, there is some chance of hitting the point.

In their paper, Bernstein and Wattenberg formalized this intuitive answer using positive infinitesimals from Robinson's NSA.[69] Their measure is based on a hyperfinite counting measure of a hyperfinite subset of the hyperextension of the sample space.[70] The non-standard result for any Lebesgue-measurable set is infinitely close to its Lebesgue measure:[71] "In particular, nonempty sets of Lebesgue measure zero will have positive infinitesimal measure." They stated that:

> Thus, for example, it is now possible to say that 'the probability of hitting a rational number in the interval $[0, \frac{1}{4})$ is exactly half that of hitting a rational number in the interval $[0, \frac{1}{2})$,' despite the fact that both sets in question have Lebesgue measure zero.

Of course, the former probability being half that of the latter also applies if both probabilities are zero, rather than infinitesimals.[72] This observation is only relevant if an additional assumption is made, for instance that the probabilities are non-zero or that the former should be smaller than the latter.

69 Observe that, in order to assign non-zero infinitesimals to point events, they have to depart from the usual application of NSA. Moreover, the function that they obtain is an external object, which means (roughly) that it does not have a counterpart within standard analysis (*cf.* section 4). On the other hand, it is possible to take the standard part of the function's output, which yields the unique real value that is closest to the hyperreal value.

70 Recall section 4 for the meaning of 'hyperfinite' and 'hyperextension.'

71 One may object against the use of measure theory to represent probability, since measures are motivated by a desire to idealize the notions of physical length, area, and volume, and not probability *per se*. Hence, the usual reservations of representing probability by measure functions, be they standard or non-standard, may apply here.

72 This observation is due to Alan Hájek, whose copy of the article I was allowed to copy.

16.3 *After 1969: Further Developments and Philosophical Discussions*

The 1970s: Further mathematical developments

PARIKH & PARNES' CONDITIONAL PROBABILITY FUNCTIONS Starting from a standard absolute probability function, the ratio formula does not always suffice to define a conditional probability function. This may fail in two ways: the probabilities may be undefined (non-measurable events) or the conditioning event may have probability zero. The non-standard absolute probability function obtained by Bernstein and Wattenberg (1969) does allow us to define a non-standard absolute probability function for all pairs of subsets of the real numbers by the usual ratio formula, provided that the conditioning event is non-empty. By taking the standard part, we obtain a real-valued function defined for all pairs of subsets of the real numbers (as long as the conditioning event is non-empty). However, Rohit Parikh and Milton Parnes (1974) remarked that the conditional probability function so obtained does not necessarily exhibit translation invariance in the following sense:

$$\forall A, B \subseteq \mathbb{R} \text{ such that } B \neq \varnothing,\ \forall x \in \mathbb{R},\ P(A + x, B + x) = P(A, B),$$

where $A + x$ is the set obtained by adding x to all elements of A and P is the standard conditional probability function obtained by applying the ratio formula to a non-standard absolute probability function as constructed by Bernstein and Wattenberg (1969) and then taking the standard part.

Parikh and Parnes did not consider non-standard conditional probability functions. Instead, they merely used NSA as a means of obtaining standard functions. Using techniques from NSA (in particular, hyperfinite sets), Parikh and Parnes constructed standard conditional probability functions, each fulfilling a number of algebraic conditions that correspond with our intuitions. Apart from a condition that entails the above criterion of translation invariance, they also obtained: (i) $P(B, B) = 1$ for all B, (ii) if $B = [0, 1]_{\mathbb{Q}}$ (the unit interval of $\mathbb{Q}$ with endpoints included) and $0 \leq a < b \leq 1$, then $P([a, b], B) = b - a$, and (iii) $P(A, B) = 0$ whenever A is finite and B is not.[73] It requires a bit more effort (choosing a suitable ideal on $\mathbb{R}$, *cf.* section 5) to obtain a function P such that the following stronger version of (iii) also holds: $P(A, B) = 0$ whenever A is countable and B is not. After proving the relevant existence theorems, they showed that the cardinality of the set of standard conditional probability functions satisfying the various combinations of properties is 2^c, with c the cardinality of the continuum.

73 Observe that these conditional probability functions violate regularity, but this should not be surprising since they are real-valued.

HENSON'S REPRESENTATION THEOREM Meanwhile, C. Ward Henson (1972) showed that for every standard, finitely additive probability measure that assigns zero to finite sets there exists a non-standard representation. Once again, the proof relies on a hyperfinite counting measure on a hyperfinite subset of the hyperextension of the sample space of the standard function. He also considered the special case in which the standard measure is countably additive. As is typical in the context of NSA, Henson showed how to apply his result in order to obtain a shorter proof of a standard result (in section 2 of his paper).[74]

LOEB MEASURE Seminal contributions to non-standard measure theory were obtained by Peter A. Loeb (1975). A good overview of this topic (up to the early 1980s) can be found in Cutland (1983). Loeb measures require more advanced technical knowledge than any of the other approaches covered in this chapter. In particular, they require non-standard models with a saturation beyond countable saturation.[75]

DE FINETTI'S RESPONSE As indicated in section 16.1, de Finetti wrote on the topic of non-Archimedean probability rankings well before the development of NSA. Although he lived long enough and was aware of the development of NSA, he never showed much interest in applying it to his own work on probability. This can be seen by inspecting his work from the 1970s.

In the second volume of his 1974 book, de Finetti famously returned to the discussion of possible events with zero probability—a topic already on his mind (and in his publications) in the 1930s. In particular, he wondered whether it is "possible to compare the zero probabilities of possible events" and whether "a union of events with zero probabilities [can] have a positive probability" (de Finetti, 1974, Vol. II, p. 117). On p. 118, he remarks that the latter question can be rephrased in terms of additivity and he distinguishes three cases: finite additivity, countable additivity, and perfect additivity "if the additivity always holds."[76] On p. 119, he discusses weak and strong coherence; of the latter he writes "This means that 'zero probability' is equivalent to 'impossibility'." However, he warns us that besides "these serious authors" who have written on this topic, there are others "who refer to zero probability as impossibility, either to simplify matters in elementary

74 See also Hofweber and Schindler (2016) for "a new and completely elementary proof of this fact."

75 In the construction of ${}^*\mathbb{R}$, we used a free ultrafilter on $\mathbb{N}$ (see part 3). This is sufficient to obtain a model with countable saturation. It is possible to fix a free ultrafilter on an infinite index set of higher cardinality. In particular, by choosing "good" ultrafilters, it is possible to arrive at the desired level of saturation in a single step (Keisler, 2010, section 10). See Hurd and Loeb (1985, pp. 104–108) for more on saturation.

76 *Cf.* ultra-additivity in the terminology of Skyrms (1983b): see section 16.3.

treatments, or because of confusion, or because of metaphysical prejudices." So, according to de Finetti, if we are careful enough not to interpret zero probability as impossibility, we do not need infinitesimal probabilities at this point—in fact, he does not mention them on these pages.

Elsewhere in his book, however, de Finetti does consider non-zero infinitesimal probabilities in relation to additivity. De Finetti (1974, p. 347) writes:

> Let us just mention that the consideration of probability as a non-Archimedean quantity would permit us to say, if we wished, that 'zero probabilities' are in fact 'infinitely small' (actual infinitesimals), and only that of the impossible event is zero. Nothing is really altered by this change in terminology, but it might sometimes be useful as a way of overcoming preconceived ideas. It has been said that to assume that $0+0+0+...+0+...=1$ is absurd, whereas, if at all, this would be true if 'actual infinitesimal' were substituted in place of zero. There is nothing to prevent one from expressing things in this way, [...]

This seems to be a welcoming invitation to adopt techniques from NSA in order to deal with infinitesimal probabilities and associated puzzles concerning their additivity. However, de Finetti continues his sentence less enthusiastically: "[...] apart from the fact that it is a useless complication of language, and leads one to puzzle over '*les infiniment petits*'."[77]

Moreover, in 1979 (as transcribed in de Finetti, 2008, Ch. 12, p. 122), a graduate student asked de Finetti about his thoughts concerning NSA. The student (referred to as 'Alpha' in the transcript) asked: "do you consider it plausible that this hierarchy of zero probabilities could be replaced by a hierarchy of actual infinitesimals in the sense of non-standard analysis?" To which de Finetti responded:

> I only attended a few talks on non-standard analysis and I have to say that I am not sure about its usefulness. On the face of it, it does not persuade me, but I think I have not delved enough into this topic in order to be able [to] give [a] well thought-out judgment. [...] I made those speculations on

77 The French expression '*les infiniment petits*' was in use since the development and popularization of the calculus; consider, for instance, the title of de l'Hôpital's 1696 book, *Analyse des Infiniment Petits pour l'Intelligence des Lignes Courbes*. The use of infinitesimals in calculus was discredited in subsequent years (in favour of epsilon-delta constructions developed in the work of Weierstrass, *cf.* section 16.1). Although NSA did much to reinstate them, this process of rehabilitation of infinitesimals was neither immediate nor uniform (and remains incomplete, even today). So, it seems that de Finetti held on to the post-Weierstrassian and pre-Robinsonian viewpoint of infinitesimals as a suspect concept, to be avoided when possible.

> infinitely small probabilities to see the extent to which the idea of a comparison between zero probabilities is plausible. However, I did not attach much importance to it and I am not sure whether one needs sophisticated theories, such as non-standard analysis, for that goal.

The 1980s: Skyrms, Lewis, and Nelson

SKYRMS ON INFINITESIMAL CHANCES Skyrms (1980) argued that propensity (for instance, the bias parameter in a binomial distribution) does not equal the limiting relative frequency (for instance, of an infinite Bernoulli process). He did so by appealing to infinitesimal probabilities (pp. 30–31):

> If we extend our language so that we can talk in it about limiting relative frequencies in an infinite sequence of trials and make a few assumptions about limiting probabilities, we can state what appears to be a more powerful version of the law of large numbers: the probability that, in a given sequence of independent and identically distributed trials, the limiting relative frequency will either fail to exist or diverge by some positive real number from the probability of the outcome is infinitesimal. Then, if our coin is flipped an infinite number of times, the probability that the limiting relative frequency fails to be one-half is infinitesimal.

He then went on to show that this viewpoint is not compatible with the idea "that infinitesimal propensity implies impossibility." The stance that Skyrms is refuting here is sometimes called the "principle of Cournot."[78]

> [T]he assumptions that get the striking version of the strong law of large numbers give us infinitesimal probability not only for the outcome sequence All Heads, but for each other definite sequence of outcomes as well. But the coin has to do something! There is nothing more probable than that something improbable will happen, but it is impossible that something impossible should happen. Small probability, even infinitesimally small probability, does not mean impossibility. Then even if, for each process, the propensity for a divergence between propensity

78 The principle of Cournot is named after Augustin Cournot, because of his writings on the notion of "physical impossibility" (of events corresponding to infinitesimal probabilities in a geometric context). The roots of the concept go back to that of "moral certainty" (practical certainty) in the work of Jacob Bernoulli. Similar ideas also arose in the work of Paul Lévy and émile Borel (which inspired de Finetti's speculations on hierarchies of infinitesimals). The name for the principle was introduced by Maurice Fréchet. For more details, see, *e.g.*, Shafer (2008).

> and relative frequency is infinitesimal, it hardly follows that the propensity for a divergence for some process, somewhere in the world, is infinitesimal. But this is just what those who wish to turn the law of large numbers into a philosophical analysis of propensity must assume.

Here, Skyrms used infinitesimal probabilities to illustrate the qualitative difference between possible events and the impossible event. In particular, in cases of equiprobability it may be certain that a highly unlikely event will occur. This seems to be diametrically opposed to Cournot's principle and similar ideas such as the Lockean thesis (but see also section 13).

LEWIS ON INFINITESIMAL CHANCES AND CREDENCES David Lewis (1980) introduced his "Principal Principle" as a way to connect subjective credences to objective chances. In this context, he discussed how infinitesimal chances lead to the introduction of infinitesimal credences (p. 269):

> The Principal Principle may be applied as follows: you are sure that some spinner is fair, hence that it has infinitesimal chance of coming to rest at any particular point; therefore (if your total evidence is admissible) you should believe only to an infinitesimal degree that it will come to rest at any particular point.

On pp. 267–268, Lewis (1980) discussed infinitesimal credences in the context of regularity (*cf.* section 16.1) and a "condition of reasonableness":

> I should like to assume that it makes sense to conditionalize on any but the empty proposition. Therefore I require that [any reasonable initial credence function] C is *regular*: $C(B)$ is zero, and $C(A/B)$ is undefined, only if B is the empty proposition, true at no worlds. You may protest that there are too many alternative possible worlds to permit regularity. But that is so only if we suppose, as I do not, that the values of the function C are restricted to the standard reals. Many propositions must have infinitesimal C-values, and $C(A \mid B)$ often will be defined as a quotient of infinitesimals, each infinitely close but not equal to zero. (See Bernstein and Wattenberg [1969].) The assumption that C is regular will prove convenient, but it is not justified only as a convenience. Also it is required as a condition of reasonableness: one who started out with an irregular credence function (and who then learned from experience by conditionalizing) would stubbornly refuse to believe some propositions no matter what the evidence in their favor.

SKYRMS ON REGULARITY AND ULTRA-ADDITIVITY Skyrms (1983b) gave an intriguing analysis of the Zenonian intuition of regularity. His text focused on length measurement, but the argument carries over to probability measures; hence, we present it in some detail. Zeno's paradox of measure is a scholarly reconstruction of an argument against plurality emerging from Zeno's four paradoxes of motion. The conclusion of this argument is that something of non-zero, finite length cannot be composed of infinitely many parts. The Zenonian argument starts by assuming the opposite: if the whole is composed of infinitely many parts, then either those parts all have no magnitude or they all have a non-zero magnitude, but then the whole would either have no magnitude or an infinite magnitude, respectively, both of which are in contradiction with the whole having a non-zero, finite length. Skyrms argued that this argument crucially relies on some implicit assumptions: that the parts all have equal size (invariance), that they are not infinitesimal (Archimedean axiom), and that we can make sense of an infinite sum of the individual magnitudes (ultra-additivity). As such, Zeno's paradox of measure has a very similar structure to the proof that shows that there is no real-valued, countably additive probability function that assigns equal probabilities to single tickets in a lottery on the natural numbers (*cf.* section 8.3): it shows that either assigning zero probability or non-zero probability to individual tickets both fail to yield a normalizable measure, because either the sum over all tickets is zero or it diverges. Analogous assumptions are in place in both arguments: an invariant partition such that the parts have equal magnitudes versus equiprobability; no infinitesimal magnitudes versus real-valued probability; and a way to make sense of infinite sums of magnitudes versus countable additivity.

Skyrms named the additivity assumption in the Zenonian argument the principle of ultra-additivity, which he specified as follows (p. 227):

> the principle that the magnitude of the whole is the sum of the magnitudes of its parts continues to hold good when we have a partition of the whole into an infinite number of parts.

This way of phrasing it—as a property known for finite quantities that is assumed to hold for infinite quantities, too—resembles Leibniz's "*souverain principe*" (see Katz & Sherry, 2012, section 4.3), which in turn can be formalized by the Transfer principle of NSA (as was explained in section 4). In this light, it is curious to observe that the term for the Zenonian principle chosen by Skyrms, ultra-additivity, resonates well within the context of NSA, which is replete with ultrafilters. (This resonance may be curious, but it need not be coincidental—given Skyrms' familiarity with NSA.)

Skyrms also argued that the step in the Zenonian argument that implicitly assumes the principle of ultra-additivity was not contested by the

school of Plato, the school of Aristotle, or the atomists. So, it appears that the principle of ultra-additivity was—possibly without reflection—widely accepted, which suggests that it represents a deeply anchored intuition about magnitudes: if finite magnitudes are to be infinitely divisible (which of course the Zenonian argument tries to refute), then it is hard to imagine for the magnitudes of the parts in the partition *not* to sum to the magnitude of the whole. Skyrms wrote (p. 235): "It is ironic that it is just here that the standard modern theory of measure finds the fallacy."

In the context of measure theory, and thus of standard probability, the principle of ultra-additivity is formalized—and thereby restricted to countable collections—in terms of CA. However, as the failure of the existence of a countably additive fair probability measure on the natural numbers demonstrates, it does not do justice to the underlying intuition of universal summability.

LEWIS ON INFINITESIMAL CHANCES In a postscript to "Causation" (an article that appeared in 1973) and in a passage that appears between brackets, Lewis (1986b, pp. 175–176) discussed infinitesimal chances and presented real-valued probabilities as a rounding off of the true hyperreal chances (with original italics):[79]

> They say that things with no chance at all of occurring, that is with probability zero, do nevertheless happen; for instance when a fair spinner stops at one angle instead of another, yet any precise angle has probability zero. I think these people are making a rounding error: they fail to distinguish zero chance from infinitesimal chance. Zero chance is *no* chance, and nothing with zero chance ever happens. The spinner's chance of stopping exactly where it did was not zero; it was infinitesimal, and infinitesimal chance is still *some* chance.

Although they are not mentioned here, Lewis' wording is very reminiscent of Bernstein and Wattenberg (1969), who wrote "there is still some chance of hitting the point." Also observe that according to the definition that we gave in the introduction, zero is an infinitesimal. Hence, what Lewis is arguing for must be called "non-zero infinitesimals" in our terminology.

NELSON'S RADICALLY ELEMENTARY PROBABILITY THEORY Previously, Edward Nelson (1977) had provided the first axiomatic approach to NSA, which he called "Internal Set Theory" (IST),[80] but he also provided an important alternative approach to infinitesimal probabilities. Nelson

79 Hájek (2012a) cites this passage and calls Lewis work on this topic "[t]he most important philosophical defence of regularity" of which he is aware (p. 414).

80 According to Luxemburg (2007, p. xi):

(1987) developed a "Radically elementary probability theory," which relies on internal probability functions: these functions can be obtained by applying the Transfer principle (recall section 4) to sequences of standard Kolmogorovian probability functions on finite domains. Internal probability functions do not assign probability values to any infinite standard sets, but only to hyperfinite sets. The resulting additivity property is hyperfinite additivity. Nelson's probability functions are regular and they admit infinitesimal values. Unlike much previous work on non-standard probability functions, this approach does not aim at providing a real-valued probability measure (by the standard part function, *cf.* section 4). Precisely by leaving out this step, this framework has the benefit of making probability theory on infinite sample spaces equally simple and straightforward as the corresponding theory on finite sample spaces.

ACKNOWLEDGMENTS

Some parts of this chapter have appeared earlier in an unpublished manuscript called "Hyperreals and their applications," which was circulated as a handout for two tutorial sessions presented at the Formal Epistemology Workshop held in 2012 in Munich, Germany. I am grateful to participants at the workshop for feedback on that manuscript. I thank Danny Vanpoucke for proofreading an earlier version of the current chapter and Mikhail Katz for detailed feedback and corrections mainly pertaining to the historical section. Finally, I thank the editors, Richard Pettigrew and Jonathan Weisberg, for helpful suggestions on improving the organization of this chapter.

This work was supported financially by two grants from the FWO (Research Foundation – Flanders) through grant numbers GoB8616N and Go66918N.

REFERENCES

Adams, E. W. (1962–63). On rational betting systems. *Archiv für mathematische Logik und Grundlagenforschung*, *6*, 7–29. Part 1 of 2.

Adams, E. W. (1959). Two aspects of the theory of rational betting odds. *Technical Report, Berkeley (Univ. of Calif.)* *1*, 9. Rotaprintvervielfältigung.

[F]rom the beginning Robinson was very interested in the formulation of an axiom system catching his non-standard methodology. Unfortunately he did not live to see the solution of his problem by E. Nelson presented in the 1977 paper entitled "Internal Set Theory".

Adams, E. W. (1964). On rational betting systems. *Archiv für mathematische Logik und Grundlagenforschung*, *6*, 112–128. Part 2 of 2.

Albeverio, S., Fenstad, J. E., Hoegh-Krøhn, R., & Lindstrøm, T. (1986). *Non-standard methods in stochastic analysis and mathematical physics*. Pure and Applied Mathematics. Orlando, FL: Academic Press.

Alexander, A. (2014). *Infinitesimal: How a dangerous mathematical theory shaped the modern world*. London, UK: Oneworld.

Anderson, R. M. (1976). A non-standard representation for Brownian motion and Itô integration. *Israel Journal of Mathematics*, *25*, 15–46.

Bair, J., Błaszczyk, P., Ely, R., Henry, V., Kanovei, V., Katz, K. U., ... Shnider, S. (2013). Is mathematical history written by the victors? *Notices of the American Mathematical Society*, *60*, 886–904.

Barrett, M. (2010). The possibility of infinitesimal chances. In E. Eells & J. H. Fetzer (Eds.), *The place of probability in science* (pp. 65–79). Boston Studies in the Philosophy of Science. Springer.

Bartha, P. (2004). Countable additivity and the de Finetti lottery. *The British Journal for Philosophy of Science*, *55*, 301–321.

Bartha, P. & Hitchcock, C. (1999). The shooting-room paradox and conditionalizing on measurably challenged sets. *Synthese*, *118*, 403–437.

Bartha, P. & Johns, R. (2001). Probability and symmetry. *Philosophy of Science*, *68*, S109–S122.

Benacerraf, P. (1965). What numbers could not be. *Philosophical Review*, *74*, 47–73.

Benci, V. & Di Nasso, M. (2003). Numerosities of labelled sets: A new way of counting. *Advances in Mathematics*, *173*, 50–67.

Benci, V., Di Nasso, M., & Forti, M. (2006). The eightfold path to nonstandard analysis. In N. J. Cutland, M. Di Nasso, & D. A. Ross (Eds.), *Nonstandard methods and applications in mathematics* (Vol. 25, pp. 3–44). Lecture Notes in Logic. Wellesley, MA: Association for Symbolic Logic, AK Peters.

Benci, V., Horsten, L., & Wenmackers, S. (2013). Non-Archimedean probability. *Milan Journal of Mathematics*, *81*, 121–151.

Benci, V., Horsten, L., & Wenmackers, S. (2018). Infinitesimal probabilities. *British Journal for the Philosophy of Science*, *69*, 509–552.

Berkeley, G. (1734). *The analyst, a discourse addressed to an infidel mathematician*. London, England: Strand.

Bernstein, A. R. & Wattenberg, F. (1969). Nonstandard measure theory. In W. A. J. Luxemburg (Ed.), *Applications of model theory to algebra, analysis and probability* (pp. 171–185). New York, NY: Holt, Rinehard and Winston.

Błaszczyk, P., Katz, M. G., & Sherry, D. (2013). Ten misconceptions from the history of analysis and their debunking. *Foundations of Science*, *18*, 43–74.

Boolos, G. S., Burgess, J. P., & Jeffrey, R. C. (2007). *Computability and logic*. 5th ed. Cambridge, UK: Cambridge University Press.

Boyer, C. (1949). *The concepts of the calculus*. Hafner Publishing Company.

Brickhill, H. & Horsten, L. (2018). Triangulating non-Archimedean probability. *The Review of Symbolic Logic*, *11*(3), 519–546.

Carlyle, T. (1845). *Oliver Cromwell's letters and speeches: With elucidations*. New York, NY: Wiley and Putnam.

Carnap, R. (1950). *Logical foundations of probability*. Chicago, IL: University of Chicago Press.

Carnap, R. (1971a). A basic system of inductive logic, part I. In R. Carnap & R. C. Jeffrey (Eds.), *Studies in inductive logic and probability* (Vol. 1). Chicago, IL: University of Chicago Press.

Carnap, R. (1971b). Inductive logic and rational decisions. In R. Carnap & R. C. Jeffrey (Eds.), *Studies in inductive logic and probability* (Vol. 1). Chicago, IL: University of Chicago Press.

Carnap, R. (1980). The problem of a more general concept of regularity. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. 2, pp. 145–155). Written in 1960. Berkeley, CA: University of California Press.

Chen, E. & Rubio, D. (2018). *Surreal decisions*. Forthcoming in *Philosophy and Phenomenological Research*; doi:10.1111/phpr.12510.

Cutland, N. (1983). Nonstandard measure theory and its applications. *Bulletin of the London Mathematical Society*, *15*, 529–589.

de Finetti, B. (1931). Sul significato soggettivo della probabilità. *Fundamenta Mathematica*, *18*, 298–329. Translated in English as "On the subjective meaning of probability" in: P. Monari and D. Cocchi (eds.), "Probabilità e Induzione; Induction and Probability" (1993) Clueb, Bologna; pp. 291–321.

de Finetti, B. (1936). Les probabilités nulles. *Bulletin de Sciences Mathématiques*, *60*, 275–288.

de Finetti, B. (1937). La prévision: Ses lois logique, ses sources subjectives. *Annales de l'Institute Henri Poincaré*, *7*, 1–68.

de Finetti, B. (1972). *Probability, induction and statistics; the art of guessing*. London, UK: Wiley.

de Finetti, B. (1974). *Theory of probability*. Translated by: A. Machí and A. Smith. London, UK: Wiley.

de Finetti, B. (2008) In A. Mura (Ed.), *Philosophical lectures on probability* (Vol. 340). Synthese Library. Introductory Essay by Maria Carla Galavotti; translated by: Hykel Hosni. London, UK: Springer.

Dedekind, R. (1888). *Was sind und was sollen die Zahlen?* Braunschweig, Germany: Vieweg.

DiBella, N. (2018). The qualitative paradox of non-conglomerability. *Synthese*, *195*, 1181–1210.

Easwaran, K. (2014). Regularity and hyperreal credences. *Philosophical Review*, *123*, 1–41.

Ehrlich, P. (2006). The rise of non-Archimedean mathematics and the roots of a misconception i: The emergence of non-Archimedean systems of magnitudes. *Archive for History of Exact Sciences*, *60*, 1–121.

Elga, A. (2004). Infinitesimal chances and the laws of nature. *Australasian Journal of Philosophy*, *82*, 67–76.

Feferman, S. (1979). Constructive theories of functions and classes. *Logic Colloquium*, *78*, 159–224.

Feferman, S. (1999). Does mathematics need new axioms? *The American Mathematical Monthly*, *106*, 99–111.

Foley, R. (2009). Beliefs, degrees of belief, and the Lockean thesis. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (Vol. 342, pp. 37–47). Synthese Library. Dordrecht, The Netherlands: Springer.

Gaifman, H. (1986). Towards a unified concept of probability. In R. B. Marcus, G. J. W. Dorn, & P. Weingartner (Eds.), *Logic, methodology and philosophy of science vii* (Vol. 114, pp. 319–350). Studies in Logic and the Foundations of Mathematics. Amsterdam, The Netherlands: Elsevier.

Goldblatt, R. (1998). *Lectures on the hyperreals; an introduction to nonstandard analysis*. Graduate Texts in Mathematics. New York, NY: Springer.

Hacking, I. (1975). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge, UK: Cambridge University Press.

Hájek, A. (2003). Waging war on Pascal's wager. *The Philosophical Review*, *112*, 27–56.

Hájek, A. (2012a). Is strict coherence coherent? *Dialectica*, *66*, 411–424.

Hájek, A. (2012b). *Staying regular?* Unpublished manuscript. Retrieved from http://philrsss.anu.edu.au/sites/default/files/Staying%20Regular.December%2028.2012.pdf

Halpern, J. Y. (2010). Lexicographic probability, conditional probability, and nonstandard probability. *Games and Economic Behavior*, *68*, 155–179.

Heath, T. L. (Ed.). (1897). *The works of Archimedes; edited in modern notation with introductory chapters*. Cambridge, UK: Cambridge University Press.

Henson, C. W. (1972). On the nonstandard representation of measures. *Transactions of the American Mathematical Society*, *172*, 437–446.

Herzberg, F. (2007). Internal laws of probability, generalized likelihoods and Lewis's infinitesimal chances—A response to Adam Elga. *British Journal for the Philosophy of Science*, *58*, 25–43.

Herzberg, F. (2010). The consistency of probabilistic regresses. a reply to Jeanne Peijnenburg and David Atkinson. *Studia Logica*, *94*, 331–345.

Hewitt, E. (1948). Rings of real-valued continuous functions I. *Transactions of the American Mathematical Society*, *64*, 54–99.

Hilbert, D. (1900). Mathematische Probleme. *Göttinger Nachrichten*, 253–297. Translated as "Mathematical Problems", *Bulletin of the American Mathematical Society*, **8**, no. 10 (1902), pp. 437–479.

Hofweber, T. (2014). Infinitesimal chances. *Philosophers' Imprint*, *14*, 1–14.

Hofweber, T. & Schindler, R. (2016). Hyperreal-valued probability measures approximating a real-valued measure. *Notre Dame Journal of Formal Logic*, *57*, 369–374.

Hoover, D. (1980). A note on regularity. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. 2, pp. 295–297). Berkeley, CA: University of California Press.

Howson, C. (2017). Regularity and infinitely tossed coins. *European Journal for Philosophy of Science*, *7*, 97–102.

Hrbáček, K. (2007). Stratified analysis? In I. van den Berg & N. Neves (Eds.), *The strength of nonstandard analysis* (pp. 47–63). Vienna, Austria: Springer.

Hurd, A. E. & Loeb, P. A. (1985). *An introduction to nonstandard real analysis*. Pure and Applied Mathematics. Orlando, FL: Academic Press.

Kanovei, V., Katz, M. G., & Mormann, T. (2013). Tools, objects, and chimeras: Connes on the role of hyperreals in mathematics. *Foundations of Science*, *18*, 259–296.

Katz, M. G. (2014). Leibniz's infinitesimals: Their fictionality, their modern implementations, and their foes from Berkeley to Russell to beyond. *Erkenntnis*, *78*, 571–625.

Katz, M. G. & Sherry, D. (2013). Leibniz's infinitesimals: Their fictionality, their modern implementations, and their foes from Berkeley to Russell to beyond. *Erkenntnis*, *78*, 571–625.

Katz, M. G. & Sherry, D. M. (2012). Leibniz's laws of continuity and homogeneity. *Notices of the American Mathematical Society*, *59*, 1550–1558.

Keisler, H. J. (2010). The ultraproduct construction. In V. Bergelson, A. Blass, M. Di Nasso, & R. Jin (Eds.), *Ultrafilters across mathematics* (Vol. 530, pp. 163–179). Contemporary Mathematics. American Mathematical Society.

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

Kemeny, J. G. (1981). Fair bets and inductive probabilities. *The Journal of Symbolic Logic*, *20*, 263–273.

Kerkvliet, T. & Meester, R. (2016). Uniquely determined uniform probability on the natural numbers. *Journal of Theoretical Probability*, *29*, 797–825.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitrechnung*. Ergebnisse der Mathematik. Translated by N. Morrison, *Foundations of probability.* Chelsea Publishing Company, 1956 (2nd ed.) Berlin, Germany: Springer.

Kolmogorov, A. N. (1948). Algèbres de Boole métriques complètes. *VI Zjazd Matematyków Polskich*, 21–30. Translated by R. C. Jeffrey as "Complete metric Boolean algebras" *Philosophical Studies* 77 pp. 57–66, 1995.

Komjáth, P. & Totik, V. (2008). Ultrafilters. *American Mathematical Monthly*, *115*, 33–44.

Kossak, R. & Schmerl, J. (2006). *The structure of models of Peano Arithmetic*. Oxford Logic Guides. Oxford, UK: Clarendon Press.

Kremer, P. (2014). Indeterminacy of fair infinite lotteries. *Synthese*, *191*, 1757–1760.

Kyburg, H. E., Jr. (1961). *Probability and the logic of rational belief*. Middletown, CT: Wesleyan University Press.

Laplace, P.-S. (1814). *Essai philosophique sur les probabilités*. 3th edition printed by V. Courcier, Paris, France, 1816. Translated by Truscott, F. W., Emory, F. L. *Philosophical Essay on Probabilities.* Wiley (1902) New York, NY. Paris, France.

Lehman, R. S. (1955). On confirmation and rational betting. *The Journal of Symbolic Logic*, *20*, 251–262.

Lewis, D. K. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability* (Vol. 2, pp. 263–293). Berkeley, CA: University of California Press.

Lewis, D. K. (1986a). Philosophical papers. (Chap. A Subjectivist's Guide to Objective Chance, Vol. 2). Oxford, UK: Oxford University Press.

Lewis, D. K. (1986b). *Philosophical papers*. Oxford, UK: Oxford University Press.

Lindley, D. V. (Ed.). (1991). *Making decisions*. 2nd edition. UK: Wiley.

Lindley, D. V. (Ed.). (2006). *Understanding uncertainty*. UK: Wiley.

Loeb, P. A. (1975). Conversion from nonstandard to standard measure spaces and applications in probability theory. *Transactions of the American Mathematical Society*, *211*, 113–122.

Łoś, J. (1955). Quelques remarques, théorèmes, et problèmes sur les classes définissables d'algèbres. In *Mathematical interpretation of formal systems (symposium, amsterdam 1954)* (Vol. 98, pp. 1–13). Studies in Logic and the Foundations of Mathematics. Amsterdam, The Netherlands: North-Holland Publishing Co.

Luxemburg, W. A. (2007). Foreword. In I. van den Berg & N. Neves (Eds.), *The strength of nonstandard analysis* (pp. v–x). Vienna, Austria: Springer.

Mancosu, P. (2009). Measuring the size of infinite collections of natural numbers: Was Cantor's theory of infinite number inevitable? *The Review of Symbolic Logic*, *2*, 612–646.

Martin-Löf, P. (1990). Mathematics of infinity. In P. Martin-Löf & G. Mints (Eds.), *Colog-88 computer logic* (Vol. 417, pp. 146–197). Lecture Notes in Computer Science. Berlin, Germany: Springer.

McCall, S. & Armstrong, D. M. (1989). God's lottery. *Analysis*, *49*, 223–224.

McGee, V. (1994). Learning the impossible. In E. Eells & B. Skyrms (Eds.), *Probability and conditionals: Belief revision and rational decision* (pp. 179–199). Cambridge, UK: Cambridge University Press.

McGee, V. (2002). Nonstandard models of true arithmetic. Lecture notes for course 'Logic II' at MIT; http://web.mit.edu/24.242/www/NonstandardModels.pdf.

Nelson, E. (1977). Internal set theory: A new approach to nonstandard analysis. *Bulletin of the American Mathematical Society*, *83*, 1165–1198.

Nelson, E. (1987). *Radically elementary probability theory*. Princeton, NJ: Princeton University Press.

Oberschelp, W. (1962–63). Über die Begründung wahrscheinlichkeitstheoretischer Axiome durch Wetten. *Archiv für mathematische Logik und Grundlagenforschung*, *6*, 35–51.

Oppy, G. (1990). On Rescher on Pascal's wager. *International Journal for Philosophy of Religion*, *30*, 159–168.

Painlevé, P. (1967). *Analyse des travaux scientifiques*. Reprinted in: "Œuvres de Paul Painlevé", Éditions du CNRS, Paris (1972), Vol. 1, pp. 72–73. Paris, France: Albert Blanchard.

Palmgren, E. (1998). Developments in constructive nonstandard analysis. *The Bulletin of Symbolic Logic*, *4*, 233–272.

Parikh, R. & Parnes, M. (1974). Conditional probabilities and uniform sets. In A. Hurd & P. Loeb (Eds.), *Victoria symposium on nonstandard analysis* (Vol. 369, pp. 177–188). Lecture Notes in Mathematics. Berlin, Germany: Springer.

Parker, M. (2013). Set size and the part–whole principle. *Review of Symbolic Logic*, *6*, 589–612.

Parker, M. (2018). *Symmetry arguments against regular probability: A reply to recent objections*. Unpublished manuscript; URL: http://philsci-archive.pitt.edu/14362/.

Pascal, B. (1670/1995). *Pensées*. Translated by A.J. Krailsheimer. Penguin Classics.

Peano, G. (1889). *Arithmetices principia, nova methodo exposita*. Translated as "The principles of arithmetic, presented by a new method" by J. Van Heijenoort in J. Van Heijenoort, editor, *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*, Harvard University

Press, Cambridge, MA (1977) 83–97; http://books.google.be/books?id=v4tBTBlU05sC&pg=PA83. Turin, Italy: Bocca.

Pedersen, A. P. (2014). Comparative expectations. *Studia Logica*, *102*, 811–848.

Perkins, E. (1981). A global intrinsic characterization of Brownian local time. *The Annals of Probability*, *9*, 800–817.

Pivato, M. (2014). Additive representation of separable preferences over infinite products. *Theory and Decision*, *77*, 31–83.

Pruss, A. (2012). Infinite lotteries, perfectly thin darts, and infinitesimals. *Thought*, *1*, 81–89.

Pruss, A. (2013). Probability, regularity, and cardinality. *Philosophy of Science*, *80*, 231–240.

Pruss, A. (2014). Infinitesimals are too small for countably infinite fair lotteries. *Synthese*, *191*, 1051–1057.

Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (Vol. 5, pp. 156–198). International library of psychology, philosophy, and scientific method. (Original paper from 1926). London, UK: Routledge & P. Kegan.

Robinson, A. (1961). Non-standard analysis. *Proceedings of the Royal Academy of Sciences, Amsterdam, ser. A*, *64*, 432–440.

Robinson, A. (1966). *Non-standard analysis*. Amsterdam, The Netherlands: North-Holland.

Schechter, E. (1997). *Handbook of analysis and its foundations*. San Diego, CA: Academic Press (Elsevier).

Schmieden, C. & Laugwitz, D. (1958). Eine Erweiterung der Infinitesimalrechnung. *Mathematisches Zeitschrift*, *69*, 1–39.

Schurz, G. & Leitgeb, H. (2008). Finitistic and frequentistic approximation of probability measures with or without σ-additivity. *Studia Logica*, *89*, 257–283.

Shafer, G. (2008). The game-theoretic framework for probability. In B. Bouchon-Meunier, C. Marsala, M. Rifqi, & R. R. Yager (Eds.), *Uncertainty and intelligent information systems* (pp. 3–15). Hackensack, NJ: World Scientific.

Shimony, A. (1955). Coherence and the axioms of confirmation. *The Journal of Symbolic Logic*, *20*, 1–28.

Skolem, T. A. (1923). Einige bemerkungen zur axiomatischen begründung der mengenlehre. *Proc. 5th Scandinaviska Matematikerkongressen, Helsingfors, July 4–7, 1922*, 217–232. Translated as "Some remarks on axiomatized set theory" by S. Bauer-Mengelberg in J. Van Heijenoort, editor, *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*, Harvard University Press, Cambridge, MA (1977) 290–301; http://books.google.be/books?id=v4tBTBlU05sC&pg=PA290.

Skolem, T. A. (1934). Über die Nicht-charakterisierbarkeit der Zahlenreihe mittels endlich oder abzählbar unendlich vieler Aussagen mit ausschliesslich Zahlenvariablen. *Fundamenta Mathematicae*, *23*, 150–161.

Skyrms, B. (1980). *Causal necessity*. New Haven, CT: Yale University Press.

Skyrms, B. (1983a). Three ways to give a probability assignment a memory. In J. Earman (Ed.), *Testing scientific theories* (Vol. 10, pp. 157–161). Minnesota Studies in the Philosophy of Science. Minneapolis: University of Minnesota Press.

Skyrms, B. (1983b). Zeno's paradox of measure. In R. S. Cohen & L. Laudan (Eds.), *Physics, philosophy and psychoanalysis: Essays in hounour of Adolf Grunbaum* (pp. 223–254). Dordrecht, The Netherlands: Reidel.

Skyrms, B. (1995). Strict coherence, sigma coherence and the metaphysics of quantity. *Philosophical Studies*, *77*, 39–55.

Stillwell, J. (1977). Concise survey of mathematical logic. *Australian Mathematical Society Journal (Series A)*, *24*, 139–161.

Stolz, O. (1883). Zur Geometrie der Alten, insbesondere über ein Axiom des Archimedes. *Mathematische Annalen*, *22*, 504–519. Based on an earlier publication in "Berichten des naturwissenschaftlich-medicinischen Vereines in Innsbruck", 1882, volume 12, p. 74.

Stolz, O. (1885). *Vorlesungen über allgemeine Arithmetik*. Leipzig, Germany: Teubner.

Tao, T. (2007–2012). Blog posts tagged "nonstandard analysis". http://terrytao.wordpress.com/tag/nonstandard-analysis/.

Tao, T. (2007). Ultrafilters, nonstandard analysis, and epsilon management. http://terrytao.wordpress.com/2007/06/25/ultrafilters-nonstandard-analysis-and-epsilon-management/.

Tao, T. (2012). A cheap version of nonstandard analysis. http://terrytao.wordpress.com/2012/04/02/a-cheap-version-of-nonstandard-analysis/.

Weintraub, R. (2008). How probable is an infinite sequence of heads? A reply to Williamson. *Analysis*, *68*, 247–250.

Wenmackers, S. (2011). *Philosophy of probability: Foundations, epistemology, and computation* (Doctoral dissertation, University of Groningen, Groningen, The Netherlands). http://philpapers.org/archive/WENPOP.

Wenmackers, S. (2012). Ultralarge and infinite lotteries. In B. Van Kerkhove, T. Libert, G. Vanpaemel, & P. Marage (Eds.), *Logic, philosophy and history of science in belgium ii; proceedings of the young researchers days 2010* (pp. 59–66). Belgium, Brussels: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.

Wenmackers, S. (2013). Ultralarge lotteries: Analyzing the lottery paradox using non-standard analysis. *Journal of Applied Logic*, *11*, 452–467.

Wenmackers, S. (2018). *Do infinitesimal probabilities neutralize the infinite utility in Pascal's wager?* Forthcoming in P. Bartha and L. Pasternack (eds.) *Classic Arguments in the History of Philosophy: Pascal's Wager*, Cambridge, UK: Cambridge University Press.

Wenmackers, S. & Horsten, L. (2013). Fair infinite lotteries. *Synthese*, *190*, 37–61.

Williamson, T. (2007). How probable is an infinite sequence of heads? *Analysis*, *67*, 173–180.

6

COMPARATIVE PROBABILITIES

Jason Konek

On the Bayesian view, belief is not just a binary, on-off matter. Bayesians model agents not as simply categorically believing or disbelieving propositions, but rather as having *degrees of confidence*, or *degrees of belief*, or *credences* in those propositions. Rather than flat out believing that your Kimchi Jjigae will turn out splendidly, you might, for example, be 0.7 confident that it will turn out splendidly. Or you might have less precise opinions on the matter. You might be more confident than not that it will turn out splendidly. You might be at least 0.6 confident and at most 0.9 confident that it will turn out splendidly. You might have any number of more or less informative opinions, but nevertheless fall short of having a *precise* credence on the matter. In that case, we say that your credences are *imprecise*.

Credences, whether precise or imprecise, play a number of important theoretical roles according to Bayesians. For example, a rational agent's credences determine *expectations* of measurable quantities—quantities like the size of the deficit 10 years hence, or the utility of an outcome—which capture her *best estimates* of those quantities. Those best estimates, in turn, typically *rationalise* or make sense of her *evaluative attitudes* and *choice behaviour*.

Suppose that you are considering donating to charity. You have credences regarding the cost of bulk food, shipping, and other matters relevant for estimating how good different donation options are. Your credences, let's imagine, determine a higher expected utility for giving cash directly to the poor than for investing in infrastructure development. These expected utilities, on the Bayesian view, capture your best estimates of how much good each option would produce. And these best estimates, in turn, rationalise or make sense of your evaluative attitudes—your opinion that direct-giving is the better action, perhaps. Evaluative attitudes, in turn, rationalise choice behaviour. In the case at hand, your evaluative attitudes rationalise your choice to give cash directly to the poor rather than invest in infrastructure development.

According to many Bayesians, *e.g.*, Koopman (1940a, 1940b), Good (1950), de Finetti (1951), Savage (1954), and Joyce (2010), certain types of doxastic attitudes—opinions of the form "X is at least as likely as Y," known as *comparative beliefs*—play an especially important role in explicating the concept of credence. These explications typically involve an important bit of mathematics known as a *representation theorem*. The aim of this

chapter is straightforward, but fundamental. We will explore three very different approaches to explicating credence using comparative beliefs and representation theorems. Along the way, we will introduce a brand new account of credence: *epistemic interpretivism*. We will also evaluate how these respective accounts stand up to the criticisms of Hájek (2009), Meacham and Weisberg (2011), and Titelbaum (2015).

The rest of the chapter proceeds as follows. Section 1 outlines the main interpretations of *comparative probability orderings*, which mirror the main interpretations of *quantitative probability functions*. Section 2 homes in on one interpretation of comparative probability in particular: the *subjectivist* interpretation. Then it briefly surveys some important representation theorems from Kraft, Pratt, and Seidenberg (1959), Scott (1964), Suppes and Zanotti (1976), and Alon and Lehrer (2014). Section 3 outlines three different "comparativist" accounts of credence: the *measurement-theoretic*, *decision-theoretic*, and *epistemic interpretivist* accounts. Comparativist accounts explain what it is to have one credal state or another in terms of subjective comparative probability relations (or comparative belief relations) and representation theorems. Epistemic interpretivism is an entirely new account of credence. So we spend a bit of time developing it. Section 4 examines criticisms of comparativist accounts by Hájek (2009), Meacham and Weisberg (2011), and Titelbaum (2015). Finally, Section 5, Section 6, and Section 7 explore the extent to which our three different approaches can withstand these criticisms. We will pay special attention to the question of whether they vindicate *probabilism*: the thesis that rational credences satisfy the probability axioms.

1 MAIN INTERPRETATIONS OF COMPARATIVE PROBABILITY

Probabilities seem to pop up all over the place. They feature in the respective explanations of all sorts of different phenomena. They help to explain, for example, *singular events*, such as the outcomes of particular experiments, particular one-off historical events, and the like. Consider some examples:

(1a) Why did the die land with a blue face up?

(1b) It has 1 blue faces and 1 red face. And it's fair. So it had a 5/6 probability of coming up blue.

(2a) Why did Rose get lung cancer?

(2b) She smoked for 30 years. And the probability of getting lung cancer if you smoke for so long is really high.

The high probability (= 5/6) of this particular die coming up blue on this particular toss helps to explain why it in fact came up blue. And the

high probability of this particular person—Rose—getting lung cancer (as a result of her 30 years of smoking) helps to explain why she in fact got lung cancer. Probabilities also help to explain why we ought to have high or low confidence in certain hypotheses. Consider:

(3a) Why should we think that Quantum Electrodynamics is true?

(3b) It's the best confirmed physical theory ever. It's extremely probable given the current evidence.

(4a) Why should we think that Jones stole the paintings?

(4b) Given his acrimonious history with the art museum's curator, the eyewitness testimony, and the DNA evidence, it's quite probable that Jones is guilty.

The extremely high probability of Quantum Electrodynamics given the current evidence at least partially explains why we ought to think that it is true. Likewise, the high probability that Jones stole the paintings given the eyewitness testimony, the DNA evidence, etc., at least partially explains why we ought to think that he is guilty. Finally, probabilities explain and rationalise our *behaviour*. For example:

(5a) Why did you bet Aadil £100 that Manchester City would win their match against Newcastle?

(5b) Have you seen Newcastle lately? They're a joke. It's extremely probable that Manchester City will win.

(6a) Why did you go to Better Food Company rather than Sainsbury's?

(6b) I wanted fresh herbs, and it's more probable that the Better Food Company will have them.

The fact that it is extremely probable, in your view, that Manchester City will win the match helps to explain why you took the bet. It also helps to rationalise or make sense of your decision. And the fact that it's more probable, in your view, that the Better Food Company will have fresh herbs than Sainsbury's helps to explain and rationalise your choice to go to the Better Food Company.

So, probabilities seem to do quite a lot of explanatory work. But no single thing is shouldering the whole explanatory load in (1)–(6). Different kinds of probability do the explaining in different examples. In (1)–(2), it is the *physical probability* or *chance* of the singular event in question that helps to explain why the event actually occurs. (See Hájek, 2009, Gillies, 2000, and Hitchcock, 2012, for discussion of different theories of chance.) In (3)–(4), it is the *logical probability* or *degree of confirmation* of the hypothesis

in question (conditional on the current evidence) that helps to explain why we ought to accept it. (See Earman, 1992, Hajek, 2008, and Paris, 2011, for discussion of Bayesian confirmation theory and some of its issues.) Finally, in (5)–(6), it is the *subjective probability*, or *degree of belief*, or *credence*, of the agent in question that helps to explain and rationalise her choice.

Formally, a *probability function* is just a particular type of real-valued function. Let Ω be a *universal set*, which we can think of as the set of "possible worlds" or "basic possibilities." And let $\mathcal{F}$ be a *Boolean algebra* of subsets of Ω, which we can think of as a set of "propositions." More carefully, we can think of each $X \in \mathcal{F}$ as the proposition that is true at each world $w \in \mathcal{F}$, and false at each $w^* \notin \mathcal{F}$. A Boolean algebra $\mathcal{F}$ of subsets of Ω has three important properties: (*i*) $\mathcal{F}$ contains Ω (*i.e.*, $\Omega \in \mathcal{F}$); (*ii*) $\mathcal{F}$ is closed under complementation (*i.e.*, if $X \in \mathcal{F}$, then $\Omega - X \in \mathcal{F}$); and (*iii*) $\mathcal{F}$ is closed under unions (*i.e.*, if $X, Y \in \mathcal{F}$, then $X \cup Y \in \mathcal{F}$). A real-valued function $p : \mathcal{F} \to \mathbb{R}$ is a *probability function* if and only if it satisfies the laws of (finitely additive) probability.

1. Normalization. $p(\Omega) = 1$.
2. Nonnegativity. $p(X) \geq p(\varnothing)$.
3. Finite Additivity. If $X \cap Y = \varnothing$, then $p(X \cup Y) = p(X) + p(Y)$.

Axiom 1 says that p must assign probability 1 to the tautologous proposition Ω. Axiom 2 says that p must assign at least as high a probability to every proposition X as it does to the contradiction $\varnothing$. Axiom 3 says that the probability that p invests in a disjunction of incompatible propositions X and Y must be the sum of the probabilities that it invests in X and Y, respectively.

The three main interpretations of probability—*physical probability*, or *chance*; *logical probability*, or *degree of confirmation*; and *subjective probability*, or *degree of belief*, or *credence*—correspond to the three main types of phenomena that we use probability functions to model. For example, according to propensity theories of chance, chance functions measure how strongly a causal system is disposed to produce one outcome or other on a particular occasion. Chance functions, on this view, are just probability functions that are used to model one type of physical system (causal systems) as having one type of gradable property (causal dispositions of varying strengths). Likewise, logical probability functions measure how strongly a body of evidence E supports or confirms a given hypothesis H. Logical probability functions are just probability functions that are used to model another type of target system (systems of propositions describing data and hypotheses) as having another type of gradable property (as having hypotheses which are supported to varying degrees by data propositions). Finally, subjective probability functions, or credence functions,

measure (roughly) how confident an agent with some range of doxastic attitudes can be said to be of various propositions. Subjective probability functions are just probability functions that are used to model yet another type of system (an agent's doxastic attitudes) as having yet another type of gradable property (as either constituting or licensing varying degrees of confidence).

Of course, sorting out the precise relationship between these various models—probability functions—and their respective target systems is a delicate task. The "interpretations" above provide only rough, first-pass descriptions of that relationship. Part of this chapter's goal is to explore the relationship between *subjective* probability functions, in particular, and the underlying system of doxastic attitudes that they model.

Just as there are a few main interpretations of *quantitative probability functions*, corresponding to the main types of phenomena that we use those probability functions to model, so too are there a few main interpretations of "comparative probability orderings." Formally, a comparative probability ordering is just particular type of relation $\succeq$ on a Boolean algebra $\mathcal{F}$ of subsets of Ω. On each of the three main interpretations, "$X \succeq Y$" means roughly that X *is at least as probable as* Y. (What *exactly* this amounts to, however, will vary from interpretation to interpretation.) Traditionally, a relation $\succeq$ on $\mathcal{F}$ is said to be a *comparative probability ordering* if and only if it satisfies de Finetti's (1964, pp. 100–101) axioms of comparative probability.

1. Nontriviality. $\Omega \succeq \varnothing$ and $\varnothing \not\succeq \Omega$.
2. Nonnegativity. $X \succeq \varnothing$.
3. Transitivity. If $X \succeq Y$ and $Y \succeq Z$, then $X \succeq Z$.
4. Totality. $X \succeq Y$ or $Y \succeq X$.
5. Quasi-Additivity. If $X \cap Z = Y \cap Z = \varnothing$, then $X \succeq Y$ if and only if $X \cup Z \succeq Y \cup Z$.

Axiom 1 says that the tautology Ω is strictly more probable than the contradiction $\varnothing$. Axiom 2 says that every proposition X is at least as probable as the contradiction $\varnothing$. Axioms 3 and 4 guarantee that $\succeq$ is a total preorder (*i.e.*, reflexive, transitive, and total). Finally, axiom 5 says that disjoining X and Y with some incompatible Z does nothing to alter their comparative probability; so X is at least as probable as Y if and only if the disjunction of X and Z is at least as probable as the disjunction of Y and Z.

The three main interpretations of comparative probability correspond to the three main types of phenomena that we use comparative probability

orderings to model. *Physical comparative probability orderings*, or *chance orderings*—on one theory of chance anyway, *viz.*, propensity theory—model causal systems. In particular, they model causal systems $\mathcal{C}$ as being more or less strongly disposed to produce one outcome or another on a particular occasion:

$$X \succeq Y$$
iff
$\mathcal{C}$ is at least as strongly disposed to produce an outcome $w \in X$ and thereby make X true as it is to produce an outcome $w^* \in Y$ and thereby make Y true.

Logical comparative probability orderings model a rather different type of target system: systems of propositions describing data and hypotheses. In particular, they model certain data D as supporting or confirming certain hypotheses H more than other data D^* support other hypotheses H^*:

$$\langle H, D\rangle \succeq \langle H^*, D^*\rangle$$
iff
D supports or confirms H at least as much as D^* supports or confirms H^*.

Finally, *subjective comparative probability orderings* model yet another type of target system: an agent's doxastic attitudes. In particular, they model agent $\mathcal{A}$ as being more or less confident that one proposition or another is true:

$$X \succeq Y$$
iff
$\mathcal{A}$ is at least as confident that X is true as she is that Y is true.

Of course, the three main interpretations of comparative probability are really *families* of interpretations. All three types of comparative probability orderings come in different flavours. For example, behaviorists like de Finetti (1931, 1964) and Savage (1954) treat subjective comparative probability orderings as particular types of *preference orderings*. To be more confident that X is true than Y is, roughly speaking, to prefer a dollar bet on X to a dollar bet on Y. They subscribe to what Jeffrey calls *the thesis of the primacy of practical reason*, which says that between belief and preference, "it is preference that is real and primary" (Jeffrey, 1987, p. 590). Hence, "belief states that correspond to identical preference rankings of propositions are in fact one and the same" (Jeffrey, 1965/1983, p. 138).

Jeffrey (1965, 2002) and Joyce (1999), in contrast, do not subscribe to this thesis. On their view, being more confident that X is true than Y involves making a peculiarly *doxastic* judgment. Such doxastic judgments partially explain and rationalise our preferences. But they do not even supervene

on preferences, let alone reduce to them. (Two agents, for example, could both be in a state of nirvana on this view, and so be indifferent between every prospect and the status quo, but nevertheless make different comparative probability judgments.) And the laws governing *rational* subjective comparative probability judgments, on this account, are not simply special cases of the laws governing rational preference. Rather, they derive from peculiarly *epistemic* considerations, *e.g.*, considerations of *accuracy*.

To have a general way of talking about comparative beliefs, without assuming that they satisfy de Finetti's axioms, let's introduce some terminology. Call any relation $\succeq$ on an algebra $\mathcal{F}$ of subsets of Ω that is used to model an agent's comparative beliefs a *comparative belief relation*. And call $\langle \Omega, \mathcal{F}, \succeq \rangle$ a *comparative belief structure*. Comparative belief relations may or may not be comparative probability orderings. That is, they may or may not satisfy de Finetti's axioms of comparative probability.

Why the hubbub about comparative belief? Why think that comparative belief relations have a particularly important role to play in modeling rational agents' doxastic states? What are they especially suited to do that precise, real-valued credence functions are not?

There are a number of common answers to these questions. The first is that comparative belief relations provide a more psychologically realistic model of agents' doxastic attitudes than precise, real-valued credence functions. Often I simply lack an opinion about which of two propositions is more plausible. I am not *more* confident that copper will be greater than £2/lb in 2025 (call this proposition C) than I am that nickel will be greater than £3/lb in 2025 (call this proposition N). Neither am I *less* confident in C than N, nor *equally* confident. I simply lack an opinion on the matter. We can model this using comparative belief relations. Just choose a relation $\succeq$ that does not rank C and N:

$$C \not\succeq N \text{ and } N \not\succeq C.$$

The incompleteness in $\succeq$ reflects my lack of opinionation. Precise credence functions, on the other hand, do not allow for this sort of lack of opinionation. *Any* agent with precise credences for C and N takes a stand on their comparative plausibility. She is either more confident in C than N, less confident in C than N, or equally confident in the two.[1]

The second answer is *evidentialist*. Not only do real agents *in fact* have sparse and incomplete opinions, but they *ought* to have such opinions. If your evidence is incomplete and unspecific, then your comparative beliefs (and your other qualitative and comparative opinions) should be correspondingly incomplete to reflect the unspecific nature of that evidence. This is the response that is most *justified*, or *warranted*, or *appropriate* in

1 See Suppes (1994, p. 19), Kyburg and Pittarelli (1996, p. 325), Kaplan (2010, p. 47), and Joyce (2010, p. 283) for similar remarks.

light of such evidence. Again, we can capture this sort of lack of opinionation using comparative belief relations, but not using precise credence functions. Having precise credences requires having total or complete comparative beliefs (as well as total conditional comparative beliefs, total preferences, and so on).[2]

The third answer is *information-theoretic*. Proponents of maximum entropy methods, for example, argue that you ought to have the *least informative* doxastic state consistent with your evidence. And according to any plausible informativeness or entropy measure for comparative beliefs, any incomplete comparative belief relation will be less informative than any extension of it.[3] As a result, minimizing informativeness will often require adopting incomplete comparative beliefs. Precise credences, however, do not allow for incomplete comparative beliefs. As Joyce puts it, adopting precise credences, in many evidential circumstances, "amounts to pretending that you have lots and lots of information that you simply don't have" (Joyce, 2010, p. 284).

The final common answer is that comparative belief is more explanatorily fundamental than precise credence. Comparative beliefs figure into the best explanation of *what precise credences are*, but not vice versa. It is worthwhile, then, exploring the various accounts of credence that aim to furnish such an explanation. We will turn our attention to them shortly. Each of these accounts, however, makes use of an important bit of mathematics known as a *representation theorem*. So our first task is to get familiar with the nuts and bolts of representation theorems.

2 REPRESENTATION THEOREMS

Suppose that Monty Hall invites you to choose one of three doors: either door a, b, or c. Behind one of these doors: a car. Behind the other two: a goat. You are more confident that the car is behind a than b, let's imagine. You are also more confident that it's behind b than c. But that is all. You do not take a stand, for example, on whether it's more likely to be behind either b or c than a, or vice versa. You abstain from judgment on all other matters.

Let w_a be the world in which the car is behind door a, w_b be the world in which the car is behind door b, and w_c be the world in which the car is

2 See Joyce (2005, p. 171) for similar remarks.

3 See Abellan and Moral (2000, 2003) for measures of entropy for imprecise probability models which might also serve as measures of entropy for comparative belief relations.

behind door c. Then we can represent you as having comparative beliefs about propositions in the following Boolean algebra:

$$\mathcal{F} = \left\{ \begin{array}{c} \{w_a, w_b, w_c\} \\ \{w_a, w_b\}, \{w_a, w_c\}, \{w_b, w_c\}, \\ \{w_a\}, \{w_b\}, \{w_c\}, \\ \varnothing \end{array} \right\}.$$

And we can represent those fragmentary comparative beliefs as follows:

$$\{w_a, w_b, w_c\} \succ \{w_a\} \succ \{w_b\} \succ \{w_c\} \succ \varnothing,$$

where $X \succ Y$ is shorthand for $X \succeq Y$ and $X \not\preceq Y$.[4]

Your comparative belief relation $\succeq$ is *not* a comparative probability ordering, *i.e.*, $\succeq$ does not satisfy de Finetti's axioms of comparative probability. Relation $\succeq$ violates Quasi-Additivity, for example, as well as totality. You are, after all, more confident that the car is behind a than b:

$$\{w_a\} \succ \{w_b\}.$$

So de Finetti's Quasi-Additivity axiom demands that you also be more confident that it is behind a or c than you are that it is behind b or c. But you abstain from judgment on the matter:

$$\{w_a, w_c\} \not\succeq \{w_b, w_c\} \quad \text{and} \quad \{w_a, w_c\} \not\preceq \{w_b, w_c\}.$$

A few back-of-the-envelope calculations suffice to show that de Finetti's axioms are necessary for *probabilistic representability*. Following Savage (1954), we say that:

$$\begin{array}{c} p \textit{ fully agrees} \text{ with } \succeq \\ \textit{iff} \\ X \succeq Y \Leftrightarrow p(X) \geq p(Y). \end{array}$$

We say that $\succeq$ is (fully) *probabilistically representable* iff there is a probability function p that fully agrees with $\succeq$. Since your comparative belief relation $\succeq$ does not satisfy de Finetti's axioms, in our little example, it is not probabilistically representable.

4 This shorthand is inadequate. You may well think $X \succeq Y$ and $X \not\preceq Y$ without thinking $X \succ Y$. Imagine for example that you recently learned that Y entails X. So you think that X is at least as likely as Y, *i.e.*, $X \succeq Y$. But you have no idea whether the entailment goes both ways. So you withhold judgment about whether Y is at least as likely as X, *i.e.*, $X \not\preceq Y$. For exactly the same reason you withhold judgment about whether X is strictly more likely than Y, *i.e.*, $X \not\succ Y$. We would do better, then, to represent your doxastic state with a *pair* of relations $\langle \succ, \succeq \rangle$. But historically one or the other has been taken as primitive. For ease of exposition, we follow de Finetti (1951), Savage (1954), and Krantz, Luce, Suppes, and Tversky (1971), who take $\succeq$ as primitive.

De Finetti (1951) famously conjectured that his axioms encode not only necessary conditions for probabilistically representability, but *sufficient* conditions as well. The question: was de Finetti right?

Let $\langle \Omega, \mathcal{F}, \succeq \rangle$ be an agent's comparative belief structure. (For now, assume that $\mathcal{F}$ is finite.) Probability functions that fully agree with $\succeq$ show that we can think of that structure $\langle \Omega, \mathcal{F}, \succeq \rangle$ numerically, so to speak. We can map the propositions X in $\mathcal{F}$ to real-valued proxies $p(X)$. And we can do so in such a way that one proxy $p(X)$ is larger than another $p(Y)$ exactly when our agent is more confident in X than Y. So the familiar "greater than or equal to" relation $\geq$ on the real numbers $\mathbb{R}$ is a mirror image of our agent's comparative belief relation $\succeq$ on $\mathcal{F}$.

Kraft et al. (1959) show that de Finetti's conjecture is false. Though de Finetti's axioms are *necessary* for probabilistic representability, they are not *sufficient*. To establish this, Kraft et al. construct a clever counterexample involving a comparative belief relation $\succeq$ over the Boolean algebra $\mathcal{G}$ of all subsets of $\Omega = \{w_a, w_b, w_c, w_d, w_e\}$. Their relation $\succeq$ satisfies de Finetti's axioms of comparative probability, and also the following inequalities:

$$\{w_d\} \succ \{w_a, w_c\}, \tag{1}$$

$$\{w_b, w_c\} \succ \{w_a, w_d\}, \tag{2}$$

$$\{w_a, w_e\} \succ \{w_c, w_d\}. \tag{3}$$

As a result, any probability function p that fully agrees with $\succeq$ satisfies the corresponding inequalities (to simplify notation, we let $p(\{w\}) = p(w)$):

$$p(w_d) > p(w_a) + p(w_c), \tag{4}$$

$$p(w_b) + p(w_c) > p(w_a) + p(w_d), \tag{5}$$

$$p(w_a) + p(w_e) > p(w_c) + p(w_d). \tag{6}$$

But any p that satisfies (4)–(6) also satisfies (7) (simply sum the inequalities).

$$p(w_b) + p(w_e) > p(w_a) + p(w_c) + p(w_d). \tag{7}$$

Notice, however, that $\{w_b, w_e\}$ and $\{w_a, w_c, w_d\}$ appear nowhere in (1)–(3). So Transitivity does not constrain how you order them. Neither do any supersets of $\{w_b, w_e\}$ or $\{w_a, w_c, w_d\}$ appear there. So Quasi-Additivity does not constrain how you order them either. Hence, for all de Finetti's axiom say, you can order $\{w_b, w_e\}$ and $\{w_a, w_c, w_d\}$ any way you please. So Kraft et al. make $\succeq$ satisfy (8).

$$\{w_b, w_e\} \preceq \{w_a, w_c, w_d\}. \tag{8}$$

But if p fully agrees with $\succeq$, then (8) requires:

$$p(w_b) + p(w_e) \leq p(w_a) + p(w_c) + p(w_d). \tag{9}$$

Lines (7) and (9), however, are jointly unsatisfiable. So no probability function p fully agrees with $\succeq$.

2.1 *Scott's Theorem*

So de Finetti's conjecture is false. De Finetti's axioms of comparative probability are *not* necessary and sufficient for probabilistic representability. Luckily, Kraft et al. (1959) and Scott (1964) provide the requisite fix. They provide stronger axioms that are both necessary and sufficient for probabilistic representability. Scott's axioms are more straightforward, so we will focus our attention on them.

Before stating Scott's theorem, it is worth noting that our formulation abuses notation a bit. Expressions 'X_i' and 'Y_i' refer both to propositions in $\mathcal{F}$, as well as their characteristic functions, *i.e.*, functions that take the value 1 at worlds w where X_i (or Y_i) is true (*i.e.*, $w \in X_i$), and take the value 1 at worlds w' where X_i is false (*i.e.*, $w' \notin X_i$). This will turn out to be a helpful bit of sloppiness.

Scott (1964) proves the following:

> **Scott's Theorem**. Every comparative belief structure $\langle \Omega, \mathcal{F}, \succeq \rangle$ (with finite $\mathcal{F}$) has a probability function $p : \mathcal{F} \to \mathbb{R}$ that fully agrees with $\succeq$ in the sense that
>
> $$X \succeq Y \Leftrightarrow p(X) \geq p(Y)$$
>
> if and only if $\succeq$ satisfies the following axioms.
>
> 1. Non-triviality. $\Omega \succ \varnothing$.
> 2. Non-negativity. $X \succeq \varnothing$.
> 3. Totality. $X \succeq Y$ or $Y \succeq X$.
> 4. Isovalence. If $X_1 + \ldots + X_n = Y_1 + \ldots + Y_n$ and $X_i \succeq Y_i$ for all $i \leq n$, then $X_i \preceq Y_i$ for all $i \leq n$ as well.

Axiom 4—sometimes called *Scott's axiom*, the *Isovalence axiom*, or the *Finite cancellation axiom*—is the one new axiom of the bunch. To see what it says, note that $X_1(w) + \ldots + X_n(w)$ counts the number of truths in the set $\{X_1, \ldots, X_n\}$ at world w. Ditto for $Y_1(w) + \ldots + Y_n(w)$. So

$$X_1 + \ldots + X_n = Y_1 + \ldots + Y_n$$

says that the two sets of propositions, $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$, contain the same number of truths *come what may*, *i.e.*, in every possible world. We call such sets of propositions *isovalent*.

In light of this, the Isovalence axiom says that if $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$ are isovalent, then you cannot think that the X_is are uniformly more plausible than the Y_is. ("Uniformly more plausible" here means that you think that X_i is at least as plausible as Y_i for all i, and that X_j is strictly more plausible than Y_j for some j.) After all, an equal number of the X_is

and Y_is are *guaranteed* to be true! So you can only think that the X_is are at least as plausible as the Y_is across the board if you think that they are *equally plausible*.

But how exactly does Scott *prove* his representation theorem? It is worth walking through the proof strategy informally. This will help interested readers dig through the mathematical minutia in Scott (1964).

Indeed, it will prove instructive to use Scott's strategy to establish something slightly stronger than Scott's theorem.

> **Generalised Scott's Theorem (GST)**. For any comparative belief structure $\langle \Omega, \mathcal{F}, \succeq \rangle$ with finite $\mathcal{F}$ and a comparative belief relation $\succeq$ that satisfies
>
> 1. Non-triviality. $\Omega \succ \varnothing$,
> 2. Non-negativity. $X \succeq \varnothing$,
>
> the following two conditions are equivalent.
>
> 3. Isovalence. If $X_1 + \ldots + X_n = Y_1 + \ldots + Y_n$ and $X_i \succeq Y_i$ for all $i \leq n$, then $X_i \preceq Y_i$ for all $i \leq n$ as well.
> 4. Strong representability. there exists a probability function $p : \mathcal{F} \to \mathbb{R}$ that *strongly agrees* with $\succeq$ in the sense that
>
> (i) $X \succeq Y \Rightarrow p(X) \geq p(Y)$,
>
> (ii) $X \succ Y \Rightarrow p(X) > p(Y)$.

Scott's theorem, as we shall see at the end of Section 2, follows fairly straightforwardly from GST. We prove the GST in the appendix.

The key insight required for proving GST is this. In the presence of Non-triviality and Non-negativity, strong representability boils down to sorting *almost desirable gambles* from *undesirable gambles*.[5] On top of this, Scott (1964) shows that sorting almost desirable from undesirable gambles is equivalent to satisfying Isovalence.[6] Figure 1 summarizes the situation.

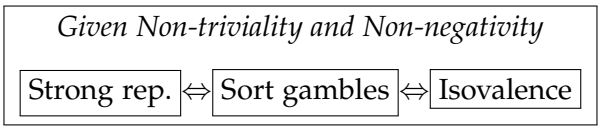


Figure 1: Logical relations between properties of $\succeq$

5 For an accessible introduction to desirable gambles, see Walley (2000). See Quaeghebeur (2014) for more detail.

6 More carefully, Scott (1964) shows that for any comparative belief relation $\succeq$ that satisfies Non-triviality and Non-negativity, satisfying Isovalence is *sufficient* for sorting almost desirable from undesirable gambles. Showing that it is *necessary* is straightforward. See the appendix for proof.

Gambles are measurable quantities $\mathcal{G} : \Omega \to \mathbb{R}$. Say that a gamble $\mathcal{G}$ is *almost desirable* relative to $\succeq$ iff it is a non-negative linear combination of almost desirable components:

$$(X_1 - Y_1), \ldots, (X_n - Y_n).$$

And say that each component $X_i - Y_i$ is almost desirable iff $X_i \succeq Y_i$. Gamble $\mathcal{G}$ is a non-negative linear combination of $(X_1 - Y_1), \ldots, (X_n - Y_n)$ just in case:

$$\mathcal{G} = \sum_i \lambda_i (X_i - Y_i)$$

for some $\lambda_1, \ldots, \lambda_n \geq 0$.

We call components $X_i - Y_i$ almost-desirable if $X_i \succeq Y_i$ because any probability function p that strongly agrees with $\succeq$ determines a non-negative expected value for $X_i - Y_i$:

$$\begin{aligned} X_i \succeq Y_i \quad &\Rightarrow \quad p(X_i) \geq p(Y_i) \\ &\Leftrightarrow \quad E_p[X_i] \geq E_p[X_i] \\ &\Leftrightarrow \quad E_p[X_i - Y_i] \geq 0. \end{aligned}$$

So if we interpret those values as payoffs in utility, then p expects $X_i - Y_i$ to be at least as good as the status quo (*i.e.*, its expected utility is non-negative).

Likewise, we call $\mathcal{G}$ almost desirable if it is a non-negative linear combination of almost-desirable components because any probability function p that strongly agrees with $\succeq$ determines a non-negative expected value for $\mathcal{G}$:

$$\begin{aligned} X_i \succeq Y_i \quad \text{for all } i \quad &\Rightarrow \quad E_p[X_i - Y_i] \geq 0 \quad \text{for all } i \\ &\Rightarrow \quad \sum_i \lambda_i E_p[X_i - Y_i] \geq 0 \\ &\Leftrightarrow \quad E_p\left[\sum_i \lambda_i (X_i - Y_i)\right] \geq 0 \\ &\Leftrightarrow \quad E_p[\mathcal{G}] \geq 0. \end{aligned}$$

Similar remarks apply to *undesirable gambles*. We call a gamble $\mathcal{G}^*$ *undesirable* relative to $\succeq$ iff it is a convex combination of undesirable components

$$(X_1^* - Y_1^*), \ldots, (X_n^* - Y_n^*),$$

so that

$$\mathcal{G}^* = \sum_i \lambda_i^* (X_i^* - Y_i^*).$$

for some $\lambda_1^*, \ldots, \lambda_n^* \geq 0$ with $\sum_i \lambda_i^* = 1$. A component $X_i^* - Y_i^*$ is undesirable iff $X_i^* \prec Y_i^*$. The reason is the same as before. Any probability

function that strongly agrees with $\succeq$ determines a negative expected value for undesirable components, as well as convex combinations of undesirable components.

Now say that $\succeq$ *sorts almost desirable gambles from undesirable ones* iff the two sets of gambles are disjoint. That is, if

$$\mathbb{A} = \{\mathcal{G} \mid \mathcal{G} \text{ is almost desirable rel. to } \succeq\}$$

and

$$\mathbb{U} = \{\mathcal{G}^* \mid \mathcal{G}^* \text{ is undesirable rel. to } \succeq\},$$

then $\succeq$ sorts almost desirable gambles from undesirable ones iff

$$\mathbb{A} \cap \mathbb{U} = \varnothing.$$

If $\succeq$ fails to sort gambles in this way, then some gamble is both almost desirable and undesirable, *i.e.*, $\mathcal{G} = \mathcal{G}^*$ for some almost desirable gamble $\mathcal{G}$ and some undesirable gamble $\mathcal{G}^*$. And if *that* is the case, then there is no probability that *strongly* agrees with it. (Moreover, since full probabilistic representability entails strong representability, there is no probability function that fully agrees with it either.) If it *were* strongly representable, then we would have both

$$E_p[\mathcal{G}] = E_p[\mathcal{G}^*]$$

and

$$E_p[\mathcal{G}^*] < 0 \leq E_p[\mathcal{G}]$$

for some probability function p.

This shows that sorting almost desirable from undesirable gambles is *necessary* for strong agreement with a probability function, which is itself *necessary* for full agreement with a probability function, *i.e.*, probabilistic representability. Scott's insight, though, is that it is also *sufficient*, in the presence of Non-triviality and Non-negativity. Given that $\succeq$ satisfies Non-triviality and Non-negativity, it sorts almost desirable from undesirable gambles *if and only if* it strongly agrees with a probability function. What's more, if $\succeq$ is total as well, then *strong* agreement is equivalent to *full* agreement. So non-trivial, non-negative, total comparative belief relations sort almost desirable from undesirable gambles *if and only if* they are probabilistically representable. See Figure 2.

To prove this, Scott uses what is known as a *hyperplane separation theorem*. The hyperplane separation theorem guarantees that for any two closed, convex, disjoint sets, there is a hyperplane that strictly separates them (Kuhn & Tucker, 1956, p. 50). Now note that $\mathbb{A}$ is the closed, convex polyhedral cone generated by the set $\{X - Y \mid X \succeq Y\}$. Likewise, $\mathbb{U}$ is the convex hull of $\{Y - X \mid X \succ Y\}$—a closed and convex set. And if $\succeq$ sorts

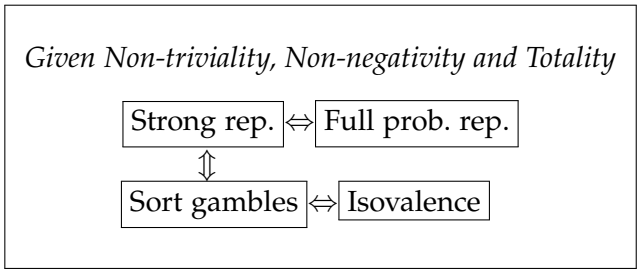


Figure 2: Logical relations between properties of $\succeq$

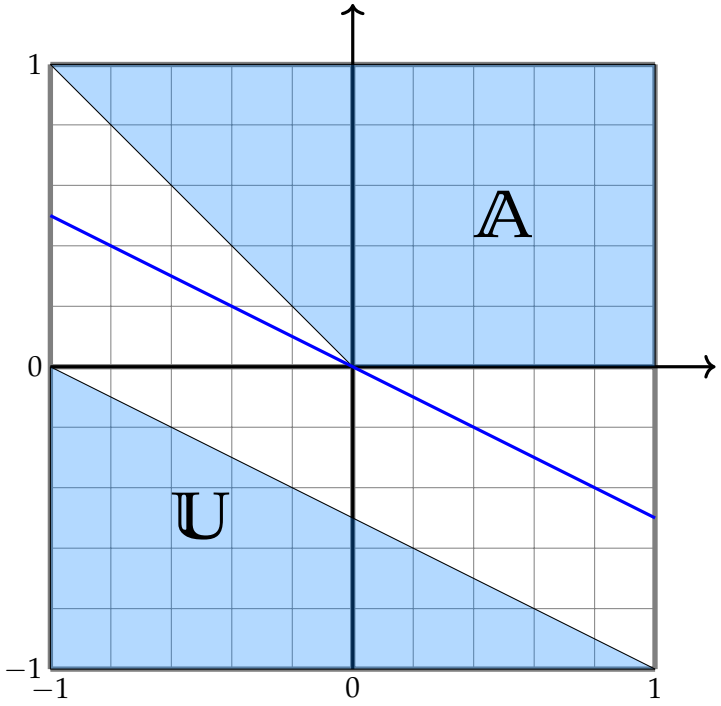


Figure 3: Hyperplane strictly separating $\mathbb{A}$ and $\mathbb{U}$

almost desirable from undesirable gambles, then they are also disjoint. So there is a hyperplane that strictly separates $\mathbb{A}$ and $\mathbb{U}$ (see Figure 3).

This hyperplane determines (in effect) an expectation operator E. Gambles $\mathcal{G}$ on one side of the hyperplane get positive expected values according to E. Gambles on the other side get negatives ones. Precisely *how* high or low $E[\mathcal{G}]$ happens to be is determined by $\mathcal{G}$'s distance from the hyperplane.

The resulting expectation operator E assigns a non-negative value to every almost desirable gamble $\mathcal{G}$ in $\mathbb{A}$, and a negative value to every undesirable gamble $\mathcal{G}^*$ in $\mathbb{U}$:

$$E[\mathcal{G}] \geq 0 \text{ for all } \mathcal{G} \in \mathbb{A},$$

$$E[\mathcal{G}^*] < 0 \text{ for all } \mathcal{G}^* \in \mathbb{U}.$$

And from this expectation operator, E, it is fairly straightforward to extract a probability function p that strongly agrees with $\succeq$. Just let $p(X) = E[X]$ for all $X \in \mathcal{F}$.[7] Then we have:

$$\begin{aligned} X_i \succeq Y_i &\Rightarrow E[X_i - Y_i] \geq 0 \\ &\Leftrightarrow E[X_i] \geq E[Y_i] \\ &\Leftrightarrow p(X_i) \geq p(Y_i). \end{aligned}$$

We also have:

$$\begin{aligned} X_i^* \prec Y_i^* &\Rightarrow E[X_i^* - Y_i^*] < 0 \\ &\Leftrightarrow E[X_i^*] < E[Y_i^*] \\ &\Leftrightarrow p(X_i^*) < p(Y_i^*). \end{aligned}$$

The upshot: sorting almost desirable from undesirable gambles is both *necessary* and *sufficient* for strong agreement with a probability function (in the presence of Non-triviality and Non-negativity).

Now for the kicker: a comparative belief relation $\succeq$—whether or not it satisfies Non-triviality and Non-negativity—sorts almost desirable from undesirable gambles (in the sense that $\mathbb{A} \cap \mathbb{U} = \varnothing$) if and only if it satisfies Isovalence.[8] Hence non-trivial and non-negative $\succeq$ are strongly representable if and only if they satisfy Isovalence. What's more, as we mentioned above, for total comparative belief relations it's easy to see that *strong* agreement with a probability function is equivalent to *full* agreement. So non-trivial, non-negative, total $\succeq$ are *fully* probabilistically representable if and only if they satisfy Isovalence. This is the main thrust of Scott's theorem.

2.2 *Varieties of Representability*

There are, of course, other types of representability besides just *strong* and *full* probabilistic representability. For example, a probability function p strongly agrees with $\succeq$ just in case it satisfies two conditions:

$$\begin{aligned} X \succeq Y &\Rightarrow p(X) \geq p(Y), \\ X \succ Y &\Rightarrow p(X) > p(Y). \end{aligned}$$

7 More methodically, the hyperplane separation theorem gives a strictly separating linear functional, ϕ. But given that $\succeq$ satisfies Non-triviality and Non-negativity, $\mathbb{A}$ and $\mathbb{U}$ have a certain structure, which guarantees that we can normalise ϕ to arrive at an expectation operator E. For example, Non-triviality ensures that $\varnothing - \Omega \in \mathcal{Y}$. Hence $\phi(\Omega) > 0$. Normalising then gives us $E[\Omega] = 1$. Similarly, Non-negativity ensures that $X - \varnothing \in \mathcal{X}$. Hence $\phi(X) \geq 0$, and in turn $E[X] \geq 0$.

8 See Scott (1964, pp. 235–6) for the proof of sufficiency. We present a simplified version of both necessity and sufficiency in the appendix.

We can pick apart these two conditions to arrive at two weaker notions of representability. Say that

$$p \textit{ almost agrees} \text{ with } \succeq$$
$$\textit{iff}$$
$$X \succeq Y \Rightarrow p(X) \geq p(Y),$$

and also that

$$p \textit{ partially agrees} \text{ with } \succeq$$
$$\textit{iff}$$
$$X \succ Y \Rightarrow p(X) > p(Y).$$

A comparative belief relation $\succeq$ is *almost representable* if there is a probability function p that almost agrees with $\succeq$. Likewise, $\succeq$ is *partially representable* if there is a probability function p that partially agrees with $\succeq$.

Kraft et al. (1959) show that $\succeq$ is almost representable if and only if it satisfies the *Almost-Cancellation axiom*.

> Almost-Cancellation. If
>
> $$X_1 + \ldots + X_n < Y_1 + \ldots + Y_n$$
>
> and $X_i \succeq Y_i$ for all $i \neq j$, then $X_j \not\succeq Y_j$.

Similarly, Adams (1965) and Fishburn (1969) show that $\succeq$ is partially representable if and only if it satisfies the *Partial-Cancellation axiom*.

> Partial-Cancellation. If $X_1 + \ldots + X_n \leq Y_1 + \ldots + Y_n$ and $X_i \succ Y_i$ for all $i \neq j$, then $X_j \not\succ Y_j$.

It does *not*, to be clear, follow from these two results that $\succeq$ is strongly representable if and only if $\succeq$ satisfies both the Almost and Partial-Cancellation axioms. Satisfying Almost and Partial-Cancellation would simply guarantee that (i) some probability function p almost agrees with $\succeq$, and (ii) some *possibly distinct* probability function q partially agrees with $\succeq$. But strong representability requires that a *single* probability function do *both* types of agreeing. It is an open question what exactly is required for strong representability. (Of course, GST identifies necessary and sufficient conditions for strong representability *given Non-triviality and Non-negativity*. But clearly neither of those conditions is itself necessary for strong representability.)

Almost, partial, and strong representability all place *negative* demands on your comparative beliefs. They require you to *avoid* certain sets of comparative beliefs. Say that

$$p \textit{ endorses} \succeq \textit{ iff } \; p(X) \geq p(Y) \Rightarrow X \succeq Y.$$

For your comparative belief relation $\succeq$ to be almost representable, you must *avoid* having weak comparative beliefs that *no probability function whatsoever* endorses. Likewise, for $\succeq$ to be partially representable, you must avoid having strict comparative beliefs that no probability function endorses. For $\succeq$ to be strongly representable, you must avoid both.

Full probabilistic representability is stronger. It makes *positive* demands as well as negative demands on your comparative beliefs. Full probabilistic representability requires your comparative beliefs to be sufficiently rich and specific that some probability function endorses *exactly* those comparative beliefs. So not only must you *avoid* comparative beliefs that are not endorsed by *any* probability function, but you must positively go in for *all* of the comparative beliefs endorsed by *some* probability function.

Imprecise representability, or *IP-representability*, strikes a balance between these previous types. Like strong representability, IP-representability places negative demands on your comparative beliefs. It requires you to avoid comparative beliefs that no probability function endorses. But like full probabilistic representability, it also makes *positive* demands on your comparative beliefs. It does not go so far as to demand that you go in for *all* of the comparative beliefs endorsed by some probability function. But it *does* say that you must already be more confident in X than Y if *every* probability function that endorses your other comparative beliefs endorses $X \succ Y$ as well. In this way, it requires you to draw out the "probabilistic consequences" of your other comparative beliefs.

Formally, a comparative belief relation $\succeq$ is imprecisely representable if and only if there is a set of probability functions $\mathcal{P}$ that fully agrees with it:

$$\begin{gathered}\mathcal{P} \textit{ fully agrees} \text{ with } \succeq \\ \textit{iff} \\ X \succeq Y \Leftrightarrow p(X) \geq p(Y) \text{ for all } p \in \mathcal{P}.\end{gathered}$$

Rios Insua (1992) and Alon and Lehrer (2014) show that $\succeq$ is IP-representable if and only if it satisfies Reflexivity, Non-negativity, Non-triviality, and the *Generalised Finite-Cancellation axiom*.

> Generalised Finite-Cancellation axiom. If
>
> $$X_1 + \ldots + X_n + \underbrace{A + \ldots + A}_{\text{k times}} = Y_1 + \ldots + Y_n + \underbrace{B + \ldots + B}_{\text{k times}}$$
>
> and $X_i \succeq Y_i$ for all $i \leq n$, then $A \preceq B$.

IP-representability is clearly stronger than strong representability. IP-representability implies strong-representability. But a strongly representable comparative belief relation $\succeq$ might fail to satisfy Reflexivity, Non-negativity, and Non-triviality. No IP-representable $\succeq$ will do so,

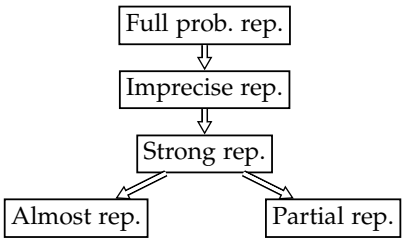


Figure 4: Logical relations between different types of representability

however. So strong-representability does not imply IP-representability. Moreover, Harrison-Trainor, Holliday, and Icard (2016) show that even for non-trivial, non-negative, and reflexive $\succeq$, IP-representability is stronger than strong representability.

To wrap up, let's taxonimise these various types of representability according to their logical strength: see Figure 4.

2.3 *Loose Ends: Infinite Algebras, Conditional Comparative Beliefs, Etc.*

Scott (1964, p. 247) claims that his theorem extends to comparative belief structures $\langle \Omega, \mathcal{F}, \succeq \rangle$ with *infinite* algebras $\mathcal{F}$, by a clever application of the Hahn-Banach Theorem. The proof however remains unpublished. Suppes and Zanotti (1976) also provide necessary and sufficient conditions for a comparative belief relation on an infinite algebra to be fully probabilistically representable. Suppes and Zanotti's axioms, however, do not directly constrain comparative beliefs. Rather, they show that $\succeq$ is probabilistically representable if and only if it is extendable to a *comparative estimation relation* over a larger set; a set containing not just *propositions*—sets of worlds, or equivalently, functions from worlds to 1 (true) or 0 (false), *i.e.*, indicator functions—but to real-valued quantities $\mathcal{Q} : \Omega \to \mathbb{R}$ more generally.

To get the rough idea, consider a travel agent. She might not have a precise estimate of how many travelers will go to Hawaii this year. (Perhaps her evidence is incomplete and ambiguous.) Likewise, she might not have a precise estimate of how many travelers will go to Acapulco. Despite this, she might well estimate that more travelers will go to Hawaii than Acapulco. Or consider the weather. Alayna might not have a precise estimate of how much rain London will receive in June. She might not have a precise estimate of how much rain Canterbury will receive in June. Despite this, she might estimate that London will receive more rain than Canterbury.

Ditto for stock prices, or the number of MPs that different parties will lose or gain in the next election, or any other quantity you might care about. You can have *comparative estimates* regarding those respective quantities—*i.e.*, estimate that one quantity $\mathcal{Q}$ will have a higher/equal/lower value

than another quantity $\mathcal{Q}^*$—without having a unique, precise best estimate for any of them.

Call a relation $\succeq^*$ on a set $\mathcal{F}^*$ of real-valued quantities defined on Ω a *comparative estimation relation* if it is used to model an agent's comparative estimates. And call a comparative estimation relation $\succeq^*$ *qualitatively satisfactory* if it satisfies the following (putative) coherence constraints.

1. $\succeq^*$ is transitive and total.
2. $\Omega \succ^* \varnothing$.
3. $X \succeq^* \varnothing$.
4. $X \succeq^* Y$ iff $X + Z \succeq^* Y + Z$.
5. If $X \succ^* Y$ then for all $W, Z \in \mathcal{F}^*$, there's an $n > 0$ such that

$$\underbrace{X + \ldots + X}_{n \text{ times}} + W \succeq^* \underbrace{Y + \ldots + Y}_{n \text{ times}} + Z.$$

Suppes and Zanotti (1976) show that a comparative belief relation $\succeq$ on an algebra $\mathcal{F}$ of propositions (indicator functions), whether $\mathcal{F}$ is finite or infinite, is fully probabilistically representable if and only if there is a comparative estimation relation $\succeq^*$ on the set $\mathcal{F}^*$ of non-negative integer-valued quantities

$$\mathcal{F}^* = \left\{ \mathcal{Q} \mid \mathcal{Q} : \Omega \to \mathbb{Z}_{\geq 0} \right\},$$

which both (i) extends $\succeq$, and (ii) is qualitatively satisfactory.

This shows that whatever latent structural defect prevents a comparative belief relation $\succeq$ from being fully probabilistically representable rears its head explicitly when you extend the relation. If $\succeq$ has this defect, then when you extend it, so that it encodes not just comparative estimates of truth-values of propositions, but also comparative estimates of the values of non-negative integer-valued quantities more generally, what you end up with—your new, larger comparative estimation relation $\succeq^*$—will violate one of Suppes and Zanotti's putative coherence constraints. And vice versa. If $\succeq$ does *not* have this latent defect, then there is *some* way of extending it that does *not* violate those constraints.[9]

Of course, the representation theorems surveyed here are just the tip of the iceberg. For example, we said that a comparative belief relation $\succeq$ is imprecisely representable if and only if there is a set of probability functions $\mathcal{P}$ that fully agrees with it. But we could explore full agreement (or almost agreement, or partial agreement, or strong agreement) with any

9 Suppes and Zanotti's axiom 5 is an "Archimedean axiom," which guarantees (roughly) that differences in one's best estimates are not "infinitely small." For a non-Archimedean theory of comparative estimation, see Pederson (2014).

number imprecise probability models: Dempster-Shafer belief functions, n-monotone Choquet capacities, coherent lower previsions/expectations, or coherent lower probabilities (*cf.* Walley, 1991, 2000; Augustin, Coolen, de Cooman, and Troffaes, 2014; Troffaes and de Cooman, 2014).

Alternatively, we could focus not on comparative belief relations, but *conditional* comparative belief relations. Hájek (2003)—following Rényi (1955), Jeffreys (1961), de Finetti (1974), and others—argues forcefully that we should treat precise conditional credence as more fundamental than precise unconditional credence. Similarly, we might treat conditional comparative beliefs of the form

$$A \mid B \succeq C \mid D$$

as more fundamental than unconditional comparative beliefs. $A \mid B \succeq C \mid D$ says that the agent in question is at least as confident in A given B as she is in C given D. We can then recover unconditional comparative belief relations from comparative ones by conditioning on the tautology, Ω:

$$A \succeq B \Leftrightarrow A \mid \Omega \succeq B \mid \Omega.$$

Say that a conditional comparative belief relation $\succeq$ on $\mathcal{F}$ is probabilistically representable if there is a conditional probability function that fully agrees with it. More carefully: there is a probability function $p : \mathcal{F} \to \mathbb{R}$ such that for any two propositions $A, B \in \mathcal{F}$, and any two non-null propositions $C, D \in \mathcal{F}$, we have

$$A \mid B \succeq C \mid D \Leftrightarrow \frac{p(A \cap B)}{p(B)} \geq \frac{p(C \cap D)}{p(D)}.$$

A proposition X is *non-null* just in case it is not just as likely as the contradiction, *i.e.*,

$$X \mid \Omega \not\approx \varnothing \mid \Omega.$$

Now we can ask: when are conditional comparative belief relations probabilistically representable? Domotor (1969) extends the results of Scott (1964) to provide necessary and sufficient conditions for probabilistic representability when $\succeq$ is defined on a finite algebra $\mathcal{F}$. Suppes and Zanotti (1982) extend the results of Suppes and Zanotti (1976) to provide necessary and sufficient conditions in the general case (whether or not $\mathcal{F}$ is finite). See Suppes (1994) for additional detail.

With a basic understanding of representation theorems and their mechanics in hand, we can now turn our attention to the central question of this chapter: do comparative beliefs and representation theorems figure into the best explanation of what precise credences are? If so, how? What do these accounts of credence look like? And how do they stand up to the criticisms of Hájek (2009), Meacham and Weisberg (2011), and Titelbaum (2015)?

3 THE COMPARATIVE BELIEF-CREDENCE CONNECTION

To a first approximation, an agent's credence function measures how confident she can be said to be in each proposition. If $c(X) = 1$, then she is maximally confident that X is true, *i.e.*, 100% confident. If $c(X) = 0$, then she is minimally confident that X is true, *i.e.*, 0% confident. If $c(X) = 2/3$, then she is more confident than not that X is true, but not quite fully confident.

But what does this really *mean*? What does it mean to say that an agent is 100%, or 80%, or 23.9556334% confident in a proposition?

We might have similar questions for *imprecise* Bayesians. Imprecise Bayesians model rational agents' opinions not with a single credence function c, but with a *set* of credence functions $\mathcal{C}$. Sets of credence functions are called *imprecise credal states* (see Mahtani's entry in this volume).[10] To a first approximation, imprecise credal states also measure how confident agents can be said to be in various propositions. But they allow for a strictly greater range of opinions than precise credal states. For example, if $c(X) = 1$ for all c in $\mathcal{C}$, then our agent is 100% confident that X is true. If, however, $0.6 \leq c(X) \leq 0.9$ for all c in $\mathcal{C}$, and nothing stronger, then she is at least 60% confident and at most 90% confident that X is true. But she has no precise level of confidence for X. Precise credence functions allow for the first sort of opinion, but not the second.

But again, what exactly does this mean? What does it mean to say that an agent is at least 60% confident and at most 90% confident in a proposition?

The history of Bayesianism is chock-full of different accounts of credence that aim to answer this question. Very roughly, we can lump them into three groups: *measurement-theoretic* accounts, *decision-theoretic* accounts, and *interpretivist* accounts. Before exploring the differences between these various accounts, it is worth emphasising one similarity. They all treat 'credence function' or 'credal state' in roughly the way Carnap treated theoretical terms more generally. They carve out some theoretical role (or set of roles) $\mathcal{R}$ as constitutive of what it is for a real-valued function, c, or a set of such functions, $\mathcal{C}$, to count as "your credal state." The better c (or $\mathcal{C}$) plays role $\mathcal{R}$, the more eligible it is as a "credal state candidate." What these accounts disagree on is *what the relevant theoretical role $\mathcal{R}$ is*.

10 Precise credal states are special cases of imprecise credal states, on the imprecise Bayesian view. Formally, $\mathcal{C}$ is precise just in case $\mathcal{C} = \{c\}$ for some credence function c.

3.1 *Measurement-Theoretic Account of Credence*

Measurement-theoretic accounts, like those of Koopman (1940a, 1940b), Good (1950), and Krantz et al. (1971), treat credal states as *mere numerical measurement systems* for comparative beliefs (or more generally, for some underlying structure of comparative and qualitative opinions). Compare: numerical measurement systems for length, mass, velocity, etc., allow engineers, scientists and the like to measure certain parts of the system of interest, perform numerical calculations, and draw inferences about other parts of the system. Imagine, for example, measuring the length of two pieces of wood arranged at a right angle, and using the Pythagorean theorem to infer how long the diagonal must be.

Similarly, on the measurement-theoretic view, *credal states* are mere numerical measurement systems. They allow you to measure certain parts of an agent's system of comparative beliefs, perform numerical calculations, and draw conclusions about what other comparative beliefs she must have (or must not have). Imagine, for example, that you elicit a sufficient number of an agent's comparative beliefs $\succeq$ to be quite confident that (i) she satisfies Scott's axioms, so that $\succeq$ fully agrees with some probability function c, and further that (ii) $c(X) = 0.3$, $c(Y) = 0.4$, and $c(X \cap Z) = c(Y \cap Z) = 0$. Given these measurements, you can use the probability axioms to calculate that $c(X \cup Z) \leq c(Y \cup Z)$. Since c fully agrees with $\succeq$, you can infer that $X \cup Z \preceq Y \cup Z$. (See Section 5 for a more complete introduction to the measurement-theoretic view.)

The upshot: just like measurement systems for physical quantities (length, etc.), credal states allow you to represent comparative beliefs in an elegant, easy-to-use, numerical fashion. And modeling comparative (and qualitative) beliefs with numbers is *useful*. Numerical measurement systems are designed specifically to reflect important structural features of the underlying target system, so that you can use them to straightforwardly extract information about one part of the system from information about other parts.

Where does this leave us? The measurement-theoretic view takes a particular stand on the nature of the theoretical role $\mathcal{R}$ that a function c (or set of functions $\mathcal{C}$) must play in order to count as "your credal state." More specifically, c (or $\mathcal{C}$) must fully agree (or almost agree, or partially agree, or strongly agree) with the agent's comparative beliefs, $\succeq$, in the way required to count as a numerical measurement system for $\succeq$. The better c (or $\mathcal{C}$) plays this role $\mathcal{R}$, the more eligible it is as a credal state candidate. Equally good measurement systems are equally eligible credal state candidates.

3.2 *Decision-Theoretic Account of Credence*

Decision-theoretic accounts of credence, like those of Ramsey (1931), de Finetti (1931, 1964), and Walley (1991), carve out a rather different theoretical role for credal states. Credal states, on these views, encode an agent's fair buying and selling prices. An agent's *fair buying price* for a gamble $\mathcal{G}$ is, roughly, the largest amount that she could pay for $\mathcal{G}$ while still leaving herself (in her own view) in at least as good a position as the status quo. An agent's *fair selling price* for a gamble $\mathcal{G}$ is, roughly, the smallest amount that someone else would have to pay her in exchange for $\mathcal{G}$ in order to leave herself (in her own view) in at least as good a position as the status quo.

To illustrate, imagine that you have an urn. The urn contains 10 balls. Each ball is either red or black. There are at least 3 black balls, and at most 7 black balls. But you have no absolutely no idea whether the urn contains 3, 4, 5, 6, or 7 black balls.

Let $\mathcal{G}$ be the gamble that pays out £10 if a random draw from the urn yields a black, and £0 otherwise. Given what you know about the contents of the urn, you would likely judge that paying a measly £1 for $\mathcal{G}$ is a good deal. Maybe £2 is a good deal too. But let's imagine that £3 is your limit. Paying any more than £3 would leave you in a situation where you are no longer, in your own view, determinately doing at least as well as the status quo. Then your fair buying price for $\mathcal{G}$ is 3. More carefully, your fair buying price for $\mathcal{G}$ is 3 iff you weakly prefer paying 3 and receiving $\mathcal{G}$ to the status quo, but not so for any amount higher than 3.

Similarly, suppose that a friend wants to buy $\mathcal{G}$ from you. They will pay you some initial amount. Then you will pay them £10 if the draw comes up black and £0 otherwise. Given what you know about the urn, you would likely judge that selling $\mathcal{G}$ to your friend for £9 is a good deal (for you, anyway). Maybe £8 is a good deal too. But let's imagine that £7 is your limit. If they offer you any less than £7, then you would be left in a position where you are no longer, in your own view, determinately doing at least as well as the status quo. Then your fair selling price for $\mathcal{G}$ is 7. More carefully, your fair selling price for $\mathcal{G}$ is 7 iff you weakly prefer receiving 7 and selling $\mathcal{G}$ to the status quo, but not so for any amount lower than 7.

On the decision-theoretic view, the principal theoretical role of an agent's credal state is to encode her fair buying and selling prices. A set $\mathcal{E}$ of real-valued functions e counts as "your credal state" just in case its *lower and upper envelope* for gambles $\mathcal{G}$,

$$\underline{\mathcal{E}}[\mathcal{G}] = \inf\left\{e(\mathcal{G}) \mid e \in \mathcal{E}\right\},$$
$$\overline{\mathcal{E}}[\mathcal{G}] = \sup\left\{e(\mathcal{G}) \mid e \in \mathcal{E}\right\},$$

are equal to your fair buying and selling prices for $\mathcal{G}$, $\mathcal{B}(\mathcal{G})$ and $\mathcal{S}(\mathcal{G})$, respectively, *i.e.*, $\underline{\mathcal{E}}[\mathcal{G}] = \mathcal{B}(\mathcal{G})$ and $\overline{\mathcal{E}}[\mathcal{G}] = \mathcal{S}(\mathcal{G})$. (Treat a single real-valued function e as the singleton $\mathcal{E} = \{e\}$.) The better $\mathcal{E}$ plays this role $\mathcal{R}$—the closer its lower and upper envelopes are to your fair buying and selling prices—the more eligible it is as a credal state candidate.

Your credal state only captures information about your beliefs, on this view, insofar as they are reflected in your fair buying and selling prices. For any proposition $X \in \mathcal{F}$, let $\mathcal{G}_X$ be the unit gamble on X, *i.e.*, the gamble that pays out £1 if X and £0 otherwise. Your lower and upper "previsions" for $\mathcal{G}_X$, $\underline{\mathcal{E}}[\mathcal{G}_X]$ and $\overline{\mathcal{E}}[\mathcal{G}_X]$ (*i.e.*, the value of the lower and upper envelopes of $\mathcal{E}$ at $\mathcal{G}_X$), encode your fair buying and selling prices for $\mathcal{G}_X$. If you are willing to pay something near £1 for a unit gamble on X ($\underline{\mathcal{E}}[\mathcal{G}_X] \approx 1$), then *for the purposes of decision-making* you are quite confident in X. If you would be happy to sell a unit gamble on X to a friend for mere pennies ($\overline{\mathcal{E}}[\mathcal{G}_X] \approx 0$), then for the purposes of decision-making you have extremely low confidence in X. If you would only buy a unit gamble on X for next to nothing ($\underline{\mathcal{E}}[\mathcal{G}_X] \approx 0$), and would only sell a unit gamble on for close to its maximum payout ($\underline{\mathcal{E}}[\mathcal{G}_X] \approx 1$), then for the purposes of decision-making you have no idea whether X is true. Your opinions are rather imprecise.

The decision-theoretic view comes in many flavours—one for each way of thinking about the preferences that determine your fair buying and selling prices. On a flat-footed behaviourist view, $\mathcal{B}(\mathcal{G})$ is your fair buying price for $\mathcal{G}$ just in case you *actually* buy $\mathcal{G}$ for $\mathcal{B}(\mathcal{G})$, and *actually* refuse to buy $\mathcal{G}$ for any higher price (or perhaps do so a sufficiently high proportion of the time). On a more sophisticated behaviourist view, $\mathcal{B}(\mathcal{G})$ is your fair buying price for $\mathcal{G}$ just in case you are *disposed* to buy $\mathcal{G}$ for $\mathcal{B}(\mathcal{G})$, and *disposed* to refuse to buy $\mathcal{G}$ for any higher price. Alternatively, we might reject behaviourism in its various guises, and say that the preferences that fix your fair buying/selling prices are irreducibly *evaluative* attitudes.

But where do comparative beliefs enter the picture? It may not appear that comparative beliefs play an especially important role in explicating the concept of credence on the decision-theoretic view. After all, on this view, an agent's credal state encodes her fair buying and selling prices. And fair buying and selling prices are fixed by one's *preferences*, not their *comparative beliefs*. Even on Savage's view, where comparative belief reduces to preference, different fragments of an agent's preference relation fix her fair buying/selling prices and comparative beliefs, respectively. Nonetheless, *rational* comparative beliefs and fair buying/selling prices hang together in a certain way (Section 6). So comparative beliefs (and representation theorems) will be important for answering the normative question, even on the decision-theoretic view.

Also worth noting: if an agent has a precise credal state $\mathcal{E} = \{e\}$, then

$$\underline{\mathcal{E}}[\mathcal{G}] = \overline{\mathcal{E}}[\mathcal{G}]$$

for all gambles $\mathcal{G}$. That is, her fair buying prices *just are* her fair selling prices. The maximum amount she is willing to pay for $\mathcal{G}$ is *precisely* the minimum amount she is willing to accept in exchange for selling $\mathcal{G}$. Agents with genuinely imprecise credal states (non-singleton $\mathcal{E}$), in contrast, may well think that buying is worthwhile only at very low prices, and selling is worthwhile only at very high prices. Imprecise Bayesians typically see this as the *proper* (or at least a *permissible*) type of evaluative attitude to bear in decision contexts where evidence is unspecific or ambiguous.

One final note: measurement-theoretic and decision-theoretic accounts of credence can be difficult to distinguish in practice. Consider a proponent of the measurement-theoretic account, such as Savage, who treats comparative belief as reducible to preference (Savage, 1954, Section 3.2). You judge that $X \succeq Y$ iff whenever you prefer one outcome to another, you also prefer getting the better outcome if X than if Y. Then certain types of measurement systems for comparative belief—*viz.*, sets of probability functions—encode fair buying and selling prices (see Section 6). Whence the difference, then, between this sort of measurement-theoretic account of credence, and a decision-theoretic account?

The difference is this. On the measurement-theoretic view, *any* numerical measurement system for $\succeq$ does the work necessary to count as "your credal state"—not just ones that encode your fair buying and selling prices. Likewise, on the decision-theoretic view, *any* numerical system that encodes your buying and selling prices counts as "your credal state." But some of those systems (*viz.* upper and lower previsions) carry too little information to determine a numerical representation of your preference relation (Walley, 2000, Section 6).

Shorter: even though *some* numerical systems do both jobs (measurement-theoretic and decision-theoretic), it is *possible* to do one without doing the other. So the two accounts make different predictions about which functions (sets of functions) count as "eligible credal state candidates."

3.3 *Interpretivist Account of Credence*

Our final account of credence is the *interpretivist account*, of the sort espoused by Lewis (1974) and Maher (1993). According to *preference-based* interpretivist accounts, like Patrick Maher's, "an attribution of probabilities and utilities is correct just in case it is part of an overall interpretation of the person's preferences that makes sufficiently good sense of them and better sense than any competing interpretation does" (Maher, 1993, p. 12). And according to Maher, if some probabilistically coherent credence function c and cardinal utility function u jointly agree with an agent $\mathcal{A}$'s preferences, in the following sense:

$$\mathcal{A} \text{ weakly prefers } \alpha \text{ to } \beta$$
$$\textit{iff}$$
$$E_c[\alpha] \geq E_c[\beta]$$

(where $E_c[\alpha]$ and $E_c[\beta]$ are the expected utilities of acts α and β relative to c and u, respectively), then c and u *perfectly rationalise* or *make sense of* that agent's preferences.

On Maher's view, both credence functions and utility functions earn their theoretical keep by rationalising *preferences*. If c and u rationalise your preferences better than any competing c^* and u^*, then c plays the appropriate theoretical role to count as "your credal state," and u plays the appropriate theoretical role to count as "your utility function." This presupposes the thesis of the primacy of practical reason. Whether or not c rationalises your comparative and qualitative beliefs, understood as irreducibly *doxastic* attitudes, is neither here nor there. What makes c "your credence function" is the fact that it helps to rationalise your preferences.

But we can distinguish another brand of interpretivism: *epistemic interpretivism*. This is a new account of credence. So we will spend a bit of time developing it.

According to epistemic interpretivism, credal states are *assignments of truth-value estimates* (or sets of such assignments) that rationalise one's *comparative beliefs* (or more generally, her comparative and qualitative opinions), understood as irreducibly doxastic attitudes. A function $c : \mathcal{F} \to \mathbb{R}$ (or set $\mathcal{C}$) counts as "your credal state" just in case it encodes truth-value estimates (or constraints on such estimates) that best rationalise or make sense of your comparative beliefs.

Spelling out epistemic interpretivism requires two things: (i) saying something about what truth-value estimates are, and (ii) explaining what it means for truth-value estimates to best rationalise a set of comparative beliefs.

Estimates are familiar enough. For example, an analyst's best estimate of Tesla's stock price 1 years hence might be $425. Your best estimate of the number of bananas in a randomly selected bunch might be 5.7. And so on. In each of these examples, there is the *agent* doing the estimating, there is the *quantity* being estimated, and there is the *estimate* of that quantity. For the purposes of spelling out epistemic interpretivism, it is the last of these that matters most.

Estimates are numbers. But not all numbers are estimates. For example, the numbers in the expression

$$1,000,000 > 2$$

are not estimates. What sorts of numbers are estimates then? Plausibly, they are numbers that are subject to a certain standard of evaluation. A

number is an estimate in a context iff it is *evaluated qua estimate* in that context. In typical contexts of evaluation, numbers like 2 in expressions like the above are not estimates because they are not evaluated qua estimates. There is no quantity that it would be better or worse for 2 to be close to. It is no better or worse for being close to the actual price of stock X, or the actual dosage of drug Y, etc. In contrast, the number at the bottom of a contractor's quote—a paradigm of an estimate—*is* evaluated qua estimate. It is quite bad, for example, if it is £20,000 off the actual price of the job.

What exactly is it to evaluate a number *qua* estimate? We will not provide a full answer here. But we can say something informative.

The type of phenomenon under consideration—evaluating an entity $\mathcal{E}$ *qua* $\mathcal{X}$—is a common one. You might be brilliant *qua* scientist, mediocre *qua* mentor, and terrible *qua* conversationalist. The reason seems to be this: scientists, mentors and conversationalists all perform characteristic functions. And you can perform some functions well while performing others poorly. Microbiologists, for example, carefully dissect tissue samples, meticulously document their experiments, write up academic papers, communicate their results at conferences, etc. Conversationalists, on the other hand, ask engaging questions, are familiar with current events, and so on. You might well dissect tissue samples masterfully, but have no idea what the news of the day is.

This suggests the following. Evaluating an entity $\mathcal{E}$ *in some capacity* $\mathcal{X}$, or *qua* $\mathcal{X}$, is a matter of evaluating $\mathcal{E}$ on the basis of how well it performs the characteristic functions $\mathcal{F}_1, \ldots, \mathcal{F}_n$ associated with $\mathcal{X}$. What to say about estimates in particular then? What characteristic functions do they serve, for example, in scientific inquiry, engineering, finance, etc.? Whatever the full answer is, the following seems non-negotiable: an estimate of quantity $\mathcal{Q}$ serves the function of approximating the true value of $\mathcal{Q}$. So *ceteris paribus* it is better the closer it is to the true value of $\mathcal{Q}$.

Note that, on the present account, for a number to count as an estimate in a context, there must be an *evaluator* in that context; an agent evaluating the number *qua* estimate. (This need not require having the concept *estimate*, or anything of the sort. Evaluating a number *qua* estimate might be fairly cognitively undemanding task.) But there need be no *estimator*; no agent producing the estimate; no agent explicitly judging that this is the best estimate of that, etc.

Thermometers provide estimates of temperature. Geiger counters provide estimates of radiation. Ditto for other measurement devices. In each of these cases, there is an estimate (38°C, 0.10mSv, etc.), but no estimator; no agent doing the estimating. Similarly, a tree's rings provide an estimate of its age. Your parents' income provides an estimate of your income. Again, estimates without estimators. And estimates, of course, do not need to be *good*. The number of tea leaves concentrated in one part of your cup

provides a (thoroughly unreliable) estimate of the number of fortunate events in your future. Once more: estimate, but no estimator.

The upshot: we can talk of estimates doing this or that—for example, rationalising a set of comparative beliefs—even if those estimates do not "belong" to anyone. Estimates without estimators.

Back to our original question: what are truth-value estimates? We have made some progress in saying what estimates are more generally. Now, following de Finetti and Jeffrey, treat a proposition X as an "indicator variable" that takes the value 1 at worlds where X is true, and 0 where X is false. Truth-value estimates, then, are simply estimates of the value, 0 or 1, that the proposition takes at the actual world.

To finish spelling out the epistemic interpretivist account of credence, we need to explain what it means for truth-value estimates to "best rationalise" a set of comparative beliefs. To get a feel for how this might work, consider an example. Grandma relies on folklore methods for predicting the weather. She feels things in her bones, observes the behaviour of the cows in the pasture, etc. You are not sure whether the weather-related opinions that Grandma comes to on this basis make much sense or not. But then you open your weather app. Lo and behold, you find a bunch of estimates—probabilities for sun, clouds, rain, etc., estimates of rainfall amount, hour-by-hour temperature estimates, etc.—that recommend thinking precisely what Grandma thinks. For example, Grandma thinks it is likelier than not to rain this evening. And the weather app recommends thinking that too. It specifies a greater than 50% probability of rain. (We will explore a few different accounts of *recommendation* shortly.)

The weather app's estimates recommend having Grandma's opinions. And these estimates are themselves eminently rational. In virtue of this, they rationalise or make sense of those opinions.

Note, however, that the weather app itself is not essential to this story. Estimates do not need an estimator. If there *exists* some rational set of estimates that recommend Grandma's opinions, then whether or not any weather app actually spits those estimates out, or any meteorologist actually judges those estimates to be best—or indeed whether any artificial or human system is in the business of explicitly estimating quantities at all—Grandma's opinions are nonetheless *rationalisable*. The rational estimates that recommend her opinions provide that rationale.

Before saying something more general about when a set of truth-value estimates best rationalises a set of comparative beliefs, we should key in on two important features of our example. The first is the *strength* of the recommendation in question. The second is the *quality* of that recommendation.

We stipulated that the weather app's estimates recommend having Grandma's opinions. This makes it seem as though recommendation is

an on-off matter. But recommendations plausibly come in degrees. You can recommend a trip to the Alps a little more strongly than a trip to Tahoe, but *much* more strongly than a trip to Cudahy, Wisconsin. In our example, the weather app's estimates most strongly recommend thinking *precisely what Grandma thinks*. We might have stipulated, however, that they recommend a similar but distinct state of opinion most strongly, and recommend Grandma's state of opinion a little less strongly. In that case, the weather app's estimates provide a fairly strong, but not maximally strong rationale for Grandma's state of opinion.

In addition to the *strength* of a recommendation, we can consider the *quality* of that recommendation. We stipulated that the weather app's estimates are eminently rational. But our weather app could have been a bit glitchy and delivered mildly irrational estimates (ones that violate the probability axioms, perhaps, but not by much). Those estimates might still recommend thinking what Grandma thinks just as strongly. But in virtue of their mild irrationality, they provide a slight lower *quality* rationale for Grandma's state of opinion.

The distinction between strength and quality is important. If Grandma's state of opinion is epistemically defective, it may turn out that no estimates unreservedly recommend it, *i.e.*, recommend it at least as strongly as any other state of opinion. Every set of estimates might recommend some other state of opinion more strongly. Nonetheless, some sets of estimates might recommend Grandma's state of opinion more strongly than others. And amongst the sets of estimates that recommend it as strongly as possible (at least as strongly as any other set of estimates), some might provide a higher *quality* recommendation than others. The extent to which a set of estimates rationalises or makes sense of a state of opinions depends on both strength and quality. To provide the *best possible rationale* for Grandma's state of opinion, for example, a set of estimates must (i) recommend that state as *strongly* as possible, and (ii) must provide the highest *quality* recommendation from amongst the sets of estimates that satisfy (i).

Let's take stock. According to epistemic interpretivism, a function $c : \mathcal{F} \to \mathbb{R}$ (or set of functions $\mathcal{C}$) counts as "your credal state" just in case it encodes truth-value estimates (or constraints on such estimates) that best rationalise or make sense of your comparative beliefs. We gave a brief account of *estimatehood* to fill this out a bit. And we quickly unpacked what it means for c (or $\mathcal{C}$) to "best rationalise" your comparative beliefs, $\succeq$. To *best* rationalise $\succeq$, c should provide *at least as strong a rationale* for $\succeq$ as any other set of truth-value estimates c^*. And on the picture sketched above, c provides a rationale for $\succeq$ by *recommending* $\succeq$. So for c to count as "your credal state," no other c^* can recommend $\succeq$ more strongly than c. Moreover, amongst the truth-value estimates that provide a maximally strong rationale for $\succeq$ (recommend it as strongly as possible), c should

provide at least as high *quality* a rationale as any other c^*. On the picture sketched above, the quality of c's rationale depends on how close c itself is to rational. So for c to count as "your credal state," no other c^* that recommends $\succeq$ as strongly as possible should be *more rational* than c. Pulling this all together, c (or C) counts as "your credal state" just in case it encodes truth-value estimates (or constraints on such estimates) that recommend your comparative beliefs as strongly as possible, and are as rational as possible whilst doing so. See Figure 5.

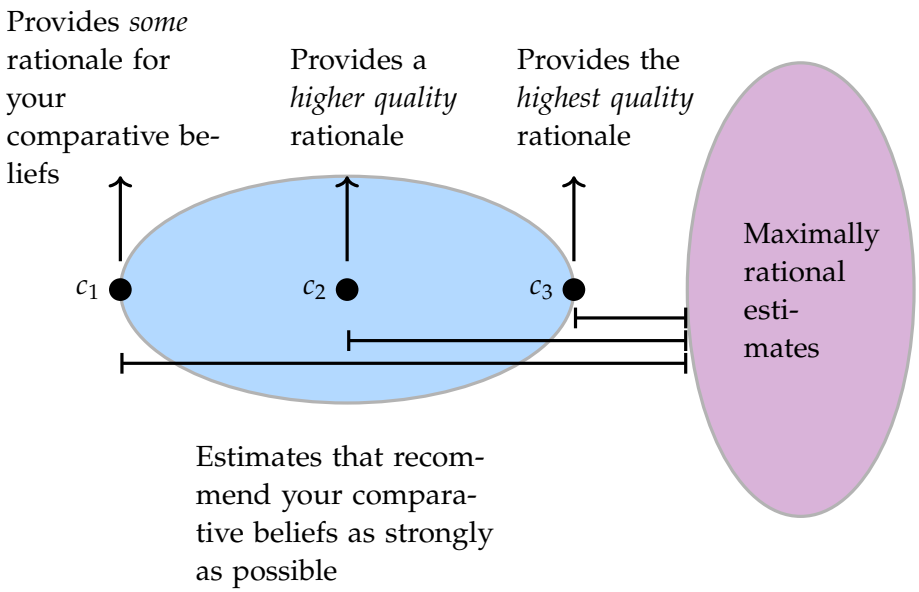


Figure 5: More rational estimates provide higher quality rationales.

The big lingering question is this: when exactly does a set of truth-value estimates *recommend* a certain set of comparative beliefs more or less strongly? There are a number of ways one could spell this out. We will not defend a particular account of recommendation here. But here are three options.

> **Metaphysical Account.** The truth-value estimates given by $c : \mathcal{F} \to \mathbb{R}$ recommend $\succeq$ to degree k iff it is *metaphysically necessary* that any agent who explicitly judges $c(X)$ to be the best truth-value estimate for X, for all $X \in \mathcal{F}$, has comparative beliefs $\succeq_c$ and $\mathcal{D}(\succeq_c, \succeq) = 1/k$, where $\mathcal{D}$ is some reasonable measure of distance between comparative belief relations.

On the metaphysical account, judging $c : \mathcal{F} \to \mathbb{R}$ to encode the best truth-value estimates for propositions in $\mathcal{F}$ *entails* having certain comparative beliefs $\succeq_c$. Since having comparative beliefs $\succeq_c$ is part and parcel of judging c best, c recommends $\succeq_c$ as strongly as possible. And c recommends

other comparative beliefs, $\succeq$, less strongly the further away they are from $\succeq_c$. See Deza and Deza (2009) and Fitelson and McCarthy (2015) for more information on measures of distance between comparative belief relations.

Our next account says that while judging c to encode the best truth-value estimates may not *entail* that you have some set of comparative beliefs or other, it nevertheless *rationally requires* you to have those beliefs. And we can use this fact to say what it is for a set of truth-value estimates to *recommend* comparative belief relations to different degrees.

> **Normative Account.** The truth-value estimates given by $c : \mathcal{F} \to \mathbb{R}$ recommend $\succeq$ to degree k iff it is *rationally required* that any agent who explicitly judges $c(X)$ to be the best truth-value estimate for X, for all $X \in \mathcal{F}$, has comparative beliefs $\succeq_c$ and $\mathcal{D}(\succeq_c, \succeq) = 1/k$, where $\mathcal{D}$ is some reasonable measure of distance between comparative belief relations.

A proponent of the normative account might treat the principles of rationality that generate the relevant requirement as properly basic components of her epistemology. Alternatively, she might provide a teleological explanation of why those principles have the normative force that they do by appealing to facts about *epistemic value* or *utility*. One final account of recommendation—the *epistemic utility account*—explains recommendation more directly in terms of epistemic value/utility facts. Informally, the epistemic utility account says that c recommends $\succeq$ to degree k just in case the most rational way of adding estimates of the *value* of comparative beliefs to the stock of truth-value estimates encoded by c involves estimating $\succeq$ to have epistemic utility k.

Let's make this a little more precise. An assignment of truth-value estimates $c : \mathcal{F} \to \mathbb{R}$ (or set $\mathcal{C}$) maps a very specific kind of measurable quantity—*propositions* or *indicator functions*—to estimates. Let $\mathbb{Q}$ be the set of all measurable quantities $\mathcal{Q} : \Omega \to \mathbb{R}$. An assignment $est : \mathbb{Q} \to \mathbb{R}$ of estimates to measurable quantities extends c just in case $c(X) = est(X)$ for all $X \in \mathcal{F}$.

To make sense of something being closer or further from rational, we need two things: an epistemic utility function $\mathcal{U}$ and laws of preference $\mathcal{L}$.

First let's talk about $\mathcal{U}$. For any assignment of truth-value estimates c, $\mathcal{U}(c, w)$ measures how epistemically valuable c is at world w. Whatever properties make truth-value estimates epistemically valuable at a world, $\mathcal{U}(c, w)$ captures the extent to which c has a good balance of these properties at w. Likewise, $\mathcal{U}(\succeq, w)$ measures how epistemically valuable comparative beliefs $\succeq$ are at world w. For a philosophically rich discussion of how to measure the epistemic value of estimates, see Joyce (2009) and Pettigrew (2016).

Laws of preference $\mathcal{L}$ are familiar from decision theory. In conjunction with $\mathcal{U}$, they specify rationally permissible ways of structuring one's preferences over options. For example, the law of dominance says that if one option o is *guaranteed* to have higher utility than another option o^*, then you ought to prefer o to o^*. Likewise, the law of (first-order) stochastic dominance says: if for any possible utility value x, o is guaranteed to have greater chance than o^* of having higher-than-x utility, then you ought to prefer o to o^*. And so on.

Let $\mathcal{T}$ be the set of rational truth-value estimates, relative to $\mathcal{U}$ and $\mathcal{L}$, *i.e.*, the set of c are not dispreferred to some other c^*. Let $\mathcal{E}$ be the set of rational estimates more generally relative to $\mathcal{U}$ and $\mathcal{L}$, *i.e.*, the set of *est* that are not dispreferred to some other *est**.

Say that *est* is the *maximally rational extension* of c to $\mathbb{Q}$ iff (i) *est* extends c to $\mathbb{Q}$, and (ii) *est* is closer to rational (*i.e.*, closer to $\mathcal{E}$) than any other *est** that extends c to $\mathbb{Q}$.

We can now state the epistemic utility account more precisely.

> **Epistemic Utility Account.** The truth-value estimates given by $c : \mathcal{F} \to \mathbb{R}$ recommends $\succeq$ to degree k iff the maximally rational extension of c to $\mathbb{Q}$, est_c, is such that such that $est_c(\mathcal{U}(\succeq)) = k$.

The basic thought here is that while c might not *directly* encode estimates of quantities other than truth-values, it nonetheless takes a stand on how to estimate those quantities. It encodes such estimates *indirectly*. There is some most rational way of adding estimates of other measurable quantities $\mathcal{Q}$ to the stock of truth-value estimates encoded by c. These estimates, $est_c(\mathcal{Q})$, are the best estimates of those quantities, from c's perspective. So, in effect, the epistemic utility account says that c recommends $\succeq$ to degree k just in case it indirectly estimates $\succeq$ to be epistemically valuable to degree k.

There are no doubt myriad unanswered questions about each of these accounts of recommendation. It is not our purpose to provide a full defense of any particular account. Just note that you can choose your favourite (or one not on the list) and slot it into our official version of epistemic interpretivism.

> **Epistemic Interpretivism.** A function c (or set $\mathcal{C}$) counts as "your credal state" iff it best rationalises your comparative beliefs $\succeq$. Moreover, c (or set $\mathcal{C}$) best rationalises $\succeq$ iff (i) it recommends $\succeq$ as strongly as possible, so that no other c^* (or set $\mathcal{C}^*$) recommends $\succeq$ to a higher degree, and (ii) c is itself closer to rational (closer to $\mathcal{T}$) than any other c^* that recommends $\succeq$ as strongly as possible, and so provides the highest quality recommendation possible.

Even setting aside questions about how to understand *recommendation*, there are various lingering questions about epistemic interpretivism. For example, one might wonder what makes comparative beliefs more or less epistemically valuable at a world, or how to measure such value. See Fitelson and McCarthy (2015) for an investigation of "additive" epistemic utility measures for comparative belief. One might also wonder what makes one set of estimates closer to rational than another. For a nuanced discussion, see Staffel (2018). We will not address these questions here. But we will evaluate epistemic interpretivism in a bit more depth in Section 7.

4 CHALLENGES TO THE RELEVANCE OF REPRESENTATION THEOREMS

We now have a number of accounts of credence on the table, however briefly sketched. These accounts purport to tell us what it means to say that an agent is x% confident in a proposition (if she has precise credences), or between y% and z% confident (if she has imprecise credences).

Proponents of these accounts use them to answer some important questions. For example, when exactly *is* there a real-valued function c (or set $\mathcal{C}$) that plays the relevant theoretical role $\mathcal{R}$ well enough to count as "your credal state"? Following Meacham and Weisberg (2011), we will call this the *characterisation question*. And why should we expect *rational* agents to have *probabilistically coherent* credences? We will call this the *normative question*.

In answering these questions, proponents typically invoke coherence constraints (on either preference or comparative belief) and representation theorems. Hájek (2009), Meacham and Weisberg (2011), and Titelbaum (2015) challenge any such approach. Whatever account of credence you adopt, they argue, there is no plausible representation-theorem-centric narrative that could answer these questions. Their objections are many. We will focus on a few central ones.

Hájek, Meacham and Weisberg, and Titelbaum all imagine that the "basic representation theorem argument" goes as follows.

1. *Coherence Constraints*. Any rational agent's comparative belief relation $\succeq$ satisfies coherence constraints ϕ.

2. *Representation Theorem*. Relation $\succeq$ satisfies constraints ϕ if and only if $\succeq$ fully agrees (or almost agrees, or partially agrees, or strongly agrees) with some probability function c (or set of probability functions $\mathcal{C}$).

C. *Probabilism*. Any rational agent has probabilistic credences (either precise credences given by c, or imprecise credences given by $\mathcal{C}$).

If successful, this argument would at least partially answer both the characterisation and normative question at once. When is there a credence function c, or a set of such functions $\mathcal{C}$, that plays the relevant theoretical role $\mathcal{R}$ well enough to count as your credal state? Whenever your comparative beliefs satisfy coherence constraints ϕ! Satisfying ϕ is a sufficient condition for having credences. And why should we expect *rational* agents to have *probabilistically coherent* credences? Because the coherence constraints ϕ are rationally mandatory. And any agent who satisfies ϕ not only has credences, but *probabilistically coherent* credences.

But this argument is *not* successful as it stands. As Eriksson and Hájek (2007), Hájek (2009), Meacham and Weisberg (2011), and Titelbaum (2015) emphasise, it does not follow from the mere fact that some probabilistically coherent credence function *fully agrees* with her comparative beliefs that she in fact *has* probabilistic credences. So the argument is invalid. Hájek puts the point as follows (*cf.* also Meacham and Weisberg, 2011, p. 14, and Titelbaum, 2015, p. 274):

> the mere *possibility* of representing you one way or another might have less force than we want; your acting *as if* the representation is true of you does not make it true of you. To make this concern vivid, suppose that I represent your preferences with *Voodooism*. My voodoo theory says that there are warring voodoo spirits inside you. When you prefer A to B, then there are more A-favouring spirits inside you than B-favouring spirits [...] I then 'prove' Voodooism: if your preferences obey the usual rationality axioms, then there exists a Voodoo representation of you. That is, you act *as if* there are warring voodoo spirits inside you in conformity with Voodooism. Conclusion: rationality requires you to have warring Voodoo spirits in you. Not a happy result. (Hájek, 2009, p. 238)

The same thing, these objectors claim, can be said about the representation theorem argument for probabilism. Just because your preferences can be *represented* as the end product of a vigorous war between the voodoo spirits inside you does not imply that you *in fact* have such spirits inside you. Similarly, just because your comparative beliefs can be *represented* as arising from precise credences c (or imprecise credences $\mathcal{C}$) does not imply that you *in fact* have such credences.

This line of criticism is not particularly concerning. The reason: no Bayesians put forward this basic "representation theorem argument." Koopman, Savage, Joyce, etc.; they all presuppose *some* account of credence or other. For example, Krantz et al. presuppose a measurement-theoretic account of credence.

> we inquire into conditions under which an ordering $\succeq$ of $\mathscr{E}$ has an order-preserving function P that satisfies Definition 2. Obviously, the ordering is to be interpreted empirically as meaning "qualitatively at least as probable as." Put another way, we shall attempt to treat the assignment of probabilities to events as a measurement problem of the same fundamental character as the measurement of, e.g., mass or momentum. (Krantz et al., 1971, pp. 199–202)

The upshot: any faithful reconstruction of the "representation theorem argument" really ought to feature an account of credence explicitly as a premise. The simple argument under attack here fails this basic test.

Of course, objectors do not focus exclusively on this simple version of the representation theorem argument. Hájek, Meacham and Weisberg, and Titelbaum all consider more sophisticated versions as well. A fairly general, more charitable way of understanding what fans of representation theorems are up to is this. Firstly, to shed some light on the characterisation question, they establish a "Bridge Theorem" which shows that the function c, or set of functions $\mathcal{C}$, outputted by their favourite representation theorem is fit to play the theoretical role $\mathcal{R}$ singled out by their favourite account of credence.

> **Bridge Theorem.** If $\succeq$ satisfies ϕ, then at least one of the probability functions c (or set of probability functions $\mathcal{C}$) whose existence is guaranteed by the *Representation Theorem* plays role $\mathcal{R}$ well enough to count as "your credal state."

Secondly, to answer the normative question, they put their favourite account of credence, their favourite representation theorem, and this bridge theorem to work in order to provide a more sophisticated argument for probabilism.

1. *Coherence Constraints*. Any rational agent's comparative belief relation $\succeq$ satisfies coherence constraints ϕ.

2. *Theory of Credence*. A real-valued function c (or set $\mathcal{C}$) counts as "your credal state" to the extent that it plays theoretical role $\mathcal{R}$. The better c (or $\mathcal{C}$) plays role $\mathcal{R}$, the more eligible it is as a "credal state candidate."

3. *Representation Theorem*. Relation $\succeq$ satisfies constraints ϕ if and only if $\succeq$ fully agrees (or almost agrees, or partially agrees, or strongly agrees) with some probability function c (or set of probability functions $\mathcal{C}$).

4. *Bridge Theorem*. If $\succeq$ satisfies ϕ, then at least one of the probability functions c (or set of probability functions $\mathcal{C}$) whose existence is

guaranteed by the *Representation Theorem* plays role $\mathcal{R}$ well enough to count as "your credal state."

C. *Probabilism*. Any rational agent has probabilistic credences (either precise credences given by c, or imprecise credences given by $\mathcal{C}$).

In Section 5–7, we will evaluate how this argument fares on each of our competing accounts of credence. But it is worth addressing some *general* concerns about this argumentative strategy here.[11]

Meacham and Weisberg worry that even if the axioms ϕ of your favourite representation theorem encode genuine coherence constraints on *rational* comparative belief, ordinary folks like you and me are not typically rational (Meacham & Weisberg, 2011, pp. 7–8).[12] Our comparative beliefs violate these constraints ϕ. So even if the Bridge Theorem is correct—even if the representation theorem in question *would* output a function c, or a set of functions $\mathcal{C}$ that deserves to be called "your credal state" *if* your comparative beliefs satisfied ϕ—it is silent about ordinary folks. The upshot: it does not help to answer the characterisation question in any interesting way. While it does specify sufficient conditions for having credences, those conditions are so demanding that they are more or less irrelevant for agents like us.

This concern, however, does not cut much ice. As we will see in Section 5–7, there is plenty to say about when ordinary folks—folks who reliably violate constraints of rationality—count as having credences on each of our competing accounts (measurement-theoretic, decision-theoretic, and epistemic interpretivist).

Meacham and Weisberg also worry that the "representation theorem argument" trivialises normative epistemology (Meacham & Weisberg, 2011, pp. 14–16). There is a gap, recall, between representability and psychological reality. Just because your comparative beliefs can be *represented* as arising from precise credences c (or imprecise credences $\mathcal{C}$) does not imply that you *in fact* have such credences. To avoid this problem, the objection goes, representation theorem arguments must *stipulatively define* an agent's credences to be given by the function c (or set $\mathcal{C}$) outputted by one's favourite representation theorem. But those theorems deliver probabilistic representations by construction. So it is simply true *by stipulative definition* that whenever an agent has credences, they are probabilistically

11 The following objections are adapted from Hájek (2009), Meacham and Weisberg (2011), and Titelbaum (2015).

12 Meacham and Weisberg are concerned primarily with representation theorems for *preference relations*. Accordingly, they focus on empirical data that shows that ordinary agents reliably violate putative coherence constraints on rational preference. For example, Kahneman and Tversky (1979) show that subjects consistently violate Savage's Independence Axiom, and Lichtenstein and Slovic (1971, 1973) show that subjects often have intransitive preferences. We adapt their concerns to the case of comparative belief *mutatis mutandis*.

coherent. Whence the normative force of probabilism then? The claim that *rational* credences are probabilistically coherent is trivial if *all* credences are probabilistically coherent *by definition*.

But again this concern need not give us much pause. We do *not* need to bridge the gap between representability and psychological reality by stipulative definition. Rather, we bridge that gap by (i) providing a theory of credence, which specifies the theoretical role $\mathcal{R}$ that a function c (or set $\mathcal{C}$) must play to count as "your credal state," and (ii) providing a bridge theorem, which establishes that some function c (or set $\mathcal{C}$) outputted by one's favourite representation theorem in fact plays role $\mathcal{R}$ sufficiently well. This strategy does *not* stipulatively define your credences as those given by c (or $\mathcal{C}$). Far from it. Establishing that c (or $\mathcal{C}$) plays $\mathcal{R}$ well enough to count as "your credal state" requires *substantive argumentation*.

It is safe, then, to put these general concerns to the side. Of course, their spectre lingers until we see the details about the relevant bridge theorems and so on (Section 5–7). We now turn our attention to evaluating how well this strategy answers the characterisation and normative questions, respectively, on each of our competing accounts of credence.

5 EVALUATING THE MEASUREMENT-THEORETIC VIEW

5.1 *Interpreting Credence Functions*

On the measurement-theoretic view, a credence function c (or set $\mathcal{C}$) is a mere numerical measurement system. It allows you to represent an agent's comparative belief structure, $\langle \Omega, \mathcal{F}, \succeq \rangle$, numerically in the following sense. Firstly, c maps the propositions X in $\mathcal{F}$ to real-valued proxies, $c(X)$. Secondly, it does so in a "structure-preserving fashion." If c fully agrees with $\succeq$, then one proxy $c(X)$ is larger than another $c(Y)$ exactly when our agent is more confident in X than Y (and $c(X) = c(Y)$ exactly when she is equally confident in X and Y):

$$X \succeq Y \Leftrightarrow c(X) \geq c(Y).$$

In this sense, the familiar "greater than or equal to" relation $\geq$ on the real numbers "preserves the structure" of our agent's comparative belief relation $\succeq$ on $\mathcal{F}$. Because of this, you can use the numerical measurement system in helpful ways. You can elicit certain comparative beliefs, infer properties of c, perform numerical calculations, and draw conclusions about what other comparative beliefs she must have (or must not have).

Similarly, a *set* of real-valued functions $\mathcal{C}$ can provide a numerical measurement system for $\succeq$. If $\mathcal{C}$ fully agrees with $\succeq$, then the c in $\mathcal{C}$ uniformly assign larger proxies to X than Y exactly when our agent is

more confident in X than Y (and uniformly assign equal proxies exactly when she is equally confident in X and Y):

$$X \succeq Y \Leftrightarrow c(X) \geq c(Y) \text{ for all } c \in \mathcal{C}.$$

Once more, this allows you to *use* the (imprecise) numerical measurement system $\mathcal{C}$ in helpful ways. You can elicit certain comparative beliefs, infer properties of $\mathcal{C}$, perform numerical calculations, and draw conclusions about what other comparative beliefs she must have (or must not have).

Weaker types of agreement yield numerical measurement systems fit for slightly different purposes. Suppose, for example, that *c strongly agrees* with $\succeq$:

$$X \succeq Y \Rightarrow c(X) \geq c(Y),$$
$$X \succ Y \Rightarrow c(X) > c(Y).$$

Such a measurement system licenses fewer inferences about $\succeq$ than fully agreeing systems. To see this, imagine that c is a probability function, X and Y are both incompatible with Z, and $X \succ Y$. Then since c strongly agrees with $\succeq$, we have $c(X) > c(Y)$. And since c is a probability function, $c(X \cup Z) > c(Y \cup Z)$. Hence $c(X \cup Z) \not\leq c(Y \cup Z)$. From this we can infer that $X \cup Z \not\preceq Y \cup Z$. But we *cannot* infer that $X \cup Z \succeq Y \cup Z$. If, on the other hand, c were to *fully* agree with $\succeq$, then we *could* make this latter inference.

To recap: the measurement-theoretic view takes a particular stand on the nature of the theoretical role $\mathcal{R}$ that a function c (or set $\mathcal{C}$) must play in order to count as "your credal state." More specifically, c (or $\mathcal{C}$) must fully agree (or almost agree, or partially agree, or strongly agree) with the agent's comparative beliefs, $\succeq$, in the way required to count as a numerical measurement system for $\succeq$. The better c (or $\mathcal{C}$) plays this role $\mathcal{R}$, the more eligible it is as a credal state candidate.

Importantly, though, *any* function c (or set $\mathcal{C}$) that plays this role $\mathcal{R}$ has equal claim to be called "your credence function," on the measurement-theoretic view. *Any* order-preserving mapping (homomorphism) from $\mathcal{F}$ into $\mathbb{R}$ is just as eligible as a credal state candidate as any other. So credence functions are not unique. Indeed, if c fully agrees (or almost agrees, or partially agrees, or strongly agrees) with $\succeq$, then any of the infinitely many strictly increasing transformations of c do so as well. So if you have *one* credence function, on this view, then you have *infinitely many*.

In addition, *interpreting* credence functions requires care, on the measurement-theoretic view. An agent's credence function does not wear its representationally significant features on its sleeve. Sorting out which features of one's credence function are representationally significant, rather than "mere artefacts," requires knowing what the "permissible

transformations" of that credence function are. That is, it requires knowing not only that b counts as "your credence function," but also what *other* functions c preserve the structure of your comparative and qualitative beliefs, and so count as "your credence function" as well. For example, on the standard Bayesian picture, an agent's credence function c is such that

$$\frac{c(X \cap Y)}{c(Y)} = c(X)$$

just in case she judges that Y is evidentially independent of X. But if credence functions c are mere numerical measurement systems for an agent's comparative beliefs $\succeq$, then properties like $c(X \cap Y)/c(X) = c(Y)$ are *not* representationally significant. They do not reflect anything *real* about the agent's doxastic state.

To see this, imagine that you take two blood tests. Let w^+_+ be the world in which both tests come back positive; w^+_- be the world in which the first comes back positive and the second negative; w^-_+ be the world in which the first comes back negative and the second positive; and w^-_- be the world in which both tests come back negative. You have comparative beliefs over propositions in the following Boolean algebra:

$$\mathcal{F} = \left\{ \begin{array}{c} \left\{w^+_+, w^+_-, w^-_+, w^-_-\right\} \\ \left\{w^+_+, w^+_-, w^-_+\right\}, \left\{w^+_+, w^+_-, w^-_-\right\} \\ \left\{w^+_+, w^-_+, w^-_-\right\}, \left\{w^+_-, w^-_+, w^-_-\right\} \\ \left\{w^+_+, w^+_-\right\}, \left\{w^+_+, w^-_+\right\}, \left\{w^+_+, w^-_-\right\} \\ \left\{w^+_-, w^-_+\right\}, \left\{w^+_-, w^-_-\right\}, \left\{w^-_+, w^-_-\right\}, \\ \left\{w^+_+\right\}, \left\{w^+_-\right\}, \left\{w^-_+\right\}, \left\{w^-_-\right\}, \\ \varnothing \end{array} \right\}.$$

In particular, your comparative beliefs are given by:

$$\begin{array}{c} \left\{w^+_+, w^+_-, w^-_+, w^-_-\right\} \succ \left\{w^+_+, w^-_+, w^-_-\right\} \succ \left\{w^+_+, w^+_-\right\} \\ \succ \left\{w^-_+, w^+_-, w^-_-\right\} \succ \left\{w^+_-, w^-_-\right\} \succ \left\{w^-_+, w^+_-\right\} \succ \left\{w^+_-\right\} \\ \succ \left\{w^+_+, w^-_+, w^-_-\right\} \succ \left\{w^+_+, w^-_-\right\} \succ \left\{w^+_+, w^-_+\right\} \succ \left\{w^+_+\right\} \\ \succ \left\{w^-_+, w^-_-\right\} \succ \left\{w^-_-\right\} \succ \left\{w^-_+\right\} \succ \varnothing. \end{array}$$

Then the probability functions b and c in Table 1 both fully agree with $\succeq$, and hence both count as a numerical measurement systems for $\succeq$. So both play the credal state role, on the measurement-theoretic view. But b is such that

$$\frac{b(\left\{w^+_+\right\})}{b(\left\{w^+_+, w^+_-\right\})} = \frac{1}{3} = b(\left\{w^+_+, w^-_+\right\}).$$

	w_+^+	w_-^+	w_+^-	w_-^-
b	$\frac{11}{36}$	$\frac{22}{36}$	$\frac{1}{36}$	$\frac{2}{36}$
c	$\frac{20}{64}$	$\frac{33}{64}$	$\frac{4}{64}$	$\frac{7}{64}$

Table 1: Probability functions b and c on $\mathcal{F}$. Both fully agree with $\succeq$.

Given that b is your credence function, the standard interpretation says: you think that the result of the first test provides no evidence one way or the other about the result of the second test. On the other hand, c is such that

$$\frac{c(\{w_+^+\})}{c(\{w_+^+, w_-^+\})} = \frac{20}{53} > \frac{24}{64} = c(\{w_+^+, w_+^-\}).$$

Given that c is your credence function, the standard interpretation says: you judge that a positive outcome on the second test *supports* or *confirms* a positive outcome on the first test. Finding out that the second test is positive *increases* your credence that the first test is positive too.

What is going on here? Answer: the "standard interpretation" reads more information into one's credence function than is actually encoded in that credence function. On the measurement-theoretic view, credence functions are nothing more than numerical measurement systems that encode the ordering determined by your comparative beliefs. (They are mere "ordinal scale" measurement systems, not "ratio scale" measurement systems.) But there is more to making judgments of evidential relevance and irrelevance than having a particular constellation of comparative beliefs. Agents who only have comparative beliefs simply are not opinionated enough to count as having opinions about evidential relevance and irrelevance. So credence functions do not reflect any such opinions.

Interpreting *imprecise* credal states requires care too. Suppose, for example, that you have opinions about the propositions in the Boolean algebra

$$\mathcal{F}^* = \{\Omega, X, \neg X, \varnothing\}.$$

Consider the precise and imprecise credal states on $\mathcal{F}^*$ given by the probability function b in Table 2 and the set of probability functions $\mathcal{C}$:

	Ω	X	$\neg X$	$\varnothing$
b	1	0.7	0.3	0

Table 2: Probability function b on $\mathcal{F}^*$

$$\mathcal{C} = \{c \mid c(\Omega) > c(X) > c(\neg X) > c(\varnothing)\}.$$

On the standard interpretation, b and $\mathcal{C}$ represent different doxastic states. An agent with credence function b is precisely 70% confident that X is true. An agent with imprecise credal state $\mathcal{C}$, in contrast, is at least 50% confident that X is true, but nothing stronger. On the standard interpretation, these are not idle differences. These differences in doxastic states are reflected in one's evaluative attitudes. For example, an agent with credence function b will have precisely the same fair buying and selling price for a unit gamble $\mathcal{G}$ on X, *viz.*, 0.7. Paying any price up to £0.7 for $\mathcal{G}$ is a good deal in her view. Selling $\mathcal{G}$ for any price over £0.7 is a good deal. But an agent with imprecise credal state $\mathcal{C}$ will have different buying and selling prices for $\mathcal{G}$. Paying any price up to £0.5 for $\mathcal{G}$ is a good deal, in her view. But selling $\mathcal{G}$ is only a determinately good deal if the buyer is willing to pay more than £1.

On the measurement-theoretic view, however, b and $\mathcal{C}$ represent *exactly the same doxastic state*. They both fully agree with the following comparative belief relation $\succeq$:

$$\Omega \succ X \succ \neg X \succ \varnothing.$$

They are both order-preserving mappings from $\mathcal{F}^*$ into the reals *that preserve exactly the same structure*. In this case, there is simply no substantive difference between being 70% confident, or 89.637% confident, or at least 50% confident that X is true. Only the comparative beliefs that b and $\mathcal{C}$ encode are psychologically real. Everything else is a "mere artefact" of one's preferred numerical measurement system. Both b and $\mathcal{C}$, and any other credal state that fully agrees with $\succeq$, plays exactly the same theoretical role: they represent the comparative beliefs captured by $\succeq$ in an elegant, easy-to-use, numerical fashion—nothing more, nothing less.

5.2 *Unary and Pluralist Variants*

We have focussed thus far on a particular *unary* variant of the measurement-theoretic view. On this view, credence functions are mere numerical measures of one's *comparative beliefs*. Having credences is nothing over and above having numerically representable comparative beliefs. You might be attracted to this view if, for example, you think that we can explain and rationalise everything important about choice and inference by appealing exclusively to comparative belief—no additional modes or types of doxastic judgment necessary. In that case, you might say: to the extent that we are willing to talk about *prima facie* distinct types of opinion—degrees of belief, full or categorical belief, etc.—they ought to ultimately reduce to comparative beliefs. Reducing those other types of opinion away will allow us to provide the simplest and most unified possible explanations of the relevant data regarding choice and inference.

But there is also a *pluralist* variant of the measurement-theoretic view, which you might find attractive if you are less optimistic about the explanatory power of comparative belief. On the pluralist version, agents have a genuine plurality of doxastic attitudes, not simply comparative beliefs. In addition to comparative beliefs, agents also have: (i) opinions about the *evidential dependence or independence* of one hypothesis on another; (ii) opinions about the *causal dependence or independence* of one variable on another; (iii) *full* or *categorical beliefs*; they may even (iv) explicitly *estimate* the values of all sorts of different variables, including the frequency of truths in a set of propositions, and the truth-values of individual propositions. Estimating, in this sense, is a matter of making a *sui generis* doxastic judgment—a type of judgment that may bear interesting relations to other types of judgments (normative relations, causal relations, etc.), but is not reducible to them. Estimating the truth-value of a proposition, in this sense, is what Jeffrey (2002) calls having an *exact judgmental probability* for the truth of that proposition.

On the pluralist measurement-theoretic view, your credence function *is* a mere numerical measurement system, but not a measure specifically *of your comparative belief relation*. Rather, on the pluralist view, you have a genuine plurality of comparative and qualitative doxastic attitudes, and your credence function is a measure of that *entire system of attitudes*.

Consider once again our blood test example. You have the following comparative beliefs:

$$\begin{gathered}\left\{w_+^+, w_-^+, w_+^-, w_-^-\right\} \succ \left\{w_+^+, w_-^+, w_-^-\right\} \succ \left\{w_+^+, w_+^-, w_-^+\right\} \\ \succ \left\{w_+^+, w_-^+\right\} \succ \left\{w_+^-, w_-^+, w_-^-\right\} \succ \left\{w_-^+, w_-^-\right\} \succ \left\{w_+^-, w_-^+\right\} \\ \succ \left\{w_-^+\right\} \succ \left\{w_+^+, w_+^-, w_-^-\right\} \succ \left\{w_+^+, w_-^-\right\} \succ \left\{w_+^+, w_+^-\right\} \\ \succ \left\{w_+^+\right\} \succ \left\{w_+^-, w_-^-\right\} \succ \left\{w_-^-\right\} \succ \left\{w_+^-\right\} \succ \varnothing.\end{gathered}$$

But now imagine that you have a wide range of comparative and qualitative opinions, not just comparative beliefs. You think, for example, that when you find out the result of the first test (positive or negative), this provides no evidence one way or the other about the result of the second test. (Perhaps the tests probe two different, unrelated conditions.) That is, you judge $\{w_+^+, w_-^+\}$ and $\{w_+^-, w_-^-\}$ to be *evidentially independent* of $\{w_+^+, w_+^-\}$ and $\{w_-^+, w_-^-\}$, and vice versa.

In addition, you have certain *full beliefs* or *categorical beliefs*. Let's suppose that you believe that the first test will come back positive. (It probes for a condition that you quite clearly have.) That is, you fully believe $\{w_+^+, w_-^+\}$. And you believe all of the logical consequences of this proposition. But you have no further full or categorical beliefs.

Finally, you judge $1/3$ to be the *best estimate* of the truth-value of the proposition that the second test will come back positive. (Recall, a proposition's truth-value is 1 if it is true and 0 if it is false.) In Jeffrey's parlance, you have a judgmental probability of $1/3$ for the proposition $\{w_+^+, w_+^-\}$.

So you have a genuine plurality of doxastic attitudes: you have comparative beliefs; you make evidential independence judgments; you have full or categorical beliefs; you also estimate the truth-values of certain propositions (you have exact judgmental probabilities). On the pluralist measurement-theoretic view, your credence function is a measure of this entire system of attitudes.

To make this more precise, let's model your doxastic attitudes using a relational structure:

$$\mathcal{A} = \left\langle \mathcal{F}, \succeq, \mathcal{I}, \mathcal{B}, \mathcal{E}_{1/3} \right\rangle.$$

$\mathcal{A}$ comprises your Boolean algebra $\mathcal{F}$ of subsets of $\Omega = \{w_+^+, w_-^+, w_+^-, w_-^-\}$, together with a *comparative belief relation* $\succeq$ on $\mathcal{F}$, an *independence relation* $\mathcal{I}$, a (unary) *belief relation* $\mathcal{B}$, and a (unary) *estimation relation* $\mathcal{E}_{1/3}$.

$\mathcal{I}$ models your evidential independence judgments. It will be convenient to think of $\mathcal{I}$ as a 3-place relation on $\mathcal{F}$:

$$\mathcal{I}(X, Y, X \cap Y)$$
iff
you judge X to be evidentially independent of Y.

Since you judge $\{w_+^+, w_-^+\}$ and $\{w_+^-, w_-^-\}$ to be independent of $\{w_+^+, w_+^-\}$ and $\{w_-^+, w_-^-\}$, and vice versa, we have:

$$\mathcal{I}(\left\{w_+^+, w_-^+\right\}, \left\{w_+^+, w_+^-\right\}, \left\{w_+^+\right\}),$$
$$\mathcal{I}(\left\{w_+^-, w_-^-\right\}, \left\{w_+^+, w_+^-\right\}, \left\{w_+^-\right\}),$$
$$\mathcal{I}(\left\{w_+^+, w_-^+\right\}, \left\{w_-^+, w_-^-\right\}, \left\{w_-^+\right\}),$$
$$\mathcal{I}(\left\{w_+^-, w_-^-\right\}, \left\{w_-^+, w_-^-\right\}, \left\{w_-^-\right\}).$$

We also have $\mathcal{I}(Y, X, X \cap Y)$ for each of these four independence judgments $\mathcal{I}(X, Y, X \cap Y)$.

Likewise, $\mathcal{B}$ models your full or categorical beliefs:

$$\mathcal{B}(X) \;\; \textit{iff} \;\; \text{you believe } X.$$

Since you believe $\{w_+^+, w_-^+\}$ and all of its logical consequences, we have:

$$\mathcal{B}(\left\{w_+^+, w_-^+, w_+^-, w_-^-\right\}),$$
$$\mathcal{B}(\left\{w_+^+, w_-^+, w_+^-\right\}),$$
$$\mathcal{B}(\left\{w_+^+, w_-^+, w_-^-\right\}),$$
$$\mathcal{B}(\left\{w_+^+, w_-^+\right\}).$$

Finally, $\mathcal{E}_{1/3}$ models your explicit estimates of truth-values:

$$\mathcal{E}_x(X)$$
$$\textit{iff}$$
you judge x to be the best estimate of the truth-value of X.

Since you judge $1/3$ to be the best estimate of the truth-value of $\{w_+^+, w_+^-\}$, we have:

$$\mathcal{E}_{1/3}\left(\left\{w_+^+, w_+^-\right\}\right).$$

On the pluralist view, your credence function is a measure of your entire system of attitudes:

$$\mathcal{A} = \langle \mathcal{F}, \succeq, \mathcal{I}, \mathcal{B}, \mathcal{E}_{1/3} \rangle.$$

It is a homomorphism—a structure-preserving mapping—that takes $\mathcal{A}$ into some numerical structure $\mathcal{A}^*$.

$$\mathcal{A}^* = \left\langle \mathbb{R}, \succeq^*, \mathcal{I}^*, \mathcal{B}^*, \mathcal{E}_{1/3}^* \right\rangle.$$

That is, your credence function c maps $\mathcal{F}$ into $\mathbb{R}$ in a way that preserves $\mathcal{A}$'s structure, so that:[13]

$$\begin{aligned} X \succeq Y &\Leftrightarrow c(X) \succeq^* c(Y), \\ \mathcal{I}(X, Y, X \cap Y) &\Leftrightarrow \mathcal{I}^*(c(X), c(Y), c(X \cap Y)), \\ \mathcal{B}(X) &\Leftrightarrow \mathcal{B}^*(c(X)), \\ \mathcal{E}_{1/3}(X) &\Leftrightarrow \mathcal{E}_{1/3}^*(c(X)). \end{aligned}$$

Which numerical structure c takes $\mathcal{A}$ into, on the measurement-theoretic view, is either a matter of convention or a matter to be decided on practical grounds. For illustrative purposes, let's choose a familiar numerical structure. Let $\succeq^*$ by the "greater than or equal to" relation, $\geq$. Let $\mathcal{I}^*$ be the standard probabilistic independence relation:

$$\mathcal{I}^*(c(X), c(Y), c(X \cap Y)) \;\; \textit{iff} \;\; c(X)c(Y) = c(X \cap Y).$$

13 We could swap full agreement for almost, or partial, or strong agreement here. Weaker notions of agreement would provide us with weaker notions of structure-preservation.

Let $\mathcal{B}^*$ be a *Lockean belief relation*, so that believed propositions X have real-valued proxies $c(X)$ that are greater than (or equal to) some threshold τ (for concreteness let $\tau = 5/6$):

$$\mathcal{B}^*(c(X)) \textit{ iff } c(X) \geq \tau.$$

Finally, let $\mathcal{E}^*_{1/3}$ be:

$$\mathcal{E}^*_{1/3}(c(X) \textit{ iff } c(X) = 1/3.$$

This ensures that for any structure-preserving measurement system, c, you explicitly judge $1/3$ to be the best estimate of X's truth-value just in case $c(X) = 1/3$.

The important observation to make is this: the pluralist view carves out a bigger job for credence functions to do than the reductive view. Credence functions must do more than preserve the order induced on $\mathcal{F}$ by your comparative belief relation. They must also preserve the structure induced by your various other doxastic attitudes: your evidential independence judgments, full or categorical beliefs, and so on. So a function $c : \mathcal{F} \to \mathbb{R}$ may well do the work required to count as "your credence function" on the unary view, but yet fall short of that mark on the pluralist view.

Consider, for example, the function $b : \mathcal{F} \to \mathbb{R}$:

$$\begin{aligned}
b(\{w^+_+, w^+_-, w^-_+, w^-_-\}) &= 1, & b(\{w^+_+, w^-_+, w^-_-\}) &= 31/64,\\
b(\{w^+_+, w^+_-, w^-_-\}) &= 60/64, & b(\{w^+_+, w^-_-\}) &= 27/64,\\
b(\{w^+_+, w^-_+, w^+_-\}) &= 57/64, & b(\{w^+_+, w^-_+\}) &= 24/64,\\
b(\{w^+_+, w^+_-\}) &= 53/64, & b(\{w^+_+\}) &= 20/64,\\
b(\{w^-_+, w^+_-, w^-_-\}) &= 44/64, & b(\{w^-_+, w^-_-\}) &= 11/64,\\
b(\{w^+_-, w^-_-\}) &= 40/64, & b(\{w^-_-\}) &= 7/64,\\
b(\{w^-_+, w^+_-\}) &= 37/64, & b(\{w^-_+\}) &= 4/64,\\
b(\{w^+_-\}) &= 33/64, & b(\varnothing) &= 0.
\end{aligned}$$

It is easy to verify that b fully agrees with $\succeq$, *i.e.*,

$$X \succeq Y \Leftrightarrow b(X) \geq b(Y).$$

So b is a real-valued measure of your comparative belief relation $\succeq$. Hence, it counts as a credence function on the unary measurement-theoretic view. It preserves the structure on $\mathcal{F}$ induced by your comparative beliefs—the only type of doxastic attitude that the unary view countenances. But it does *not* play the theoretical role required to count as a credence function on the pluralist measurement-theoretic view. To do *that*, it must also preserve

the structure on $\mathcal{F}$ induced by your various other doxastic attitudes: your independence judgments, full or categorical beliefs, and so on. But b falls short of that mark.

For example, you think that the outcome of the first test provides no evidence about the outcome of the second. But b does not treat $\{w_+^+, w_-^+\}$ and $\{w_+^+, w_+^-\}$, for example, as independent, in the way specified by $\mathcal{I}^*$:

$$b\left(\left\{w_+^+, w_-^+\right\}\right) b\left(\left\{w_+^+, w_+^-\right\}\right) = \frac{53}{64} \cdot \frac{24}{64} = \frac{159}{512} \neq \frac{160}{512} = \frac{20}{64} = b\left(\left\{w_+^+\right\}\right).$$

Similarly, you believe that the first test will come back positive. That is, you fully believe $\{w_+^+, w_-^+\}$. But b does not treat $\{w_+^+, w_-^+\}$ as believed, in the way specified by the Lockean belief relation $\mathcal{B}^*$. It maps $\{w_+^+, w_-^+\}$ to a real-valued proxy $b(\{w_+^+, w_-^+\}) = 53/64 \approx 0.828$ below the threshold $\tau = 5/6 \approx 0.833$ required for full or categorical belief.

Finally, you judge $1/3$ to be the best estimate of the truth-value of the proposition that the second test will come back positive. You have a judgmental probability of $1/3$ for the proposition $\{w_+^+, w_+^-\}$. But b fails to map $\{w_+^+, w_+^-\}$ to the real-valued proxy set aside by $\mathcal{E}^*_{1/3}$ for such propositions, *viz.*, $1/3$. Instead, $b(\{w_+^+, w_+^-\}) = 24/64 = 0.375$.

The upshot: while b preserves the structure on $\mathcal{F}$ induced by your comparative beliefs, it fails to preserve the additional structure induced by your various other doxastic attitudes: your evidential independence judgments, full or categorical beliefs, and so on. So while b *does* count as one of your (infinitely many) credence functions on the unary measurement-theoretic view, it does *not* count as one on the pluralist measurement-theoretic view.

In contrast, the function $c : \mathcal{F} \to \mathbb{R}$ of Figure 6 counts as "your credence function" on both the unary and pluralist views. (The interested reader may verify this for herself.)

To recap: the measurement-theoretic view stakes out a particular position on the theoretical role $\mathcal{R}$ that a function c (or set of functions $\mathcal{C}$) must play in order to count as "your credal state." It says that c (or $\mathcal{C}$) must fully agree (or almost agree, or partially agree, or strongly agree) with your comparative and qualitative opinions—comparative beliefs, evidential independence judgments, full or categorical beliefs, etc.—in the way required to count as a numerical measure of that *entire system of attitudes*. The better c (or $\mathcal{C}$) plays this role $\mathcal{R}$, the more eligible it is as a credal state candidate. On the unary measurement-theoretic view, the fundamental type of doxastic attitude is *comparative belief*. So credal states are numerical measures of comparative beliefs. On the pluralist measurement-theoretic view, you have a genuine plurality of doxastic attitudes. So credal states are numerical measures of a more highly structured system of attitudes.

$$
\begin{aligned}
c(\{w_+^+, w_-^+, w_+^-, w_-^-\}) &= 1, & c(\{w_+^+, w_+^-, w_-^-\}) &= \frac{14}{36},\\
c(\{w_+^+, w_-^+, w_-^-\}) &= \frac{35}{36}, & c(\{w_+^+, w_-^-\}) &= \frac{13}{36},\\
c(\{w_+^+, w_+^-, w_-^+\}) &= \frac{34}{36}, & c(\{w_+^+, w_+^-\}) &= \frac{12}{36},\\
c(\{w_+^+, w_-^+\}) &= \frac{33}{36}, & c(\{w_+^+\}) &= \frac{11}{36},\\
c(\{w_+^-, w_-^+, w_-^-\}) &= \frac{25}{36}, & c(\{w_+^-, w_-^-\}) &= \frac{3}{36},\\
c(\{w_-^+, w_-^-\}) &= \frac{24}{36}, & c(\{w_-^-\}) &= \frac{2}{36},\\
c(\{w_+^-, w_-^+\}) &= \frac{23}{36}, & c(\{w_+^-\}) &= \frac{1}{36},\\
c(\{w_-^+\}) &= \frac{22}{36}, & c(\varnothing) &= 0.
\end{aligned}
$$

Figure 6: Credence function c

To streamline our discussion, we will focus on the the unary variant of the measurement-theoretic view going forward.

5.3 *The Characterisation and Normative Questions*

How does the unary measurement-theoretic account answer the characterisation question? When exactly *is* there a function c (or set $\mathcal{C}$) that fully agrees (or almost agrees, or partially agrees, or strongly agrees) with your comparative belief relation $\succeq$ in the way required to count as a numerical measure of $\succeq$?

We explored a partial answer to this question earlier. Scott (1964) proves that there is a *probability function* that *fully agrees* with your comparative belief relation $\succeq$ just in case $\succeq$ satisfies Non-Triviality, Non-Negativity, Totality, and Isovalence. Rios Insua (1992) and Alon and Lehrer (2014) prove that there is a *set* of probability functions that *fully agrees* with $\succeq$ just in case $\succeq$ satisfies Reflexivity, Non-negativity, Non-triviality, and the Generalised Finite-Cancellation axiom. Kraft et al. (1959) proves that there is a probability function that *almost agrees* with $\succeq$ just in case it satisfies Almost-Cancellation. Adams (1965) and Fishburn (1969) prove that there is a probability function that *partially agrees* with $\succeq$ just in case it satisfies Partial-Cancellation. Finally, in proving the Generalised Scott Theorem, we identified sufficient conditions for the existence of a probability function that *strongly agrees* with $\succeq$: Non-Triviality, Non-Negativity, and Isovalence.

Pinning down necessary and sufficient conditions for strong representability is an open problem.

These representation theorems tell us what it takes to count as having *probabilistically coherent credences*, on the measurement-theoretic view. But they do not answer the more general characterisation question: when is your comparative belief relation sufficiently well-behaved for you to count as having credences *full stop*, coherent or not?

Krantz et al. (1971) provide an answer. They show that a comparative belief relation $\succeq$ fully agrees with a real-valued function c if and only if $\succeq$ is a weak order, *i.e.*, $\succeq$ satisfies Transitivity and Totality (Krantz et al., 1971, p. 15, Theorem 1). So if a real-valued function c counts as a structure-preserving numerical measure of $\succeq$ just in case c fully agrees with $\succeq$, and if precise credence functions *just are* structure-preserving numerical measures of $\succeq$, then we now know exactly when you count as having precise credences *full stop*. You count as having precise credences just in case $\succeq$ satisfies Transitivity and Totality.

Weaker notions of agreement set weaker standards for "structure preservation." They thereby make it easier for a real-valued function (or set of functions) to count as a structure-preserving numerical measurement system for $\succeq$. In turn, your comparative beliefs need not satisfy such strict constraints for you to count as having credences. For example, *every* comparative belief relation $\succeq$ *almost* agrees with a real-valued function c. So if all that is required for structure-preservation is almost-agreement, then *nothing whatsoever* is required of $\succeq$ for you to count as having credences. Any comparative belief relation will do. More interestingly, $\succeq$ *strongly* agrees with a real-valued function c if and only if $\succeq$ satisfies *weak transitivity* (see the appendix for proof).[14]

Weak Transitivity. If $X \succeq Y_1 \succeq \ldots \succeq Y_n \succeq Z$, then $X \not\prec Z$.

So if structure-preservation requires strong-agreement and nothing more, then you count as having precise credences just in case $\succeq$ satisfies Weak Transitivity.

What then of Meacham and Weisberg's concern? They claim that the axioms of typical representation theorems for comparative belief are *so* demanding that only perfectly rational agents could possibly satisfy them. So even if those axioms do encode sufficient conditions for having credences, they are more or less irrelevant for irrational agents like us. They leave entirely open whether *our* comparative beliefs are ever well-behaved enough for *us* to count as having credences.

But your comparative beliefs need not satisfy the axioms of Scott's Theorem (or the Almost-Cancellation axiom, or the Partial-Cancellation axiom, etc.) for you to count as having credences. Such axioms encode

14 The proof strategy for this theorem is due to Catrin Campbell-Moore.

necessary and sufficient conditions for having *probabilistic* credences. Probabilistic credence functions, however, are not the only credence functions in town. Your comparative beliefs only need to satisfy weaker constraints, such as Weak Transitivity, to count as having credences *tout court*. Weak Transitivity is not nearly as demanding as Scott's axiom.

It is also worth nothing that even though Scott's axiom and the like *seem* complicated, it is *not* obvious that they are excessively difficult for agents like us to satisfy. It may be computationally intensive to run a diagnostic program which continually checks your comparative beliefs for violations of Scott's axiom. And *if* we had to run such a program to reliably satisfy Scott's axiom, then you might well expect that limited agents like us typically violate it. But no such program is necessary. Nature is replete with cheap solutions to seemingly computationally intensive problems. This is one main lesson of the embodied cognition movement in cognitive science.[15] Agents like us might well use computationally cheap strategies, rather than demanding diagnostic programs, in order to minimise violations of Scott's axiom and other coherence constraints.

Meacham and Weisberg also worry that the measurement-theoretic view and its ilk count the wrong functions as eligible credal state candidates (Meacham & Weisberg, 2011, p. 5). On the (unary) measurement-theoretic view, any of the infinitely many numerical measurement systems for $\succeq$ count as equally eligible credal state candidates. But some clearly are more eligible than others. For example, suppose that Holmes has opinions about finitely many propositions, *e.g.*, about whether Moriarty is in London, etc. Then Holmes is struck on the head. The blow does not change Holmes' comparative beliefs. He is still more confident that Moriarty is in London than Paris, and so on. But it does raise his confidence that Moriarty is in London. Then clearly *something* has changed about which functions are the most eligible candidates for counting as Holmes' credence function. But on the measurement-theoretic view, nothing at all has changed.

One of two things is going on here. Option 1: the objection tacitly presupposes that the measurement-theoretic view simply misidentifies the theoretical role $\mathcal{R}$ that a function c (or set $\mathcal{C}$) must play in order to count as "your credal state." That is a meaty, substantive debate, and we will not explore it any further. Option 2: the objection tacitly presupposes that Holmes makes explicit judgments about the best estimates of truth-values, or something of the sort. But that assumes pluralism. And the pluralist

15 Consider, for example, the "outfielder's problem" (Clark, 2015, p. 12). It might seem miraculous that baseball players manage to catch fly balls if doing so involves: (i) estimating the position of a ball at various time points; (ii) using this information to estimate the ball's trajectory; (iii) calculating where the ball will land on the basis of its trajectory. This is computationally intensive! Luckily, there is a computationally cheap solution. You can just move your body in a way that keeps the ball centred in your visual field. This strategy uses the agent's body to reduce computational demand.

measurement-theoretic view simply does not say that any of the infinitely many numerical measurement systems of Holmes' comparative beliefs are equally eligible candidates for counting as Holmes' credence function.

So much for the characterisation question. How does the unary measurement-theoretic account answer the normative question? Why should we expect *rational* agents to have *probabilistically coherent* credences?

How we answer the normative question depends on what we say about structure preservation. If we say, for example, that c must *fully agree* with $\succeq$ to count as a structure-preserving numerical measure of $\succeq$, and in turn count as "your credal state," then the following argument answers the normative question.

1. *Coherence Constraints*. Any rational agent's comparative belief relation $\succeq$ satisfies Non-Triviality, Non-negativity, Totality, and Isovalence.

2. *Theory of Credence*. A real-valued function c (or set $\mathcal{C}$) counts as "your credal state" just in case it is a structure-preserving numerical measure of $\succeq$, *i.e.*, just in case it plays the "structure-preservation role" $\mathcal{R}$. And c preserves the structure of $\succeq$ just in case c *fully agrees* with $\succeq$.

3. *Scott's Theorem*. Relation $\succeq$ satisfies Non-Triviality, Non-negativity, Totality, and Isovalence if and only if $\succeq$ fully agrees with some probability function c.

4. *Bridge Theorem*. If $\succeq$ satisfies Non-Triviality, Non-negativity, Totality, and Isovalence, then there is some probability function c that plays role $\mathcal{R}$ well enough to count as "your credal state." (*From 2 and 3*)

C. *Probabilism*. Any rational agent has probabilistic credences. (*From 1 and 4*)

Now, you might quibble with premise 1. You might doubt whether Totality, for example, encodes a genuine constraint of rationality. In that case, we might weaken our putative coherence constraints by adopting less demanding standards for structure preservation. For example, if we say that structure preservation requires only *strong* agreement with $\succeq$, rather than *full* agreement, then we can offer the following argument.

1*. *Coherence Constraints*. Any rational agent's comparative belief relation $\succeq$ satisfies Non-Triviality, Non-negativity, and Isovalence.

2*. *Theory of Credence*. A real-valued function c (or set $\mathcal{C}$) counts as "your credal state" just in case it is a structure-preserving numerical measure of $\succeq$, *i.e.*, just in case it plays the "structure-preservation

role" $\mathcal{R}$. And c preserves the structure of $\succeq$ just in case c *strongly agrees* with $\succeq$.

3*. *Corollary of GST*. If $\succeq$ satisfies Non-Triviality, Non-negativity, and Isovalence, then $\succeq$ strongly agrees with some probability function c.

4*. *Bridge Theorem*. If $\succeq$ satisfies Non-Triviality, Non-negativity, and Isovalence, then there is some probability function c that plays role $\mathcal{R}$ well enough to count as "your credal state." (*From 2* and 3**)

C*. *Probabilism*. Any rational agent has probabilistic credences. (*From 1* and 4**)

Each type of agreement (full, strong, almost, partial) yields a different variant of this argument. Whether you find any of them compelling will depend on (i) which putative coherence constraints you find plausible or implausible (premise 1), and (ii) what type of agreement is required for credence functions to play any auxiliary theoretical roles you deem important (premise 2).

At this point, you might be a bit suspicious. Doesn't this argument trivialise probabilism? True enough, you might say, the probability functions outputted by Scott's theorem are fit to play the "credal state role" $\mathcal{R}$ on measurement-theoretic view. But that is because we reverse engineered $\mathcal{R}$ so that Scott's theorem outputs *exactly* the right sorts of functions to play $\mathcal{R}$! We *stipulatively defined* $\mathcal{R}$ to be the role of preserving the structure of $\succeq$. Then we *stipulatively defined* structure-preservation to be a matter of *fully agreeing with* $\succeq$. But given these stipulative definitions, it follows *trivially* that the probability functions outputted by Scott's theorem play $\mathcal{R}$ well enough to count as "your credal state." Probabilism seems less like a substantive normative thesis, then, and more like a trivial consequence of stipulative definitions.

This suspicion is doubly off the mark. Firstly, the measurement-theoretic account of credence puts forward a *substantive* claim about the principal theoretical role of credence functions c (and imprecise credal states $\mathcal{C}$). It is motivated by the thought that our opinions are qualitative. At bottom, we have opinions like: comparative beliefs, full beliefs, etc. And the best way to understand the numbers that we use to describe these qualitative attitudes is in exactly the same way that we understand the numbers that we use to describe length, mass, volume, etc., *viz.*, as numerical measurement systems. Whether this is right or wrong, it is surely no *stipulative definition*. Secondly, as we have already emphasised, representability by a probability function is strictly stronger than representability by a real-valued function. Establishing that the stronger axioms (*e.g.*, Scott's axioms) encode genuine constraints of rationality, rather than merely the weaker axioms (*e.g.*, Transitivity and Totality) is non-trivial. As a result, establishing *probabilism* is

non-trivial, even if we simply grant the measurement theorist her account of credence.

You might also be concerned that the strategy above only establishes half of probabilism. If successful, it establishes that all rational agents have probabilistic credences. But it does *not* establish that rational agents have *only* probabilistic credences. On the measurement-theoretic view, any agent that counts as having a credence function at all in fact has a plurality of credence functions. If she is rational, then at least one of these will be probabilistically coherent. But many will not be. If c is a probability function that fully agrees with $\succeq$ (or almost agrees, or partially agrees, or strongly agrees), then any of the infinitely many strictly increasing transformations of c do so as well. These transformations will not in general be probability functions.

But this auxiliary thesis—that no rational agent has a probabilistically incoherent credence function—is not particularly interesting, on the measurement-theoretic view. The reason: nothing interesting hinges on whether some incoherent function (or set of functions) is fit to play the "credal state role" for you. On the measurement-theoretic view, credence functions are mere numerical measurement systems for comparative belief; systems which allow you to measure certain parts of an agent's comparative belief relation $\succeq$ and draw inferences about other parts of $\succeq$. *Probabilistic* measurement systems are particularly useful for this end. Probability functions have nice properties; properties that simplify the calculations necessary to draw inferences about $\succeq$. Whether or not some unhelpful, incoherent measurement system exists is neither here nor there.[16] If some such system exists, who cares! It's not hurting anyone. The interesting question is *whether the useful things exist*.

But if an agent has incoherent credences, doesn't this come at some cost to *her*? Doesn't it hurt *her*? De Finetti (1964) shows that any agent with incoherent credences is Dutch bookable, *i.e.*, susceptible to sure loss at the hands of a clever bettor. And Joyce (1998, 2009) shows that any agent with incoherent credences is accuracy-dominated, *i.e.*, there are distinct

16 Of course, not *all* incoherent measurement systems are unhelpful. For example, suppose that b is a probability function and fully agrees with $\succeq$. Let $c(X) = e^{b(X)}$. Then c fully agrees with $\succeq$. But while b satisfies Finite Additivity:

$$b(X \cup Y) + b(X \cap Y) = b(X) + b(Y),$$

c satisfies Finite Multiplicativity:

$$c(X \cup Y) \cdot c(X \cap Y) = c(X) \cdot c(Y).$$

Note, though, that c is no less "helpful" than b. All of the theorems of probability theory can be rewritten in terms of a multiplicative scale rather than an additive scale. So c could be used to facilitate inference about $\succeq$ just as well as b. For analogous remarks regarding additive and mutliplicative measures in physics, see (Krantz et al., 1971, p. 100).

(coherent) credences that are guaranteed to be closer to the truth than hers. Aren't these costs—pragmatic and epistemic—that any agent with incoherent credences must pay?

No. Not on the measurement-theoretic view. De Finetti assumes that if c counts as your credence function, then $c(X)$ is both your fair buying and selling price for a unit gamble on X. But this is simply not so on the measurement-theoretic view. Credence functions represent your comparative beliefs $\succeq$ in an elegant, easy-to-use, numerical fashion—nothing more, nothing less. It is simply not the *job* of a credence function to capture your fair buying (or selling) prices. We can*not* read your fair buying and selling prices off of c in any straightforward fashion. Indeed, to infer *anything* about your betting behaviour from c, we need decision-theoretic norms that specify how rational *comparative beliefs* and *preferences* hang together. For example, following (Savage, 1954, Section 3.2), we might suggest the following.

> **Coherence**. If $X \succ Y$, then you ought to prefer to stake good outcomes on X than Y. More carefully, if you strictly prefer outcome o to o^*, and $X \succ Y$, then you ought to strictly prefer A to B:
>
> $$A = [o \text{ if } X,\ o^* \text{ if } \neg X]\,,$$
> $$B = [o \text{ if } Y,\ o^* \text{ if } \neg Y]\,.$$
>
> Moreover, you ought to be willing to sacrifice some small amount ϵ to exchange A for B.

If you satisfy *Coherence*, and c fully agrees with $\succeq$, then we can use c to infer *something* about your betting behaviour. For example, if $c(X) = 0.7$ and $c(Y) = 0.6$, then we can infer that you prefer to let £1 ride on X than on Y, and would even be willing to pay some small amount to exchange the first gamble for the second. But we *cannot* infer that your fair buying (selling) price for X is £0.7, or that your fair buying (selling) price for Y is £0.6.

Without this crucial assumption—that credences encode fair buying/selling prices—we cannot provide a de Finetti-style Dutch book argument to show that no rational agent has incoherent credences. Having an incoherent credence function does not mean that you have incoherent fair buying/selling prices, and hence does not mean that your buying/selling prices render you Dutch-bookable.

In a similar fashion, Joyce assumes that if c counts as your credence function, then $c(X)$ is your best estimate of X's truth-value. Moreover, the accuracy of these estimates is what makes your doxastic state better or worse from the epistemic perspective. (Accuracy is the *principal* source of epistemic value, anyway.) But again, this is not so on the measurement-theoretic view. Credence functions are mere numerical measures of com-

parative belief relations. It is simply not the *job* of a credence function to capture your best estimates of truth values, on the measurement-theoretic view. The upshot: having an incoherent credence function does not mean that you in any sense have incoherent *truth-value estimates*; so it does not mean having *accuracy-dominated* truth-value estimates; so it does not mean having a doxastic state that is epistemic-value-dominated.

Finally, one might level a criticism similar to Meacham and Weisberg's (2011, pp. 19–20) criticism of Lyle Zynda. Zynda is a proponent of the unary measurement-theoretic account (Zynda, 2000, pp. 66–68).[17] On Zynda's view, there are comparative beliefs—agents are more confident in some propositions than others—but there are no additional modes or types of doxastic judgment. To the extent that we countenance talk of fully believing a proposition, or believing something *much more strongly* than something else, this better ultimately reduce to talk about comparative beliefs.

Meacham and Weisberg object that comparative beliefs lack the structure required to explain everything about choice and inference that we would like to explain. So even if the measurement theorist provides *some* reason to expect rational agents to have probabilistic credences, the background picture of the basic stock of doxastic attitudes available to such agents is too impoverished for their arguments to cut much ice.

For example, if we buy the unary measurement-theoretic account, then the well-known problem of interpersonal utility comparisons rears its head as a problem of interpersonal *credal* comparisons. Just as it makes no sense to say that Ashan desires chocolate ice cream more strongly than Bilal does, on the measurement-theoretic account (since there is no common scale one which their preferences are measured), similarly it makes no sense to say that Ashan is more confident that it will rain than Bilal is. But, at least in certain cases, it seems that we need such facts to explain choice behaviour. Why did Ashan grab his umbrella but Bilal did not? One possible explanation: both are more confident than not that it will rain, but Ashan is more confident than Bilal. On the the unary measurement-theoretic account, such explanations are unavailable, Meacham and Weisberg argue. More generally:

> the extra-ordinal structure contained in the standard Bayesian picture of degrees of belief is not idle. Magnitudes encode important features of our degrees of belief, and if we abandon this structure, degrees of belief lose much of their utility. (Meacham & Weisberg, 2011, p. 20)

17 Like Maher, Zynda subscribes to the thesis of the primacy of practical reason (*cf.*, Zynda 2000, p. 55). Credence functions are numerical measures of comparative beliefs. But preferences are the real thing. Comparative beliefs reduce to preferences.

You might not think that there is much to this line of criticism. For example, Ashan might think that rain is just as likely as picking a black ball at random from an urn containing 99 black balls and 1 white ball. Bilal, in contrast, might think that rain is just as likely as picking a black ball at random from an urn containing 51 black balls and 49 white balls.[18] These individual comparative belief facts help to explain why Ashan grabbed his umbrella but Bilal did not at least as well as the purported interpersonal fact that Ashan is more confident than Bilal. It is not obvious, then, that there is any genuine problem of interpersonal credal comparisons to resolve.

Even if you do think there is something to this line of criticism, note that it is not an objection to the measurement-theoretic account of credence *per se*. It is only an objection to the *unary* measurement-theoretic account. A pluralist faces no such problems. Of course, in answering the normative question, a pluralist cannot simply appeal to Scott's theorem. Scott's theorem only shows that comparative belief relations with certain properties are probabilistically representable. The pluralist must appeal to a representation theorem that shows that a *more comprehensive system of doxastic attitudes* with certain properties is probabilistically representable. But there is no principled reason for thinking that such representation theorems are not forthcoming.

6 EVALUATING THE DECISION-THEORETIC VIEW

On the decision-theoretic view, the principal theoretical role of an agent's credal state is to encode her fair buying and selling prices. Recall, an agent's fair buying price for a gamble $\mathcal{G}$ is the largest amount $\mathcal{B}(\mathcal{G})$ that she could pay for $\mathcal{G}$ without making herself worse off. She pays $\mathcal{B}(\mathcal{G})$, receives $\mathcal{G}$, and is no worse than the status quo, in her own view. Her fair selling price for $\mathcal{G}$ is the smallest amount $\mathcal{S}(\mathcal{G})$ that someone else would have to pay her in exchange for $\mathcal{G}$ to avoid being worse off. She receives $\mathcal{S}(\mathcal{G})$, commits to shelling out $\mathcal{G}$'s payoff, and is no worse than the status quo, in her own view.

Gambles are measurable quantities $\mathcal{G} : \Omega \to \mathbb{R}$. For simplicity, we will assume that $|\Omega| = n$, and treat gambles as vectors in $\mathbb{R}^n$. When we model a gamble as a vector

$$\mathcal{G} = \langle \mathfrak{g}_1, \ldots, \mathfrak{g}_n \rangle ,$$

18 Both de Finetti (1931) and Koopman (1940a) use "partition axioms" to extract quantitative information from belief relations in roughly this way. For a recent approach along these lines, see Elliott (2018). You might also model agents as having comparative estimation relations, as explored in §2.3. Comparative estimation relations allow for a much richer and explanatorily powerful set of doxastic attitudes than comparative belief relations.

we do so by specifying the net effect $\mathfrak{g}_i$ that the gamble has on our agent's level of total wealth in world w_i. For example, suppose you let £100 ride on red at the roulette table. Let $w_1, \ldots, w_i$ be the worlds in which the ball lands on red (you net £100), and $w_{i+1}, \ldots, w_n$ be the worlds in which it does not (you net −£100). Then we model your gamble as follows:

$$\mathcal{G} = \left\langle \underbrace{100, \ldots, 100}_{i \text{ times}}, \underbrace{-100, \ldots, -100}_{(n\text{-}i) \text{ times}} \right\rangle.$$

For any proposition $X \in \mathcal{F}$, we model a unit gamble on X by the characteristic vector $x = \langle \mathbf{x}_1, \ldots, \mathbf{x}_n \rangle$ of X, *i.e.*, the vector with $\mathbf{x}_i = 1$ if $w_i \in X$ and $\mathbf{x}_i = 0$ if $w_i \notin X$. And for any $\mathbf{a} \in \mathbb{R}$, we model the "constant gamble" that pays out £$\mathbf{a}$ in every world by the constant vector $a = \langle \mathbf{a}, \ldots, \mathbf{a} \rangle$.

Following Walley (1991), we can specify an agent's fair buying and selling prices using sets of *almost-desirable gambles*. Say that a gamble $\mathcal{G}$ is *almost desirable for an agent* iff she weakly prefers $\mathcal{G}$ to $\langle 0, \ldots, 0 \rangle$, *i.e.*, the status quo. Let $\mathbb{D} \subseteq \mathbb{R}^n$ be the set of gambles that she finds almost desirable.

Now we can specify her fair buying and selling price for $\mathcal{G}$ ($\mathcal{B}(\mathcal{G})$ and $\mathcal{S}(\mathcal{G})$, respectively) in terms of $\mathbb{D}$. Let

$$\mathcal{B}(\mathcal{G}) = \sup \left\{ \mathbf{a} \mid \mathcal{G} - a \in \mathbb{D} \right\}.$$

Taking the gamble $\mathcal{G} - a$ is equivalent to paying £$\mathbf{a}$ for $\mathcal{G}$. So $\mathcal{B}(\mathcal{G})$ is the largest amount that she could pay for $\mathcal{G}$ while leaving herself in a position that she weakly prefers to the status quo, *i.e.*, her fair buying price for $\mathcal{G}$. Likewise, let

$$\mathcal{S}(\mathcal{G}) = \inf \left\{ \mathbf{a} \mid a - \mathcal{G} \in \mathbb{D} \right\}.$$

Taking the gamble $a - \mathcal{G}$ is equivalent to receiving £$\mathbf{a}$ and shelling out $\mathcal{G}$'s payoff. So $\mathcal{S}(\mathcal{G})$ is the smallest amount that someone else would have to pay her in exchange for $\mathcal{G}$ while leaving herself in a position that she weakly prefers to the status quo, *i.e.*, her fair selling price for $\mathcal{G}$.

Talk of *both* fair buying *and* fair selling prices is actually a bit redundant. Note that

$$\begin{aligned} -\mathcal{B}(-\mathcal{G}) &= -\sup \left\{ \mathfrak{a} \mid -\mathcal{G} - a \in \mathbb{D} \right\} \\ &= \inf \left\{ -\mathbf{a} \mid -\mathcal{G} - a \in \mathbb{D} \right\} \\ &= \inf \left\{ \mathbf{a} \mid -\mathcal{G} + a \in \mathbb{D} \right\} \\ &= \mathcal{S}(\mathcal{G}). \end{aligned}$$

Taking $-\mathcal{G}$ from someone (*they* shell out $-\mathcal{G}$'s payoff to *you*) is nothing more than you offering $\mathcal{G}$ to them (*you* shell out $\mathcal{G}$'s payoff to *them*). And

paying a *negative* amount to someone for some good is really nothing more than them paying you a *positive* amount (and vice versa: taking a negative amount is nothing more than you paying a positive amount). The smaller the positive amount that they pay you, the bigger the negative amount you pay them. So the negative of the biggest amount that you would pay to take $-\mathcal{G}$, *i.e.*, $-\mathcal{B}(-\mathcal{G})$, is just another way of describing the smallest amount that you would need to be paid to offer $\mathcal{G}$.

We will just talk of your fair buying prices henceforth. But these really capture both your fair buying and selling prices.

Say that a set $\mathcal{E}$ of real-valued functions $e : \mathbb{R}^n \to \mathbb{R}$ *encodes your fair buying prices* iff its *lower envelope* for $\mathcal{G}$,

$$\underline{\mathcal{E}}[\mathcal{G}] = \inf\left\{e(\mathcal{G}) \mid e \in \mathcal{E}\right\},$$

is equal to $\mathcal{B}(\mathcal{G})$ when $\mathcal{B}(\mathcal{G})$ is defined, and is undefined when it is not. Say that a set of probability functions $\mathcal{C}$ encodes your fair buying prices just in case its corresponding set of expectation operators $\mathcal{E}_{\mathcal{C}} = \left\{E_c \mid c \in \mathcal{C}\right\}$ does so:

$$\underline{\mathcal{E}_{\mathcal{C}}}[\mathcal{G}] = \inf\left\{E_c[\mathcal{G}] \mid c \in \mathcal{C}\right\}.$$

Finally, say that your fair buying prices are *probabilistic* iff some set of probability functions encodes them.

How does the decision-theoretic account answer the characterisation question? When exactly *is* there a real-valued function c (or a set of such functions $\mathcal{C}$) that encodes your fair buying and selling prices? Answer: always.

Say that a real-valued function $e : \mathbb{R}^n \to \mathbb{R}$ *dominates your fair buying prices* iff $e(\mathcal{G}) \geq \mathcal{B}(\mathcal{G})$ whenever $\mathcal{B}(\mathcal{G})$ is defined. Let $\mathcal{E}^*$ be the set of real-valued functions that dominate your fair buying prices, *i.e.*,

$$\mathcal{E}^* = \left\{e \mid e(\mathcal{G}) \geq \mathcal{B}(\mathcal{G}) \text{ if } \mathcal{B}(\mathcal{G}) \text{ is defined}\right\}.$$

Then $\mathcal{E}^*$ encodes your fair buying prices, whatever they are. Hence $\mathcal{E}^*$ counts as "your credal state" according to the decision-theoretic view.

So there are no demanding constraints that an agent must satisfy in order to have credences, on this view. Having credences is dead easy. And clearly it is perfectly possible to have non-probabilistic credences.

So much for the characterisation question. How does the decision-theoretic account answer the normative question? Why should we expect *rational* agents to have *probabilistic* credences?

The story here is considerably more tricky. One might expect standard Dutch book arguments to provide an answer. De Finetti (1964) shows that for a specific sort of agent—one whose fair buying prices are equal to her fair selling prices, *i.e.*, $\mathcal{B}(\mathcal{G}) = \mathcal{S}(\mathcal{G})$—having non-probabilistic fair buying prices renders you Dutch bookable (susceptible to sure loss at

the hands of a clever bettor). One can see essentially the same result by considering (Walley, 1991, 3.3.3a). Walley shows that an agent's fair buying prices are not Dutch bookable (avoid sure loss) iff they are dominated by the expectation operator of some probability function. And in the special case under consideration—fair buying prices equal fair selling prices—one's fair buying prices are dominated in this way just in case they are probabilistic, *i.e.*, encoded by some set of probabilities $\mathcal{C}$. The upshot: in this special case—fair buying prices equal fair selling prices—an agent is not Dutch bookable (avoids sure loss) iff there is some set of probabilities $\mathcal{C}$ that encodes her fair buying prices, and hence counts as "her credal state." So if rationality requires avoiding sure loss, then we have good reason to expect this very special kind of agent to have probabilistic credences.

You might hope, then, that such a Dutch book argument could show quite generally that rational agents have probabilistic credences. But your hopes would be in vain. An agent avoids sure loss iff there is some set of probabilities $\mathcal{C}$ whose expectations for gambles uniformly dominate her fair buying prices for those gambles, *i.e.*, $E_c[\mathcal{G}] \geq \mathcal{B}(\mathcal{G})$ for all $c \in \mathcal{C}$ and all gambles $\mathcal{G}$. When an agent's fair buying and selling prices come apart, this can happen even when there is no set of probabilities $\mathcal{C}^*$ that actually *encodes* her fair buying prices.[19] Bottom line: non-Dutch-book-ability (avoiding sure loss) does not require having probabilistic credences.

Having probabilistic credences, in the decision-theoretic sense (*i.e.*, some set of probabilities that encodes your fair buying prices), is equivalent to something stronger than non-Dutch-book-ability—what Walley calls "coherence." Your fair buying prices are *coherent* iff they satisfy the following axioms.

1. Accept Sure Gains. $\mathcal{B}(\mathcal{G}) \geq \inf \mathcal{G}$.
2. Homogeneity. $\mathcal{B}(\lambda\mathcal{G}) = \lambda\mathcal{B}(\mathcal{G})$ for $\lambda \geq 0$.
3. Superlinearity. $\mathcal{B}(\mathcal{G} + \mathcal{G}^*) \geq \mathcal{B}(\mathcal{G}) + \mathcal{B}(\mathcal{G}^*)$.

Axiom 1 forbids you from paying at most £1 for $\mathcal{G}$ when $\mathcal{G}$ is guaranteed to payoff either £2, £3, or £4, for example. It says that your maximum buying price for $\mathcal{G}$ must be at least £2. Axiom 2 says that your fair buying price for a gamble $\mathcal{G}$ that is guaranteed to pay 2 (or 10, or 58.97) times another gamble $\mathcal{G}^*$ should be 2 (or 10, or 58.97) times your fair buying price for $\mathcal{G}^*$. Axiom 3 says that your fair buying price for a package of bets

19 Consider, for example, an agent whose fair buying price for any gamble $\mathcal{G}$ is $\inf \mathcal{G} - \epsilon$. For any non-constant $\mathcal{G}$, $\mathcal{B}(\mathcal{G}) = \inf \mathcal{G} - \epsilon < \sup \mathcal{G} - \epsilon = -\inf -\mathcal{G} - \epsilon < -\inf -\mathcal{G} + \epsilon = \mathcal{S}(\mathcal{G})$. But clearly $\mathcal{B}[\mathcal{G}] < \underline{\mathcal{E}_{\mathcal{C}}}[\mathcal{G}]$ for any set of probability functions $\mathcal{C}$. So the lower envelope of $\mathcal{E}_{\mathcal{C}}$ dominates her fair buying prices. Hence she avoids sure loss. But no such $\mathcal{C}$ *encodes* her fair buying prices.

should be at least as great as the sum of your fair buying prices for each of the bets in the package.

To reiterate: coherence is strictly stronger than avoiding sure loss. Walley (1991, Section 2.4) provides examples of fair buying prices that avoid sure loss (are not Dutch bookable), but nevertheless are not coherent. (Every coherent set of fair buying prices, in contrast, avoids sure loss.) So Dutch book or sure loss considerations do not give us good reason to think that, quite generally, rational agents have probabilistic credences.

All is not lost, though. Even if Dutch books arguments don't do the trick, another argument might. For example, in the spirit of Icard (2016) and Fishburn (1986, p. 338), we might propose constraints of rationality governing how one's comparative beliefs and preferences, or judgments of almost-desirability, ought to hang together. In particular, we might suggest that the set $\mathbb{D}$ of gambles that an agent finds almost desirable (*i.e.*, that she weakly prefers to the status quo) ought to be exactly the set $\mathcal{D}$ of gambles that are almost desirable relative to her comparative belief relation $\succeq$.

Belief-Preference Coherence. $\mathbb{D} = \mathcal{D}$.

Recall, a gamble $\mathcal{G}$ is almost desirable relative to $\succeq$ iff it is a non-negative linear combination of components

$$(X_1 - Y_1), \ldots, (X_n - Y_n)$$

which are such that $X_i \succeq Y_i$. $\mathcal{G}$ is a non-negative linear combination of $(X_1 - Y_1), \ldots, (X_n - Y_n)$ just in case

$$\mathcal{G} = \sum_i \lambda_i (X_i - Y_i)$$

for some $\lambda_1, \ldots, \lambda_n \geq 0$.

The basic thought here is this. $X_i - Y_i$ is the gamble that pays out £1 if X_i is true and $-$£1 if Y_i is true. You ought to weakly prefer this to the status quo iff you are at least as confident that X_i is true as Y_i. Moreover, you ought to think that any package of such bets, even if their stakes are scaled up or down by a positive constant, is almost-desirable; you ought to weakly prefer it to the status quo. And nothing more. Your comparative beliefs give you no reason to determinately prefer any other gamble to the status quo.

Now suppose that rationality not only demands comparative beliefs and preferences hang together as per Belief-Preference Coherence, but that it also demands that comparative beliefs on their own satisfy the Generalised Finite-Cancellation axiom.

Generalised Finite-Cancellation. If

$$X_1 + \ldots + X_n + \underbrace{A + \ldots + A}_{k \text{ times}} = Y_1 + \ldots + Y_n + \underbrace{B + \ldots + B}_{k \text{ times}}$$

and $X_i \succeq Y_i$ for all $i \leq n$, then $A \preceq B$.

Perhaps pragmatic considerations other than Dutch book or sure loss considerations establish that rational comparative beliefs satisfy GFC.[20] Or perhaps *epistemic* considerations establish this. Perhaps, for example, comparative beliefs that satisfy GFC epistemic-utility-dominate ones that do not, or something of the sort. For now, let's just leave an IOU for the justification of GFC.

Supposing that rational comparative beliefs satisfy GFC, we can now provide some reason to think that *quite generally* rational agents have probabilistic credences.

1. *Coherence Constraints*. Any rational agent's comparative beliefs satisfy GFC. Moreover, her comparative beliefs and preferences, or judgments of almost-desirability, jointly satisfy Belief-Preference Coherence.

2. *Theory of Credence*. A set of real-valued functions $\mathcal{C}$ count as "your credal state" just in case they encode your fair buying prices.

3. *IP-Representability Theorem*. Relation $\succeq$ satisfies GFC iff $\succeq$ is IP-representable, *i.e.*, $\succeq$ fully agrees with some set of probability functions $\mathcal{C}$.

4. *Bridge Theorem*. If $\succeq$ is IP-representable and satisfies Belief-Preference Coherence, then the maximal set of probability functions $\mathcal{C}^*$ that fully agrees with $\succeq$ also encodes your fair buying prices, and hence counts as "your credal state." (*Premise 2, Appendix*)

C. *Probabilism*. Any rational agent has probabilistic credences. (*From 1, 3, and 4*)

Does this argument trivialise probabilism? Of course not. It relies on the decision-theoretic account of credence—a substantive, highly non-trivial thesis. Moreover, even if we simply grant the decision-theoretic account of credence, it is no trivial consequence that rational agents have probabilistic credences. Having credences is easy. You have credences whatever your fair-buying prices are. But having *probabilistic* credences requires satisfying some demanding axioms (Belief-Preference Coherence, GFC). Establishing that these axioms encode genuine constraints of rationality is non-trivial. As a result, establishing probabilism is non-trivial.

20 Icard (2016) shows that an agent who satisfies Belief-Preference Coherence avoids sure loss iff her comparative beliefs are strongly representable. Strong representability is weaker than GFC. So we need something other than sure loss considerations to show that rational comparative beliefs satisfy GFC.

You might be concerned that our little argument only establishes half of probabilism. It shows that rational agents have probabilistic credences. But it does *not* show that rational agents have *only* probabilistic credences. Indeed, it cannot do so. We gave a recipe earlier for constructing a credal state for any agent. Just take the set of real-valued functions that dominate her fair buying prices. This set will encode those prices, and so count as "her credal state," on the decision-theoretic view. But it is not a set of probability functions.

But this converse thesis—*viz.*, that for any rational agent, no set of real-valued functions with non-probabilistic members counts as "her credal state"—is not theoretically interesting, on the decision-theoretic view. The benefits of having probabilistic credences (avoiding sure loss, coherence) accrue to any agent whose fair buying prices are encoded by some set of probabilities. Whether or not some non-probabilistic set encodes them as well is neither here nor there. Nothing of theoretical import hangs on it.

Finally, you might once again complain that the decision-theoretic account presupposes that rational agents only have comparative beliefs; no additional modes or types of doxastic judgment. But this stock of basic doxastic attitudes is too sparse. It is insufficient to explain everything about choice and inference that we would like to explain. So arguments that presuppose it are weak.

A similar response to the one in Section 5 will suffice. There is good reason to think this criticism lacks punch. And even if you buy the criticism, it is not an objection to the decision-theoretic account of credence *per se*. It is only an objection to the *unary* variant of this account. A pluralist faces no such problems. Of course, a pluralist must say more about how other types of doxastic attitudes—not just comparative beliefs—ought to hang together with judgments of almost-desirability. In addition, she must provide a more sophisticated IP-representability theorem and bridge theorem. But these are not in-principle problems. They are requests to cash in an IOU.

What does the scorecard look like? Whether the decision-theoretic account provides a compelling answer to the normative question depends in part on whether those IOUs can be replaced by theorems. But there is no special reason to think this task cannot be done. In addition, epistemic utility theorists, *e.g.*, Joyce (1998, 2009) and Pettigrew (2016), worry that this story provides an incomplete picture of our reasons to have probabilistic credences. A complete picture would provide a purely *epistemic* rationale for having imprecise credences.[21] Nevertheless, some form of the argument presented here might help illuminate *some* of our reasons for having probabilistic credences.

21 Epistemic utility theorists get off the boat early by rejecting the decision-theoretic account of credence.

7 EVALUATING THE EPISTEMIC INTERPRETIVIST VIEW

On the epistemic interpretivist view, a function (or set of functions) counts as "your credal state" just in case it encodes truth-value estimates that best rationalise or make sense of your comparative beliefs, understood as irreducibly-doxastic attitudes.

More formally, a function $c : \mathcal{F} \to \mathbb{R}$ (or a set of such functions $\mathcal{C}$) counts as "your credal state" iff its truth-value estimates best rationalise your comparative beliefs $\succeq$ (or on the pluralist version: your comparative and qualitative opinions more generally). For c (or $\mathcal{C}$) to *best rationalise* $\succeq$, it must satisfy two conditions: (i) c must recommend $\succeq$ as strongly as possible, so that no other c^* (or set $\mathcal{C}^*$) recommends $\succeq$ to a higher degree; (ii) c must be closer to rational (closer to the set $\mathcal{T}$ of rational assignments of truth-value estimates) than any other c^* that recommends $\succeq$ as strongly as possible—this ensures that c provides the highest quality recommendation possible.

To illustrate this view, consider two concrete cases. In case 1, you have comparative beliefs over propositions in the following Boolean algebra:

$$\mathcal{F} = \{\varnothing, \{w_1\}, \{w_2\}, \{w_1, w_2\}\}.$$

In particular, your comparative beliefs are given by:

$$\varnothing \prec \{w_1\} \prec \{w_2\} \prec \{w_1, w_2\}.$$

Question: which function c (or set $\mathcal{C}$) encodes truth-value estimates (or constraints on estimates) that best rationalise your comparative beliefs $\succeq$, and hence counts as "your credal state," according to epistemic interpretivism? To provide a concrete answer, we will need to make a few substantive assumptions about recommendation, rationality, and the like.

In Section 3.3, we outlined three accounts of recommendation—the metaphysical, normative, and epistemic utility accounts—which aim to explain when and how an assignment of truth-value estimates c (or set of assignments $\mathcal{C}$) recommends comparative beliefs $\succeq$ more or less strongly. For simplicity, we will assume the metaphysical account in what follows. Recall, on the metaphysical account, c (or $\mathcal{C}$) recommends $\succeq$ as strongly as possible just in case explicitly judging c (or $\mathcal{C}$) to encode the best (constraints on) estimates of truth-values *metaphysically entails* having precisely the comparative beliefs given by $\succeq$. Moreover, we will assume that judging c to be best entails having a specific set of comparative beliefs, *viz.*, the comparative beliefs $\succeq_c$ that *fully agree* with c. Likewise, we will assume that judging a *set* of assignments $\mathcal{C}$ (constraints on truth-value estimates) to be best entails having the comparative beliefs $\succeq_{\mathcal{C}}$ that fully agree with $\mathcal{C}$. Finally, we will assume that the set $\mathcal{T}$ of rational assignments

of truth-value estimates is fairly inclusive. In particular, we will adopt the radical subjective Bayesian assumption that $\mathcal{T}$ is the set of probability functions. Likewise, the rational *constraints* on truth-value estimates (imprecise truth-value estimates) are just the *sets* of probability functions.

Back to our question then. Which function c (or set $\mathcal{C}$) encodes (constraints on) truth-value estimates that best rationalise your comparative beliefs $\succeq$, and hence counts as "your credal state," according to epistemic interpretivism?

Firstly, note that any probability function c with $0.5 < c(\{w_2\}) < 1$ is such that

$$c(\varnothing) < c(\{w_1\}) < c(\{w_2\}) < c(\{w_1, w_2\}),$$

and hence fully agrees with $\succeq$. So any such c recommends $\succeq$ as strongly as possible, on the metaphysical account. Secondly, note that each such a c is probabilistically coherent. So it is maximally rational, according to our radical subjective Bayesian assumption. Hence any of these probability functions counts as "your credal state," according to epistemic interpretivism.

Similarly, note that any *set* of probability functions $\mathcal{C}$ with $0.5 < c(\{w_2\}) < 1$ for all $c \in \mathcal{C}$ fully agrees with $\succeq$. So any such $\mathcal{C}$ recommends $\succeq$ as strongly as possible, on the metaphysical account. And each such $\mathcal{C}$ is maximally rational, according to our radical subjective Bayesian assumption. So any of these sets of probability functions $\mathcal{C}$ counts as "your credal state," according to epistemic interpretivism.

On a pluralist version of epistemic interpretivism, according to which your credal state does not simply best rationalise your *comparative beliefs*, but rather best rationalises a *broader set of comparative and qualitative opinions*, we might be able to winnow down the set of candidate credal states more than this. Likewise, on a more sophisticated version of epistemic interpretivism according to which your credal state does not simply best rationalise your *current* opinions, but rather is part of a package that best rationalises your opinions *over time*, we might be able to winnow down this set even further. But as it stands, epistemic interpretivism allows for a lot of slack in what counts as your credences. It allows for a great many ties between maximally eligible credal states. This, however, is as it should be. Given what epistemic interpretivists take the principle theoretical role of credal states to be—their job is to best rationalise your comparative and qualitative opinions—and given how few comparative beliefs you actually have, various sets of truth-value estimates (and constraints on such estimates) play that role equally well.

Consider one more concrete case. In case 2, your comparative beliefs are given by:

$$\varnothing \approx \{w_1\} \approx \{w_2\} \prec \{w_1, w_2\}.$$

Which function c (or set $\mathcal{C}$) encodes (constraints on) truth-value estimates that best rationalise your comparative beliefs $\succeq$, and hence counts as "your credal state," according to epistemic interpretivism?

To answer this question, first note that your comparative belief relation is clearly not probabilistically representable. No probability function fully agrees with $\succeq$. Nor is it imprecisely representable. So no probability function c, or set of probability functions $\mathcal{C}$, recommends $\succeq$ as strongly as possible. But there *are* non-probabilistic assignments of truth-value estimates that fully agree with $\succeq$, and hence recommend it as strongly as possible. In particular, any c with $c(\varnothing) = c(\{w_1\}) = c(\{w_2\}) = x$, $c(\{w_1, w_2\}) = y$ and $x < y$ is such that

$$c(\varnothing) = c(\{w_1\}) = c(\{w_2\}) < c(\{w_1, w_2\}),$$

and hence fully agrees with $\succeq$. So any such c recommends $\succeq$ as strongly as possible, on the metaphysical account.

But which one of these provides the *highest quality* recommendation of $\succeq$? That is, which one is closest to rational? Since the rational assignments of truth-value estimates are exactly the probability functions (by assumption), the question really is: which of these assignments of truth-value estimates is closest to probabilistically coherent (*i.e.*, closest to the set of all probability functions)?

To answer this question, we need to plump for some measure of "closeness" or "proximity." One natural choice: squared Euclidean distance. Squared Euclidean distance is the "Bregman divergence" generated by a very popular measure of accuracy, *viz.*, the Brier Score. It captures one attractive way of thinking about how close two sets of truth-value estimates are in terms of how similar their degree of accuracy is expected to be. If we plump for squared Euclidean distance as our measure of "closeness" or "proximity," then the assignment of truth-value estimates that best rationalises your comparative beliefs is given by:

$$\begin{aligned} c(\varnothing) = c(\{w_1\}) = c(\{w_2\}) &= 1/3, \\ c(\{w_1, w_2\}) &= 1. \end{aligned}$$

So this function c counts as "your credal state," according to epistemic interpretivism. What's more, assuming that supersets $\{c, c^*, \ldots\}$ of $\{c\}$ with the c^* all strictly less rational than c are *themselves* less rational than $\{c\}$, we have that c is the *unique* function that counts as "your credal state."

So much for the nuts and bolts of epistemic interpretivism. What about the characterisation question? When exactly *is* there a real-valued function c (or a set of such functions $\mathcal{C}$) that encodes truth-value estimates which best rationalise your comparative beliefs, and hence counts as "your credal state"?

It depends. It depends which account of recommendation you plump for. It depends which assignments of truth-value estimates you count as rational. It depends how you measure proximity to the set of rational assignments of truth-value estimates. Different commitments on these fronts will yield different answers to the question of which truth-value estimates rationalise which comparative beliefs. But for concreteness, suppose we stick with the metaphysical account and radical subjective Bayesian assumption we made earlier. In that case, whenever your comparative beliefs $\succeq$ fully agree with a real-valued function c, that function will recommend $\succeq$ as strongly as possible. So some such function will *best* rationalise $\succeq$ (or near enough).[22] And a comparative belief relation $\succeq$ fully agrees with a real-valued function c if and only if $\succeq$ satisfies Transitivity and Totality (Krantz et al., 1971, Theorem 1). So some real-valued function c best rationalises $\succeq$, and hence counts as "your credal state," whenever $\succeq$ satisfies Transitivity and Totality.

Conversely, if $\succeq$ violates either Transitivity or Totality, then no single real-valued function fully agrees with $\succeq$. Hence no single function recommends $\succeq$ as strongly as possible. If there is some *set* of real-valued functions that fully agrees with $\succeq$, then that set recommends $\succeq$ more strongly than any single function, on the metaphysical view. So while no single function will count as "your credal state," some set (the most rational one that fully agrees with $\succeq$) will do so (or near enough). It is an interesting question when exactly there is such a set of functions (not necessarily probability functions) that fully agrees with $\succeq$.

The overarching lesson here is a familiar one. Your comparative beliefs need not satisfy Scott's axioms—axioms which might seem rather demanding on their face—in order to count as having credences, according to epistemic interpretivism. Given our working assumptions (the metaphysical account, etc.), such axioms encode necessary and sufficient conditions for having precise *probabilistic* credences. But it is perfectly possible to have non-probabilistic credences. As we saw in case 2, if your comparative beliefs are given by

$$\varnothing \approx \{w_1\} \approx \{w_2\} \prec \{w_1, w_2\},$$

then the following non-probabilistic real-valued function c counts as "your credal state," according to the epistemic interpretivist:

$$c(\varnothing) = c(\{w_1\}) = c(\{w_2\}) = 1/3,$$
$$c(\{w_1, w_2\}) = 1.$$

Your comparative beliefs only need to satisfy relatively weak constraints to count as having credences *tout court*.

22 For any $\epsilon > 0$, we can pick some c that recommends $\succeq$ such that any other c^* that does so as well is no more than ϵ-closer to coherent.

On to the normative question then. Why should we expect *rational* agents to have *probabilistic* credences?

The epistemic interpretivist's answer to the normative question is much more complicated than our previous accounts' answers. For the epistemic interpretivist, the normative question breaks into (at least) three subquestions.

1. Why should we expect that for any rational agent, there is some probability function that fully agrees with her comparative beliefs?

2. Why should we expect it to be the case that the rational assignments of truth-value estimates are exactly the probability functions?

3. Why should we expect any probability function that fully agrees with comparative beliefs $\succeq$ to recommend $\succeq$ as strongly as possible?

Scott's theorem tells us that comparative beliefs $\succeq$ fully agree with some probability function c iff $\succeq$ satisfies Non-Triviality, Non-negativity, Totality, and Isovalence. So answering question 1 amounts to defending the claim that rational comparative beliefs satisfy Non-Triviality, Non-negativity, Totality, and Isovalence. Of course, you might doubt whether "structural axioms" like Totality encode genuine constraints of rationality. In that case, the epistemic interpretivist might defend the Generalised Finite-Cancellation axiom and argue that rational agents have *imprecise* probabilistic credences.

To answer question 2, the epistemic interpretivist might rely on results from epistemic utility theory. For example, she might endorse an austere conception of rationality, according to which rationality requires you to prefer one assignment of truth-value estimates b to another c just in case b is *guaranteed* to be more accurate than c. Then she might appeal to Joyce (1998, 2009), Predd et al. (2009), Schervish, Seidenfeld, and Kadane (2009), and Pettigrew (2016), who show that any non-probabilistic b is accuracy-dominated by some probabilistic c, *i.e.*, the truth-value estimates encoded by c are guaranteed to be more accurate than those encoded by b. No probabilistic c, in contrast, is even weakly accuracy-dominated. So the probability functions are exactly the rational assignments of truth-value estimates (not rationally dispreferred to any other assignment), on the austere conception of rationality.

Finally, to answer question 3, the epistemic interpretivist must defend an account of recommendation and various auxiliary claims. For example, a proponent of the epistemic utility account must defend various substantive claims about the nature of epistemic value. In particular, she must defend the claim that on any reasonable measure of epistemic utility for comparative beliefs, all probability functions expect the comparative belief relations that fully agree with them to have maximal epistemic utility.

We will not provide answers to questions 1–3 here. But if the epistemic interpretivist can answer them satisfactorily, then she can offer up something like the following argument for probabilism.

1. *Coherence Constraints*. Any rational agent's comparative belief relation $\succeq$ satisfies Non-Triviality, Non-negativity, Totality, and Isovalence.

2. *Theory of Credence*. An assignment of truth-value estimates c counts as "your credal state" iff it best rationalises your comparative beliefs $\succeq$. Moreover, c best rationalises $\succeq$ iff (i) it recommends $\succeq$ as strongly as possible, and (ii) c is itself closer to rational than any other c^* that recommends $\succeq$ as strongly as possible.

3. *Accuracy Argument*. An assignment of truth-value estimates c is rational iff c is a probability function. (*Accuracy-dominance theorem, austere conception of rationality*)

4. *Scott's Theorem*. Relation $\succeq$ satisfies Non-Triviality, Non-negativity, Totality, and Isovalence if and only if $\succeq$ fully agrees with some probability function c.

5. *Theory of Recommendation*. An assignment of truth-value estimates c recommends $\succeq$ to degree k iff the maximally rational extension of c to $\mathbb{Q}$, est_c, is such that $est_c(\mathcal{U}(\succeq)) = k$.

6. *Bridge Theorem I*. If $\succeq$ fully agrees with a probability function c, then c recommends $\succeq$ as strongly as possible. (*Premise 5, austere conception of rationality, auxiliary assumptions about epistemic utility*)

7. *Bridge Theorem II*. If $\succeq$ fully agrees with a probability function c, then c not only recommends $\succeq$ as strongly as possible, but is also rational. Hence c counts as "your credal state." (*From 2, 3 and 6*)

C. *Probabilism*. Any rational agent has probabilistic credences. (*From 1, 4 and 7*)

Does this argument trivialise probabilism? Obviously not! It is positively baroque! And unlike the measurement-theoretic and decision-theoretic accounts, the epistemic interpretivist can plausibly argue that rational agents have *only* probabilistic credences. To see this, suppose that your comparative beliefs are given by the rational comparative belief relation $\succeq$. By premise 1, $\succeq$ satisfies Non-Triviality, Non-negativity, Totality, and Isovalence. So by premises 4 and 7, there is some probability function c that fully agrees with $\succeq$, and hence best rationalises $\succeq$. Now, while many non-probabilistic assignments of truth-value estimates b will also fully agree with $\succeq$, and hence recommend $\succeq$ as strongly as possible, none will

provide as high a quality recommendation for $\succeq$. Hence none will *best rationalise* $\succeq$. The reason: any non-probabilistic b is accuracy-dominated by some probabilistic c, and hence less rational than c (premise 3). So b provides a weaker rationale for $\succeq$ than c. Hence b fails to count as "your credal state," according to the epistemic interpretivist.

Finally, you might return to the complaint that this account presupposes that rational agents only have comparative beliefs—an overly austere, explanatorily deficient stock of doxastic attitudes. But again, this is only an objection to a *unary* version of epistemic interpretivism (to the extent that it has bite at all). It has no force against pluralist variants. Of course, pluralist variants face a range of unanswered questions. When exactly does a set of truth-value estimates rationalise a more comprehensive system of doxastic attitudes (comparative beliefs, full beliefs, opinions about evidential and causal dependence and independence, etc.)? Why should we expect reasonable measures of epistemic utility for these more comprehensive systems to satisfy a suitably generalised version of strict propriety? And when exactly are these more comprehensive systems of doxastic attitudes probabilistically representable? But these are new research questions bubbling up on the boundary of an active research programme. There is no principled reason for thinking that they do not have adequate answers.

Before wrapping up, it is worth highlighting one additional virtue of epistemic intepretivism. At the outset, we mentioned a Bayesian platitude about credences. Joyce puts the platitude as follows: "in the probabilistic tradition, *the* defining fact about credences is that they are used to estimate quantities that depend on truth-values" (Joyce, 2009, pp. 268–9). A rational agent's credences determine *expectations* of measurable quantities—quantities like the size of the deficit 10 years hence, or the utility of an outcome—which capture her *best estimates* of those quantities. Those best estimates, in turn, typically *rationalise* or make sense of her *evaluative attitudes* and *choice behaviour*.

Shorter: credences capture estimates that provide rationalising explanations.

Epistemic interpretivism is much better positioned than the measurement-theoretic or decision-theoretic views to vindicate this platitude. On the measurement-theoretic view, credence functions are just mappings from propositions to real numbers that preserve the structure of your comparative beliefs. They do not encode estimates, or any other quantity that might plausibly play a role in rationalising your doxastic attitudes, evaluative attitudes, or choice behaviour. On the decision-theoretic view, credence functions are just numerical systems that encode your fair buying and selling prices. But having fair buying and selling prices is *nothing over*

and above having certain kinds of preferences. So they are hardly fit to *rationalise* preferences.[23]

In contrast, epistemic interpretivism directly identifies your credence function c with the assignment of truth-value estimates that best rationalises your comparative beliefs $\succeq$. And however we spell out what it is for truth-value estimates to best rationalise comparative beliefs, we can apply a similar story to preferences and choice behaviour. Consider, for example, the epistemic utility account of recommendation from Section 3.3. On this view, we start with c, and we add estimates of other measurable quantities $\mathcal{Q}$ to the stock of truth-value estimates encoded by c in the most rational way possible. In particular, we add estimates of the epistemic utility of comparative belief relations. The larger the estimate of $\succeq$'s epistemic utility, the more strongly c recommends $\succeq$.

We can tell exactly the same story about how your credences rationalise preferences and choice behaviour. We start with the truth-value estimates c that best rationalise your comparative beliefs, and we add estimates of the value of *actions*, for example, in the most rational way possible. The larger the estimate of an action's value, the more strongly c recommends it. In turn, the more strongly it rationalises choosing that action.

The moral: epistemic interpretivism appears to have the resources to vindicate core tenets of Bayesianism that other accounts have trouble with.

8 CONCLUDING REMARKS

Many Bayesians take *comparative belief* to be crucial for spelling out what it is to have a degree of confidence, or degree of belief, or credence. And they typically appeal to *representation theorems* when answering foundational questions about credence. We have explored three different accounts—measurement-theoretic, decision-theoretic, and epistemic interpretivist—that utilise comparative beliefs and representation theorems in order to answer two such questions: *the characterisation question*, *i.e.*, when exactly an agent counts as having credences, and *the normative question*, *i.e.*, why we should expect rational agents to have probabilistic credences. Hájek (2009), Meacham and Weisberg (2011), and Titelbaum (2015) pose some pressing challenges to accounts of this sort: they make the bar for having credences so high that very few real agents clear it, they trivialise probabilism, and so on. But we found that suitably sophisticated versions of each of our three accounts handle these challenges fairly well. There is more work to be done in filling these accounts out. But wholesale scepticism about the role of comparative belief and representation theorems in providing an account of credence seems premature.

23 For a similar criticism of behaviourism, see Joyce (1999, Section 1.3).

9 APPENDIX

Choose any comparative belief structure $\langle \Omega, \mathcal{F}, \succeq \rangle$ with finite $\mathcal{F}$. Assume without loss of generality that $|\Omega| = n$.

Theorem 3 (Generalised Scott's Theorem) *Suppose $\succeq$ satisfies the following two conditions.*

1. NON-TRIVIALITY. $\Omega \succ \varnothing$.
2. NON-NEGATIVITY. $X \succeq \varnothing$.

Then the following two conditions are equivalent.

3. ISOVALENCE. *If $X_1 + \ldots + X_n = Y_1 + \ldots + Y_n$ and $X_i \succeq Y_i$ for all $i \leq n$, then $X_i \preceq Y_i$ for all $i \leq n$ as well.*
4. STRONG REPRESENTABILITY. *There exists a probability function $p : \mathcal{F} \to \mathbb{R}$ that strongly agrees with $\succeq$ in the sense that:*
   (i) $X \succeq Y \Rightarrow p(X) \geq p(Y)$,
   (ii) $X \succ Y \Rightarrow p(X) > p(Y)$.

Proof. Let

$$\mathbb{A} = \left\{ \sum_i \lambda_i (X_i - Y_i) \mid \lambda_i \geq 0 \text{ and } X_i \succeq Y_i \right\}$$

and

$$\mathbb{U} = \left\{ \sum_i \lambda_i (Y_i - X_i) \mid \lambda_i \geq 0, \sum_i \lambda_i = 1, \text{ and } X_i \succ Y_i \right\}.$$

First we will show that $\mathbb{A} \cap \mathbb{U} = \varnothing$ iff $\succeq$ satisfies Isovalence. Then we will show that if $\succeq$ satisfies Non-Triviality and Non-Negativity, then $\mathbb{A} \cap \mathbb{U} = \varnothing$ iff $\succeq$ is strongly representable.

Suppose that $\succeq$ satisfies Isovalence. So if

$$X_1 + \ldots + X_t = Y_1 + \ldots + Y_t$$

and $X_i \succeq Y_i$ for all $i \leq t$, then $X_i \preceq Y_i$ for all $i \leq t$ as well.

Suppose for reductio that $\mathbb{A} \cap \mathbb{U} \neq \varnothing$. So there is some $\mathcal{G} \in \mathbb{A} \cap \mathbb{U}$. Hence

$$\mathcal{G} = \sum_{i \leq m} \lambda_i (X_i - Y_i) = \sum_{i \leq k} \delta_i (B_i - A_i),$$

where $\lambda_i \geq 0$ and $X_i \succeq Y_i$ for all $i \leq m$; likewise $\delta_i \geq 0$, $\sum_i \delta_i = 1$, and $A_i \succ B_i$ for all $i \leq k$. So

$$\lambda_1 (X_1 - Y_1) + \ldots + \lambda_m (X_m - Y_m) + \delta_1 (A_1 - B_1) + \ldots + \delta_k (A_k - B_k) = 0.$$

Let

$$X_i = \left\langle x_1^i, \ldots, x_n^i \right\rangle.$$

Likewise for Y_i, A_i, and B_i. Then the equality above gives us a system of n homogenous linear equations with rational coefficients:

$$(x_1^1 - y_1^1)\lambda_1 + \ldots + (x_1^m - y_1^m)\lambda_m + (a_1^1 - b_1^1)\delta_1 + \ldots + (a_1^k - b_1^k)\delta_k = 0,$$
$$\vdots$$
$$(x_n^1 - y_n^1)\lambda_1 + \ldots + (x_n^m - y_n^m)\lambda_m + (a_n^1 - b_n^1)\delta_1 + \ldots + (a_n^k - b_n^k)\delta_k = 0.$$

Since this system of equations has rational coefficients, it has a rational solution if it has any solution, by Gauss' method. So we can rewrite

$$\lambda_1(X_1 - Y_1) + \ldots + \lambda_m(X_m - Y_m) + \delta_1(A_1 - B_1) + \ldots + \delta_k(A_k - B_k) = 0$$

as

$$\frac{\alpha_1}{\beta_1}(X_1 - Y_1) + \ldots + \frac{\alpha_m}{\beta_m}(X_m - Y_m) + \frac{\phi_1}{\psi_1}(A_1 - B_1) + \ldots + \frac{\phi_k}{\psi_k}(A_k - B_k) = 0.$$

Multiplying through and rearranging gives us

$$\begin{aligned}(\alpha_1\beta_2\ldots\beta_m\psi_1\ldots\psi_k)(X_1 - Y_1)&\\ + \ldots + (\alpha_m\beta_1\ldots\beta_{m-1}\psi_1\ldots\psi_k)(X_m - Y_m)&\\ + (\phi_1\psi_2\ldots\psi_k\beta_1\ldots\beta_m)(A_1 - B_1)&\\ + \ldots + (\phi_k\psi_1\ldots\psi_{k-1}\beta_1\ldots\beta_m)(A_k - B_k) = 0.&\end{aligned}$$

This in turn gives us

$$\begin{aligned}&(\alpha_1\beta_2\ldots\beta_m\psi_1\ldots\psi_k)X_1 + \ldots + (\alpha_m\beta_1\ldots\beta_{m-1}\psi_1\ldots\psi_k)X_m\\ &\quad + (\phi_1\psi_2\ldots\psi_k\beta_1\ldots\beta_m)A_1 + \ldots + (\phi_k\psi_1\ldots\psi_{k-1}\beta_1\ldots\beta_m)A_k\\ &\quad = (\alpha_1\beta_2\ldots\beta_m\psi_1\ldots\psi_k)Y_1 + \ldots + (\alpha_m\beta_1\ldots\beta_{m-1}\psi_1\ldots\psi_k)Y_m\\ &\qquad + (\phi_1\psi_2\ldots\psi_k\beta_1\ldots\beta_m)B_1 + \ldots + (\phi_k\psi_1\ldots\psi_{k-1}\beta_1\ldots\beta_m)B_k.\end{aligned}$$

But recall, $X_i \succeq Y_i$ for all $i \leq m$, and $A_i \succeq B_i$ for all $i \leq k$. So by Isovalence we must have $X_i \preceq Y_i$ for all $i \leq m$ and $A_i \preceq B_i$ for all $i \leq k$. But since $A_i \succ B_i$ for all $i \leq k$, we have $A \not\preceq B_i$ for all $i \leq k$. Contradiction.

Therefore $\mathbb{A} \cap \mathbb{U} = \varnothing$.

Conversely, suppose that $\mathbb{A} \cap \mathbb{U} = \varnothing$. Suppose for reductio that $\succeq$ violates Isovalence. So there are $X_1, \ldots, X_t, Y_1, \ldots, Y_t \in \mathcal{F}$ such that

$$X_1 + \ldots + X_t = Y_1 + \ldots + Y_t.$$

$X_i \succeq Y_i$ for all $i \leq t$, and $X_j \succ Y_j$ for some $j \leq t$. Assume without loss of generality that $X_i \approx Y_i$ for all $i \neq j$. Then

$$\sum_{i \neq j} X_i - Y_i = Y_j - X_j.$$

Let

$$\mathcal{G} = \sum_{i \neq j} X_i - Y_i = Y_j - X_j.$$

Then $\mathcal{G} \in \mathbb{A} \cap \mathbb{U}$. Contradiction.

Therefore $\succeq$ must satisfy Isovalence.

This establishes that $\mathbb{A} \cap \mathbb{U} = \varnothing$ iff $\succeq$ satisfies Isovalence. Now we will show that if $\succeq$ satisfies Non-Triviality and Non-Negativity, then $\mathbb{A} \cap \mathbb{U} = \varnothing$ iff $\succeq$ is strongly representable.

Suppose that $\succeq$ satisfies Non-Triviality and Non-Negativity. So $\Omega \succ \varnothing$ and $X \succeq \varnothing$ for all $X \in \mathcal{F}$.

Now suppose that $\mathbb{A} \cap \mathbb{U} = \varnothing$. Note that $\mathbb{A}$ is the closed, convex polyhedral cone generated by the set $\{X - Y \mid X \succeq Y\}$. Likewise, $\mathbb{U}$ is the convex hull of $\{Y - X \mid X \succ Y\}$—a closed and convex set. So the hyperplane separation theorem of Kuhn and Tucker (1956, p. 50) guarantees that there is a linear functional E that strictly separates $\mathbb{A}$ and $\mathbb{U}$ in the sense that

$$E[\mathcal{G}] \geq 0 \text{ for all } \mathcal{G} \in \mathbb{A},$$
$$E[\mathcal{G}^*] < 0 \text{ for all } \mathcal{G}^* \in \mathbb{U}.$$

Since $\Omega \succ \varnothing$, $\varnothing - \Omega = -\Omega \in \mathbb{U}$. Hence $E[-\Omega] < 0$, which is the case iff $E[\Omega] > 0$.

Since $X \succeq \varnothing$ for all $X \in \mathcal{F}$, $X - \varnothing = X \in \mathbb{A}$. Hence $E[X] \geq 0$.

Now let

$$p(X) = \frac{E[X]}{E[\Omega]}$$

for all $X \in \mathcal{F}$. Obviously p satisfies Normalization and Non-negativity, since

$$p(\Omega) = \frac{E[\Omega]}{E[\Omega]} = 1$$

and

$$p(X) \geq p(\varnothing) \text{ iff } \frac{E[X]}{E[\Omega]} \geq \frac{E[\varnothing]}{E[\Omega]} \text{ iff } E[X] \geq 0.$$

Moreover, if $X \cap Y = \varnothing$, then $X \cup Y = X + Y$. So

$$p(X \cup Y) = \frac{E[X + Y]}{E[\Omega]} = \frac{E[X]}{E[\Omega]} + \frac{E[Y]}{E[\Omega]} = p(X) + p(Y).$$

So p satisfies Finite Additivity. Hence p is a probability function. And it follows straightforwardly that p strongly agrees with $\succeq$:

$$\begin{aligned} X \succeq Y &\Rightarrow E[X - Y] \geq 0 \\ &\Leftrightarrow E[X] \geq E[Y] \\ &\Leftrightarrow p(X) \geq p(Y), \end{aligned}$$

and

$$\begin{aligned} X \succ Y &\Rightarrow E[Y - X] < 0 \\ &\Leftrightarrow E[X] > E[Y] \\ &\Leftrightarrow p(X) > p(Y). \end{aligned}$$

Conversely, suppose that $\succeq$ is strongly representable. So there is some probability function p such that strongly agrees with $\succeq$ in the sense that

(i) $X \succeq Y \Rightarrow p(X) \geq p(Y)$,

(ii) $X \succ Y \Rightarrow p(X) > p(Y)$.

Suppose for reductio that $\mathbb{A} \cap \mathbb{U} \neq \varnothing$. So there is some $\mathcal{G} \in \mathbb{A} \cap \mathbb{U}$. Hence

$$\mathcal{G} = \sum_{i \leq m} \lambda_i (X_i - Y_i) = \sum_{i \leq k} \delta_i (B_i - A_i),$$

where $\lambda_i \geq 0$ and $X_i \succeq Y_i$ for all $i \leq m$; likewise, $\delta_i \geq 0$, $\sum_i \delta_i = 1$, and $A_i \succ B_i$ for all $i \leq k$. Let

$$E_p[\mathcal{V}] = \sum_{w_i \in \Omega} p(w_i) v_i,$$

where $\Omega = \{w_1, \ldots, w_n\}$ and $\mathcal{V} =< v_1, \ldots, v_n >$. Once more let

$$X_i = \left\langle x_1^i, \ldots, x_n^i \right\rangle.$$

Likewise for Y_i, A_i, and B_i. Then

$$\begin{aligned} E_p[\mathcal{G}] &= \sum_{w_i \in \Omega} p(w_i) \left[\sum_{j \leq m} \lambda_j (x_i^j - y_i^j) \right] \\ &= \sum_{j \leq m} \lambda_j \left[\sum_{w_i \in \Omega} p(w_i)(x_i^j - y_i^j) \right] \\ &= \sum_{j \leq m} \lambda_j (p(X_j) - p(Y_j)). \end{aligned}$$

Since $X_j \succeq Y_j$ for all $j \leq m$, $p(X_j) \geq p(Y_j)$. Hence

$$E_p[\mathcal{G}] \geq 0.$$

But we also have

$$\begin{aligned} E_p[\mathcal{G}] &= \sum_{w_i \in \Omega} p(w_i) \left[\sum_{j \leq k} \lambda_j (b_i^j - a_i^j) \right] \\ &= \sum_{j \leq k} \lambda_j \left[\sum_{w_i \in \Omega} p(w_i)(b_i^j - a_i^j) \right] \\ &= \sum_{j \leq k} \lambda_j (p(B_j) - p(A_j)). \end{aligned}$$

Since $A_j \succ B_j$ for all $j \leq k$, $p(A_j) > p(B_j)$. Hence

$$E_p[\mathcal{G}] < 0.$$

Contradiction. Therefore $\mathbb{A} \cap \mathbb{U} = \varnothing$.

So far we have established that $\mathbb{A} \cap \mathbb{U} = \varnothing$ iff $\succeq$ satisfies Isovalence. Moreover, we have established that if $\succeq$ satisfies Non-Triviality and Non-Negativity, then $\mathbb{A} \cap \mathbb{U} = \varnothing$ iff $\succeq$ is strongly representable. This suffices to prove GST. ■

Theorem 4 *Suppose that $\succeq$ is IP-representable and satisfies Belief-Preference Coherence. Let $\mathcal{C}$ be the maximal set of probability functions $\mathcal{C}$ that fully agrees with $\succeq$. Let $\mathcal{E}_\mathcal{C} = \{E_c \mid c \in \mathcal{C}\}$. This $\mathcal{C}$ encodes your fair buying prices in the sense that*

$$\underline{\mathcal{E}_\mathcal{C}}[\mathcal{G}] = \inf\left\{E_c[\mathcal{G}] \mid c \in \mathcal{C}\right\}$$

is equal to your fair buying price for $\mathcal{G}$, $\mathcal{B}(\mathcal{G})$, when $\mathcal{B}(\mathcal{G})$ is defined, and is undefined when it is not.

Proof. Suppose that $\succeq$ is IP-representable. So there is a set of probability functions that fully agrees with it. Let $\mathcal{C}$ be the maximal set of probability functions $\mathcal{C}$ that fully agrees with $\succeq$. So

$$X \succeq Y \Leftrightarrow c(X) \geq c(Y) \text{ for all } c \in \mathcal{C}.$$

And if $\mathcal{C}^*$ fully agrees with $\succeq$, then $\mathcal{C}^* \subseteq \mathcal{C}$.

First, note that $\mathcal{C}$ must be the set $\mathcal{B}$ of all probability functions b that almost agree with $\succeq$:

$$\mathcal{B} = \left\{b \mid X \succeq Y \Rightarrow b(X) \geq b(Y)\right\}.$$

Obviously $\mathcal{C} \subseteq \mathcal{B}$. To see that $\mathcal{B} \subseteq \mathcal{C}$, choose $b \in \mathcal{B}$. Suppose for reductio that $b \notin \mathcal{C}$.

> *Case 1.* For all $X, Y \in \mathcal{F}$, if $c(X) \geq c(Y)$ for all $c \in \mathcal{C}$, then $b(X) \geq b(Y)$. In that case, $\mathcal{C}^* = \mathcal{C} \cup \{b\}$ fully agrees with $\succeq$. But then $\mathcal{C}$ is not maximal. Contradiction.
>
> *Case 2.* For some $X, Y \in \mathcal{F}$, $c(X) \geq c(Y)$ for all $c \in \mathcal{C}$, but $b(X) < b(Y)$. In that case, since $\mathcal{C}$ fully agrees with $\succeq$, $X \succeq Y$. But since b almost agrees with $\succeq$, this implies $b(X) \geq b(Y)$. Contradiction.

Hence $\mathcal{B} = \mathcal{C}$.

Second, note that since $\succeq$ is IP-representable, $\succeq$ satisfies Non-Triviality and Non-Negativity. That is, $\Omega \succ \varnothing$ and $X \succeq \varnothing$ for all $X \in \mathcal{F}$.

Now suppose that $\succeq$ also satisfies Belief-Preference Coherence. So the set $\mathcal{A}$ of gambles that our agent finds almost desirable (*i.e.*, that she weakly

prefers to the status quo) is exactly the set $\mathbb{A}$ of gambles that are almost desirable relative to $\succeq$:

$$\mathcal{A} = \mathbb{A} = \left\{ \sum_i \lambda_i (X_i - Y_i) \mid \lambda_i \geq 0 \text{ and } X_i \succeq Y_i \right\}.$$

Our agent's fair buying price for a gamble $\mathcal{G}$ is

$$\mathcal{B}(\mathcal{G}) = \sup \left\{ \mathfrak{a} \mid \mathcal{G} - a \in \mathcal{A} \right\}.$$

Our aim is to show that $\underline{\mathcal{E}_{\mathcal{C}}}[\mathcal{G}] = \mathcal{B}(\mathcal{G})$ if $\mathcal{B}(\mathcal{G})$ is defined, and undefined if not. We will start by first showing that $\mathcal{A} = \mathcal{A}^*$ where

$$\mathcal{A}^* = \left\{ \mathcal{G} \mid E_c[\mathcal{G}] \geq 0 \text{ for all } c \in \mathcal{C} \right\}.$$

Suppose that $\mathcal{G} \in \mathcal{A}$. So

$$\mathcal{G} = \sum_{i \leq m} \lambda_i (X_i - Y_i),$$

where $\lambda_i \geq 0$ and $X_i \succeq Y_i$ for all $i \leq m$. Again let

$$X_i = \left\langle x_1^i, \ldots, x_n^i \right\rangle.$$

Likewise for Y_i. Choose $c \in \mathcal{C}$. Then

$$\begin{aligned} E_c[\mathcal{G}] &= \sum_{w_i \in \Omega} c(w_i) \left[\sum_{j \leq m} \lambda_j (x_i^j - y_i^j) \right] \\ &= \sum_{j \leq m} \lambda_j \left[\sum_{w_i \in \Omega} c(w_i)(x_i^j - y_i^j) \right] \\ &= \sum_{j \leq m} \lambda_j (c(X_j) - c(Y_j)). \end{aligned}$$

Since $X_j \succeq Y_j$ for all $j \leq m$, $c(X_j) \geq c(Y_j)$. So

$$E_c[\mathcal{G}] \geq 0.$$

Therefore $E_c[\mathcal{G}] \geq 0$ for all $c \in \mathcal{C}$. So $\mathcal{G} \in \mathcal{A}^*$.

Now suppose that $\mathcal{G} \in \mathcal{A}^*$. Suppose for reductio that $\mathcal{G} \notin \mathcal{A}$.

Note that $\mathcal{A}$ (= $\mathbb{A}$) is the closed, convex polyhedral cone generated by the set $\{X - Y \mid X \succeq Y\}$. So the hyperplane separation theorem of Kuhn and Tucker (1956, p. 50) guarantees that there is a linear functional E that strictly separates this point $\mathcal{G} \notin \mathcal{A}$ from $\mathbb{A}$ in the sense that

$$E[\mathcal{V}] \geq 0 \text{ for all } \mathcal{V} \in \mathbb{A},$$

but

$$E[\mathcal{G}] < 0.$$

The proof of Theorem 1 shows how to use E to construct a probability function b that almost (indeed, strongly) agrees with $\succeq$. And b is such that

$$E_b[\mathcal{V}] = \sum_{w_i \in \Omega} b(w_i)v_i = \sum_{w_i \in \Omega} \frac{E(w_i)}{E(\Omega)} v_i = \frac{E[\mathcal{V}]}{E(\Omega)}$$

for any gamble $\mathcal{V}$. So $E_c[\mathcal{V}] \geq 0$ iff $E[\mathcal{V}] \geq 0$. In particular, then, since $E[\mathcal{G}] < 0$, $E_b[\mathcal{G}] < 0$ as well.

But since b almost agrees with $\succeq$, $b \in \mathcal{B} = \mathcal{C}$. Since $E_b[\mathcal{G}] < 0$, $\mathcal{G} \notin \mathcal{A}^*$. Contradiction.

This establishes that $\mathcal{A} = \mathcal{A}^*$. Now we will show that $\underline{\mathcal{E}_{\mathcal{C}}}[\mathcal{G}] = \mathcal{B}(\mathcal{G})$ if $\mathcal{B}(\mathcal{G})$ is defined, and undefined if not.

$$\begin{aligned}
\underline{\mathcal{E}_{\mathcal{C}}}[\mathcal{G}] &= \inf\left\{E_c[\mathcal{G}] \mid c \in \mathcal{C}\right\} \\
&= \sup\left\{\mathfrak{a} \mid E_c[\mathcal{G}] \geq \mathfrak{a} \text{ for all } c \in \mathcal{C}\right\} \\
&= \sup\left\{\mathfrak{a} \mid E_c[\mathcal{G} - \mathfrak{a}] \geq 0 \text{ for all } c \in \mathcal{C}\right\} \\
&= \sup\left\{\mathfrak{a} \mid \mathcal{G} - \mathfrak{a} \in \mathcal{A}^*\right\} \\
&= \sup\left\{\mathfrak{a} \mid \mathcal{G} - \mathfrak{a} \in \mathcal{A}\right\} \\
&= \mathcal{B}(\mathcal{G}).
\end{aligned}$$

■

Theorem 5 *A relation $\succeq$ strongly agrees with a real-valued function c if and only if $\succeq$ satisfies weak transitivity:*

WEAK TRANSITIVITY. *If $X \succeq Y_1 \succeq \ldots \succeq Y_n \succeq Z$, then $X \not\prec Z$.*

Proof. The left-to-right direction is trivial. So suppose that $\succeq$ satisfies weak transitivity. For any $X \in \mathcal{F}$, let

$$\Phi_X = \{X\} \cup \left\{Z \mid X \succeq Y_1 \succeq \ldots \succeq Y_n \succeq Z \text{ for some } Y_1 \succeq \ldots \succeq Y_n \in \mathcal{F}\right\}.$$

Let $c : \mathcal{F} \to \mathbb{R}$ be defined by

$$c(X) = |\Phi_X|.$$

We must show:

(i) $A \succeq B \Rightarrow c(A) \geq c(B)$,

(ii) $A \succ B \Rightarrow c(A) > c(B)$.

Assume that $A \succeq B$. Choose $Z \in \Phi_B$. Either $Z = B$ or

$$B \succeq Y_1 \succeq \ldots \succeq Y_n \succeq Z$$

for some $Y_1 \succeq \ldots \succeq Y_n \in \mathcal{F}$. So either

$$A \succeq Z$$

or

$$A \succeq B \succeq Y_1 \succeq \ldots \succeq Y_n \succeq Z.$$

Either way, $Z \in \Phi_A$. Hence $\Phi_B \subseteq \Phi_A$. As a result

$$c(A) = |\Phi_A| \geq |\Phi_B| = c(B).$$

Now suppose that $A \succ B$. As before, $\Phi_B \subseteq \Phi_A$. But now note that while $A \in \Phi_A$, $A \notin \Phi_B$.

To see this, suppose for reductio that $A \in \Phi_B$. Then either $A = B$ or

$$B \succeq Y_1 \succeq \ldots \succeq Y_n \succeq A$$

for some $Y_1 \succeq \ldots \succeq Y_n \in \mathcal{F}$. If $A = B$, then $A \succ A$, *i.e.*, $A \succeq A$ but $A \not\preceq A$. Contradiction. If $B \succeq Y_1 \succeq \ldots \succeq Y_n \succeq A$, then by weak transitivity, $B \not\prec A$. Contradiction.

So $A \in \Phi_A$ but $A \notin \Phi_B$. Hence $\Phi_B \subset \Phi_A$. As a result

$$c(A) = |\Phi_A| > |\Phi_B| = c(B).$$

■

REFERENCES

Abellan, J. & Moral, S. (2000). A non-specificity measure for convex sets of probability distributions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *8*, 357–367.

Abellan, J. & Moral, S. (2003). Maximum entropy for credal sets. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *11*, 587–597.

Adams, E. W. (1965). Elements of a theory of inexact measurement. *Philosophy of Science*, *32*(205–228).

Alon, S. & Lehrer, E. (2014). Subjective multi-prior probability: A representation of a partial likelihood relation. *Journal of Economic Theory*, *151*, 476–92.

Augustin, T., Coolen, F., de Cooman, G., & Troffaes, M. (2014). *Introduction to imprecise probabilities*. Wiley Series in Probability and Statistics. Wiley.

Clark, A. (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.), *Open mind* (Vol. 7, *T*). Frankfurt am Main: MIND Group.

de Finetti, B. (1931). Sul significato soggesttivo della probabilita. *Fund. Mathm*, *17*, 298–329.

de Finetti, B. (1951). La "logica del plausibile" secondo la concezione di polya. In *Atti della xlii riunione* (pp. 227–236). Rome: Societa Italiana per il Progresso delle Scienze.

de Finetti, B. (1964). Foresight: Its logical laws, it subjective sources (1937). In H. E. Kyburg Jr. & H. E. Smokler (Eds.), *Studies in subjective probability* (Vol. 7, pp. 93–158). Wiley.

de Finetti, B. (1974). *Theory of probability*. Wiley.

Deza, M. & Deza, E. (2009). *Encyclopedia of distances*. Heidelberg. Springer.

Domotor, Z. (1969). *Probabilistic relational structures and their applications* (tech. rep. No. 144). Institute for Mathematical Studies in the Social Sciences, Stanford University.

Earman, J. (1992). *Bayes or bust? a critical examination of bayesian confirmation theory*. Cambridge: MIT Press.

Elliott, E. (2018). Comparativism and the measurement of partial belief. *Ms.*

Eriksson, L. & Hájek, A. (2007). What Are Degrees of Belief? *Studia Logica: An International Journal for Symbolic Logic*, *86*(2), 183–213.

Fishburn, P. C. (1969). Weak qualitative probability on finite sets. *Annals of Mathematical Statistics*, *40*, 2118–2126.

Fishburn, P. C. (1986). The axioms of subjective probability. *Statistical Science*, *1*(3), 335–345.

Fitelson, B. & McCarthy, D. (2015). Toward an epistemic foundation for comparative confidence. *Ms.*

Gillies, D. (2000). Varieties of propensity. *British Journal for the Philosophy of Science*, *51*, 807–835.

Good, I. J. (1950). *Probability and the weighing of evidence*. London: Griffin.

Hajek, A. (2008). Arguments for – or against – probabilism. *British Journal for the Philosophy of Science*, *59*(4), 793–819.

Hájek, A. (2003). What Conditional Probability Could Not Be. *Synthese*, *137*(3), 273–323.

Hájek, A. (2009). Argument for-or against-probabilism? In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (Vol. 342). Springer.

Harrison-Trainor, M., Holliday, W. H., & Icard, T. F. (2016). A note on cancellation axioms for comparative probability. *Theory and Decision*, *80*(1), 159–166.

Hitchcock, C. (2012). *Cause and chance*. Ms.

Icard, T. (2016). Pragmatic considerations on comparative probability. *Philosophy of Science*, *83*, 348–370.

Jeffrey, R. (1965/1983). *The logic of decision* (2nd). University of Chicago Press.

Jeffrey, R. (1965). *The logic of decision*. University of Chicago Press.

Jeffrey, R. (1987). Indefinite probability judgment: A reply to levi. *Philosophy of Science*, *54*(4), 586–591.

Jeffrey, R. (2002). *Subjective probability: The real thing*. Cambridge: Cambridge University Press.

Jeffreys, H. (1961). *Theory of probability* (3rd). Oxford: Clarendon Press.

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, *65*(4), 575–603.

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge University Press.

Joyce, J. M. (2005). How probabilities reflect evidence. *Philosophical Perspectives*, *19*, 153–178.

Joyce, J. M. (2009). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (Vol. 342). Dordrecht: Springer.

Joyce, J. M. (2010). A defense of imprecise credences in inference and decision making. *Philosophical Perspectives*, *24*(1), 281–323.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–91.

Kaplan, M. (2010). In defense of modest probabilism. *Synthese*, *176*(1), 41–55.

Koopman, B. O. (1940a). The axioms of algebra of intuitive probability. *Annals of Mathematics*, *41*(2), 269–292.

Koopman, B. O. (1940b). The bases of probability. *Bulletin of the American Mathematical Society*, *46*, 763–774.

Kraft, C. H., Pratt, J. W., & Seidenberg, A. (1959). Intuitive probability on finite sets. *The Annals of Mathematical Statistics*, *30*(2), 408–419.

Krantz, D., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement vol. i: Additive and polynomial representations*. New York: Academic Press.

Kuhn, H. & Tucker, A. (Eds.). (1956). *Linear inequalities and related systems*. Princeton University Press.

Kyburg, H. & Pittarelli, M. (1996). Set-based bayesianism. *IEEE Transactions on Systems, Man, and Cybernetics*, *26*, 324–339.

Lewis, D. (1974). Radical interpretation. *Synthese*, *23*, 331–344.

Lichtenstein, S. & Slovic, P. (1971). Reversals of preferences between bids and choices in gambling decisions. *Experimental Psychology*, *89*, 46–55.

Lichtenstein, S. & Slovic, P. (1973). Response-induced reversals of preference in gambling decisions: An extended replication in las vegas. *Journal of Experimental Psychology*, *101*(1), 16–20.

Maher, P. (1993). *Betting on theories*. Cambridge University Press.

Meacham, C. & Weisberg, J. (2011). Representation theorems and the foundations of decision theory. *Australasian Journal of Philosophy*, *89*(4), 641–663.

Paris, J. (2011). Pure inductive logic. In L. Horsten & R. Pettigrew (Eds.), *The continuum companion to philosophical logic* (pp. 428–449). London: Continuum International Publishing Group.

Pederson, A. P. (2014). Comparative expectations. *Studia Logica*, *102*, 811–848.

Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford: Oxford University Press.

Predd, J., Seiringer, R., H. Lieb, E., Osherson, D., Poor, H. V., & R. Kulkarni, S. (2009). Probabilistic coherence and proper scoring rules. *IEEE Transaction on Information Theory*, *55*(10), 4786–4792.

Quaeghebeur, E. (2014). Introduction to imprecise probabilities. In T. Augustin, F. Coolen, G. de Cooman, & M. Troffaes (Eds.), (Chap. 1, pp. 1–27). Wiley.

Ramsey, F. P. (1931). Truth and probability. In *The foundations of mathematics and other logical essays*. New York: Humanities Press.

Rényi, A. (1955). On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarum Hungaricae*, *6*, 286–335.

Rios Insua, D. (1992). On the foundations of decision making under partial information. *Theory and Decision*, *33*(1), 83–100.

Savage, L. (1954). *The foundations of statistics*. Wiley.

Schervish, M., Seidenfeld, T., & Kadane, J. (2009). Proper scoring rules, dominated forecasts, and coherence. *Decision Analysis*, *6*(4), 202–221.

Scott, D. (1964). Measurement structures and linear inequalities. *Journal of Mathematical Psychology*, *1*(2), 233–247.

Staffel, J. (2018). *Unsettled thoughts: A theory of degrees of rationality*. Oxford: Oxford University Press.

Suppes, P. (1994). Qualitative theory of subjective probability. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 17–37). Chichester, UK: John Wiley and Sons.

Suppes, P. & Zanotti, M. (1976). Necessary and sufficient conditions for existence of a unique measure strictly agreeing with a qualitative probability ordering. *Journal of Philosophical Logic*, *5*(3), 431–438.

Suppes, P. & Zanotti, M. (1982). Necessary and sufficient qualitative axioms for conditional probability. *Z. Wahrsch. Verw. Gebiete*, *60*, 163–169.

Titelbaum, M. (2015). *Fundamentals of bayesian epistemology*. Oxford: Oxford University Press.

Troffaes, M. C. M. & de Cooman, G. (Eds.). (2014). *Lower previsions*. Wiley.

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. New York: Chapman and Hall.

Walley, P. (2000). Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, *24*(2–3), 125–148.

Zynda, L. (2000). Representation theorems and realism about degrees of belief. *Philosophy of Science*, *67*(1), 45–69.

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 (sections 2–3). 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 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 are 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" of this handbook.[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 just 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, we 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 recent works on this issue, see Arló-Costa and Pedersen (2012), Lin and Kelly (2012), and Leitgeb (2014). Also see the chapter "Full and Partial Belief" of this handbook.

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 (no matter whether 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 and Rational Monotonicity are logically independent. 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', 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). (To anticipate, it will be very much like looking for the right analysis of knowledge in epistemology.) 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, which is a sociological fact that I do not know how to explain.

I have to confess that many belief revision theorists (including me) have incurred the debt without working hard enough to pay it off. Anyway, 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

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

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, ..., and then receives information ϕ_n. A belief revision strategy is meant to tell one how to change beliefs given any relevant history of inquiry. Accordingly:

> 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:

> 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:

> 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:

h = the hypothesis "all ravens are black";
b_i = "the i-th observed raven is black";
n_i = "the i-th observed raven is non-black."

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 most 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 we feel 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 of potential counterexamples.

But we should not overemphasize the importance of this instrumental role in philosophy. A representation result is sometimes an 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 the following. A partial, sound axiomatization already starts to facilitate the search for potential counterexamples. It would be great if we also have 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 a hard choice as to how much time to invest in trying to prove a representation conjecture only for the sake of this instrumental purpose.

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

(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 that:

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

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 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.

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 $\max(|\phi|, \geq)$. That is:

> 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:

> 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 $I_vF_bF_s$ 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 total, shown in Figure 1. The arrows

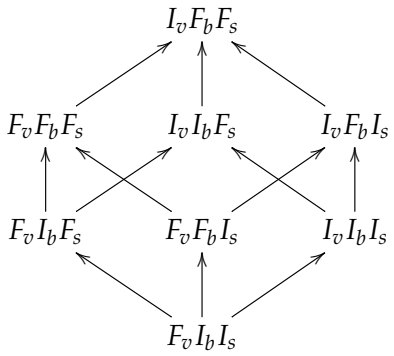


Figure 1: Hasse diagram of the Three Composers problem

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 $\geq$:

- let $I_vF_bF_s$ 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 $I_vF_bF_s$ 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: $I_vF_bF_s$. Then the agent receives the first information E, that v and b are compatriots. So the worlds incompatible with that information are to be ruled out, as shown on the left hand side of Figure 2. At this stage, the agent believes that the actual world is one of the two most plausible worlds: $F_vF_bF_s$ and $I_vI_bF_s$. Then the agent receives

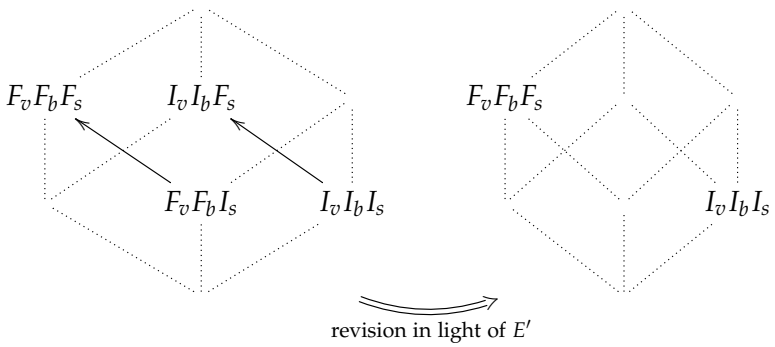


Figure 2: Revising in light of E, and then E'

the second information E', that v and s are compatriots. So the worlds incompatible with that information are to be ruled out, as shown on the right hand 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:

$$\geq \longrightarrow \geq_{*(\phi_1)} \longrightarrow \geq_{*(\phi_1,\phi_2)} \longrightarrow \geq_{*(\phi_1,\phi_2,\phi_3)} \longrightarrow \cdots$$

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

$$\begin{array}{ccccccccc} \geq & \longrightarrow & \geq_{*(\phi_1)} & \longrightarrow & \geq_{*(\phi_1,\phi_2)} & \longrightarrow & \geq_{*(\phi_1,\phi_2,\phi_3)} & \longrightarrow & \cdots \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \\ S() & & S(\phi_1) & & S(\phi_1,\phi_2) & & S(\phi_1,\phi_2,\phi_3) & & \cdots \end{array}$$

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 1), without changing the ordering of the worlds inside X (condition 2), nor changing the ordering of the worlds outside X (condition 3). 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 $w >_{*(\ldots,\phi)} u$. Some others think that even the requirement of canonical revision is too strong: Stalnaker (2009) proposes a counterexample, which we will discuss in section 5.3.

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? This is essentially a decision problem, and we will need some decision theory to help us out.

Recall the Raven Problem of section 3.3: "are all ravens black?" To choose among belief revision strategies for solving that problem, let us draw a decision 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

	h	$\neg h, n_1$	$\neg h, b_1, n_2$	...	$\neg h, b_1, b_2, \ldots, n_{101}$	...
$S^{\text{ock}}_{\text{non-skep}}$						
$S^{\text{non-ock}}_{\text{non-skep}}$						
S_{skep}						

Table 1: A decision table for the Raven Problem

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—in the present case, learning whether all ravens are black. Then, with all those outcomes specified, we will use a decision rule to sort out the options that are "OK" from those that are not "OK."

The three strategies listed in the decision table are defined as follows. The *skeptic* 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, for any set X of sentences. 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.

A non-skeptic *Ockham* strategy is a strategy that starts from asking one to believe just the logical consequences of the accumulated information, but after observing sufficiently many black ravens in a row without any counterexample, it asks one to believe h, the simpler hypothesis between h and $\neg h$, following a form of *Ockham's Razor*. For example, consider the following strategy:

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

This strategy says that 100 black ravens suffice for the inductive leap. We can replace 100 with another positive integer, which would generate another non-skeptic Ockham strategy.

A non-skeptic *non-Ockham* 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 at a later point. For example, consider the following strategy:

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

What makes it non-Ockham is the first clause. Replacement of 50 and 100 with other numbers m and n (where $m < n$) would generate other non-skeptic non-Ockham 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 decision table to think about.

Next: fill the cells with specifications of 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 the unique answer in $\{h, \neg h\}$ that is true given C, and to hold on to that answer henceforth. To be 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 "Yes" in a cell means: "yes, the strategy will

	h	$\neg h, n_1$	$\neg h, b_1, n_2$	…	$\neg h, b_1, b_2, \ldots, n_{101}$	…
$S^{\text{ock}}_{\text{non-skep}}$						
$S^{\text{non-ock}}_{\text{non-skep}}$						
S_{skep}	No	Yes	Yes	…	Yes	…

Table 2: Decision table for the Raven Problem continued

learn whether h is true given the column." Similarly, "No" 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 "$\neg h, b_1, \ldots, n_{i+1}$": the right answer is obtained, and held on to, beginning from the $(i+1)$-th observation, because n_{i+1} entails $\neg h$. It is not hard to verify that the cells left blank in the above should all be filled with "Yes."

We want to think about, not just whether a strategy will learn, but also how well it learns. Consider this situation: one believes X and then comes to believe something else that contradicts X and then comes (back!) to believe X. In that case, say that one has an *opinion cycle*. The more opinion cycles a strategy incurs, the worse it performs. Now, for each cell, let us specify (i) whether the strategy will learn and (ii) how many opinion cycles it will incur: see Table 3. For example, given the column "$\neg h, b_1, \ldots, b_i, n_{i+1}$" with $i \geq 100$, the non-skeptic non-Ockham strategy will generate 1 opinion cycle: it asks one to believe $\neg h$ on the 50th observation, believe h on the 100th, and switch back to the belief in $\neg h$ on the $(i+1)$-th, which forms an opinion cycle.

With our decision table complete, it is time to apply a decision rule to sort out the strategies that are "OK." Let us try applying the so-called

	h	$\neg h, n_1$	$\neg h, b_1, n_2$	...	$\neg h, b_1, b_2, \ldots, n_{101}$	...
$S^{\text{ock}}_{\text{non-skep}}$	(Yes, 0)	(Yes, 0)	(Yes, 0)	...	(Yes, 0)	...
$S^{\text{non-ock}}_{\text{non-skep}}$	(Yes, 0)	(Yes, 0)	(Yes, 0)	...	(Yes, 1)	...
S_{skep}	(No, 0)	(Yes, 0)	(Yes, 0)	...	(Yes, 0)	...

Table 3: Decision table for the Raven Problem complete

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 "OK" iff[17] its worst outcome is one of the best among the worst outcomes of the options on the table. Namely, Maximin asks one to *maxi*mize the *min*imal payoff. Presumably, learning is better than failing to learn, and less opinion cycles are better than more opinion cycles. So we identify the worst outcomes as in Table 4. It follows that, according to Maximin, only

	worst outcome
$S^{\text{ock}}_{\text{non-skep}}$	(Yes, 0)
$S^{\text{non-ock}}_{\text{non-skep}}$	(Yes, 1)
S_{skep}	(No, 0)

Table 4: Worst possible outcomes for the Raven Problem

the non-skeptic Ockham strategy is "OK."

The above considers only three specific revision strategies—just for the sake of illustration. It is straightforward to cover all possible revision strategies for solving the Raven Problem: just add a row to the decision table for each of those strategies, and specify the outcomes in the new cells. Once that is done, we can apply the Maximin rule to the fully completed decision table, and single out the revision strategies that Maximin judges to be "OK." These "OK" revision strategies form a set, i.e. a formal theory of belief revision. If we only consider the two performance criteria just presented—(i) whether a strategy will learn and (ii) how many opinion cycles it will incur—then Maximin favors only the non-skeptical Ockham strategies.

To sum up: a belief revision theory can be constructed in terms of the learning performances of belief revision strategies, together with decision-theoretic tools such as decision tables, decision rules, and preference relations between outcomes. This idea admits of many possible implementations:

17 In case you want to be more careful: to make the Maximin rule compatible with the Weak Dominance principle, 'iff' should be weakened to 'only if'.

- *We may consider drawing the decision table in a different way.*

  Have we considered all the relevant, possible columns? Are the columns fine-grained enough? If a column C is so unspecific that it does not determine the total number of opinion cycles that a strategy S will incur, should we fine-grain column C into more specific possibilities?

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

  We have only talked about whether one will learn and how many opinion cycles one will produce. 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 one conjecture a false answer? how fast will the true answer be learned?

- *We may consider other decision rules.*

  How about other decision rules like Minimax Regret, Maximax, or even Maximization of Expected Utility if this does not beg the inductive skeptic's question?

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.

The learning-theoretic analysis presented above is just a "baby version" for the sake of illustration. It is adapted from Genin and Kelly (2015) and Kevin T Kelly, Genin, and Lin (2016), which build on Schulte (1999) and Kevin T Kelly (2007). Also see Kevin T Kelly (1999) for an application of learning-theoretic analysis to iterated belief revision, where we care about the possibility of receiving mutually contradictory pieces of information.

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 entry on ranking theory in 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:
>
> (A) Verdi is Italian;
>
> (B) Bizet is French;

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).

(*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,

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', 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—although it is not clear to me how this can be done.

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.

There is, of course, an issue whether this line of response can, or cannot, be developed successfully to save Preservation.[20] I have to confess that I am unable to see how the defenders of Preservation can succeed. So, 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.

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

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 $\neg M$.
>
> *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 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 also given independently, though we may assume 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?

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:

> α = the conjunction of what Carla says and what Dora says;
> β = 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*

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.

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 following forms:

> (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:

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

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 $\div$, $\perp$, 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 this:

> 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 the AGM axioms and, hence, 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:

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

> 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 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 $B\bot\neg\phi$. 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 the contracted belief set *should* be the intersection of the epistemically most important candidates? That is, why the concepts of belief contraction and epistemic importance are 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 (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). Decision rule D judges every strategy not in T to be inferior to some strategy in T.
>
> PREMISE (II). If decision rule D judges a strategy S to be inferior to some other strategy, then S 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.

Now we can turn the learning-theoretic analysis in section 4.6 into an explicit argument for a belief revision theory. Let D be the Maxmin rule, T be the set of all belief revision strategies for the Raven Problem except the following two kinds: the skeptic strategies S_{skep} and the non-skeptic non-Ockham strategies $S^{\text{non-ock}}_{\text{non-skep}}$. Recall that strategies of these two kinds are judged by the Maximin rule to be inferior to some strategy in T, e.g. some non-skeptic Ockham strategy $S^{\text{ock}}_{\text{non-skep}}$.

So we have formulated an explicit argument. Since that argument is valid, let us turn to worries about its two premises.

The more urgent worry is about Premise (II): which decision rule is the right one to apply? In section 4.6, we apply the Maximin decision rule. But is Maximin the right decision rule for singling out rational or epistemically justified strategies of belief revision? Many decision theorists think 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?", and "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. So, to make a decision without begging the skeptic 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 Wald's (1950) Maximin foundation of statistical inference.

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. Kevin T 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.

Let us now turn to Premise (I). Even if Maximin (or some other qualitative decision rule) is the right one to apply when we respond to the inductive skeptic's challenge, does it really rule out the skeptic strategy and the non-skeptic non-Ockham strategy? What if Maximin is *misapplied* in section 4.6? Here are some possibilities of misapplication to think about.

- First, is there a missing column? Think about the unfortunate possibility in which (a) not all ravens are black, (b) we will observe one black raven after another *ad infinitum*, but (c) we will never observe a nonblack raven. We did *not* put this column into the decision table in section 4.6. Are we justified in missing that column? After we add this additional column, will Maximin still favor the non-skeptic Ockham strategy?

- Second, have the outcomes in the cells been specified with sufficient detail? Why care just about whether a strategy will learn and how many opinion cycles will be produced? If we add more epistemic considerations into the cells, will Maximin still favor the non-skeptic Ockham strategy?

It remains to explore how these two questions may be answered. See Kevin T Kelly (2007) for further discussion.

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 claims that a belief revision theory is intended to make (sections 5 and 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), and Gillies (2004).

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):

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

To be more precise: given any simple belief revision strategy S, we can use the bridge principle to define a nonmonotonic consequence relation $\mid\!\sim_S$ as follows: $\phi \mid\!\sim_S \psi$ iff $\psi \in S(\phi)$. Conversely, given any nonmonotonic consequence relation $\mid\!\sim$, we can use the bridge principle to define a simple belief revision strategy $S_{\mid\!\sim}$ as follows: $S_{\mid\!\sim}(\phi) =_{\text{def}} \{\psi : \phi \mid\!\sim \psi\}$ and $S_{\mid\!\sim}() =_{\text{def}} S_{\mid\!\sim}(\top)$, 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 $\mathcal{Q}$'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, and I as a belief revision theorist, 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 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.

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. & Kelly, K. T. [Kevin T]. (2015). Theory choice, theory change, and inductive truth-conduciveness. *Studia Logica*, 1–41.

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.

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.

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

Kelly, K. T. [Kevin 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. [Kevin 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.

Leitgeb, H. (2014). The stability theory of beliefthe stability theory of beliefhannes leitgeb. *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. [Kevin 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.

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 *Causation in decision, belief change, and statistics* (pp. 105–134). Springer.

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.

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

8

RANKING THEORY

Franz Huber

In epistemology ranking theory is a theory of belief and its revision. It studies how an ideal doxastic agent should organize her beliefs and conditional beliefs at a given moment in time, and how she should revise her beliefs and conditional beliefs across time when she receives new information. In this entry I will first present some background, most notably the AGM theory of belief revision (Alchourrón, Gärdenfors, & Makinson, 1985). In order to motivate the introduction of ranking theory I will then focus on the problem of *iterated* belief revisions. After presenting the elements of ranking theory (Spohn, 1988, 2012) I will show how this theory solves the problem of iterated belief revisions. I will conclude by sketching two areas of future research and by mentioning applications of ranking theory outside epistemology. Along the way we will see how ranking theory, a theory of belief, compares to subjective probability theory or Bayesianism, which is a theory of partial beliefs or degrees of belief.

1 INTRODUCTION

Sophia believes many things, among others that it will rain on Tuesday, that it will be sunny on Wednesday, and that weather forecasts are always right. Belief revision theory tells Sophia how to revise her beliefs when she learns that the weather forecast for Tuesday and Wednesday predicts rain. As we will see, this depends on the details of her beliefs, but under one way of filling in the details she should keep her belief that it will rain on Tuesday and give up her belief that it will be sunny on Wednesday. To state in full detail how Sophia should revise her beliefs when she learns new information we need a representation of her old beliefs and of the new information she receives.

In this entry I will focus on ideal doxastic agents who do not suffer from the computational and other physical limitations of ordinary doxastic agents such as people and computer programs. These ideal doxastic agents get to voluntarily decide what to believe (and to what degree of numerical precision); they never forget any of their (degrees of) beliefs; and they always believe all logical and conceptual truths (to a maximal degree). We may perhaps define a (doxastic or cognitive) agent to be *ideal* just in case any (cognitive) action that is physically possible is an action that is possible for her. Such ideal agents ought to do exactly that which they ought to do if they could, where the 'can' that is hidden in the 'could'

expresses possibility for the agent, not metaphysical possibility. Hence the principle that *Ought Implies Can* does not put any constraints on what an ideal agent should do, and on what an ideal doxastic agent should believe.

Belief revision theory models belief as a qualitative attitude towards sentences or propositions: the ideal doxastic agent believes a proposition, or she disbelieves the proposition by believing its negation, or she suspends judgment with respect to the proposition and its negation. This is different from the theory of subjective probabilities, also known as *Bayesianism* (Easwaran 2011a, 2011b; Titelbaum, this volume; Weisberg 2011; Wenmackers, this volume), where belief is modeled as a quantitative attitude towards a proposition: the ideal doxastic agent believes a proposition to a specific degree, her degree of belief, or credence, for the proposition. However, we will see that, in order to adequately model conditional beliefs and iterated belief revisions, ranking theory also models the ideal agent's doxastic state with numbers, and thus more than just the set of propositions she believes. Genin (this volume) discusses the relation between belief and degree of belief.

2 BELIEF REVISION

Spohn (1988, 1990) develops ranking theory in order to fix a problem that besets the AGM theory of belief revision. In order to provide some background for ranking theory I will first present the AGM theory. Ranking theory will then arise naturally out of the AGM theory. The latter theory derives its name from the seminal paper by Alchourrón et al. (1985). Comprehensive overviews can be found in Gärdenfors (1988), Gärdenfors and Rott (1995), Rott (2001), and Lin (this volume).

One version of the AGM theory of belief revision represents the ideal doxastic agent's old beliefs, her doxastic state at a given moment in time, by a set of sentences from some formal language, her *belief set*, together with an *entrenchment ordering* over these sentences. The entrenchment ordering represents how firmly the ideal doxastic agent holds the beliefs in her belief set. It represents the details of the ideal agent's doxastic state. The new information is represented by a single sentence. The AGM theory distinguishes between the easy case, called *expansion*, and the general case, called *revision*. In expansion the new information does not contradict the ideal doxastic agent's old belief set and is simply added. In revision the new information may contradict the ideal doxastic agent's old belief set. The general case of revision is difficult, because the ideal doxastic agent has to turn her old belief set, which is assumed to be consistent, into a new belief set that contains the new information and is consistent. Usually the general case is dealt with in two steps. In a first step, called *contraction*, the old belief set is cleared of everything that contradicts the new information.

In a second step one simply expands by adding the new information. This means that the difficult doxastic task is handled by contraction, which turns the general case of revision into the easy case of expansion.

A formal language $\mathcal{L}$ is defined inductively, or recursively, as follows. $\mathcal{L}$ contains the contradictory sentence $\ulcorner\bot\urcorner$ and all elements of a countable set of propositional variables $PV = \{\ulcorner P\urcorner, \ulcorner Q\urcorner, \ulcorner R\urcorner, \ldots\}$. Furthermore, whenever A and B are sentences of $\mathcal{L}$, then so are the negations of A and of B, $\ulcorner\neg A\urcorner$ and $\ulcorner\neg B\urcorner$, respectively, as well as the conjunction of A and B, $\ulcorner(A \wedge B)\urcorner$. Finally, nothing else is a sentence of $\mathcal{L}$. The new information is represented by a single sentence A from $\mathcal{L}$. The ideal agent's doxastic state is represented by a set of sentences, her belief set $\mathcal{B} \subseteq \mathcal{L}$, plus an entrenchment ordering $\preceq$ for $\mathcal{B}$. The entrenchment ordering, which represents the details of the ideal doxastic agent's beliefs, orders the agent's beliefs according to how reluctant she is to give them up: the more entrenched a belief, the more reluctant she is to give it up.

The entrenchment ordering does most of the work in a revision of the agent's beliefs. Suppose the agent receives new information that contradicts her belief set. Since the new belief set that results from the revision has to be consistent, some of the old beliefs have to go. The entrenchment ordering determines which beliefs have to go first: the least entrenched beliefs are the beliefs that have to go first. If giving up those is not enough to restore consistency, the beliefs that are next in the entrenchment ordering have to go next. And so on. The beliefs that would be given up last are the most entrenched ones. According to Maximality, they are the tautological sentences, which are always believed and never given up, because doing so cannot restore consistency. On the other end of the spectrum are the least entrenched sentences. According to Minimality, they are the sentences the agent does not even believe to begin with. These sentences do not belong to the agent's belief set and so are gone before the revision process has even begun.

If one sentence logically implies another sentence, then, according to Dominance, the latter cannot be more entrenched than the former, as giving up the belief in the latter sentence is to also give up the belief in the former sentence. Dominance implies that the entrenchment ordering is *reflexive*: every sentence is at least as entrenched as itself. According to Conjunctivity, two sentences cannot both be more entrenched than their conjunction: one cannot give up one's belief in a conjunction without giving up one's belief in at least one of the conjuncts. In combination with Dominance, Conjunctivity implies that the entrenchment ordering is *connected*: any two sentences can be compared to each other in terms of their comparative entrenchment. That is, either the first sentence is at least as entrenched as the second sentence, or the second sentence is at least as entrenched as the first sentence, or both. Finally, to ensure that the

entrenchment ordering is a well-behaved ordering relation, it is assumed to be *transitive* by Transitivity.

More precisely, where $\vdash$ is the logical consequence relationship on $\mathcal{L}$ and $Cn(\mathcal{B}) = \{A \in \mathcal{L} : \mathcal{B} \vdash A\}$ is the set of logical consequences of $\mathcal{B}$ (and $\emptyset$ is the empty set $\{\}$), the entrenchment ordering has to satisfy the following postulates. For all sentences A, B, and C from $\mathcal{L}$:

$\preceq$1. If $A \preceq B$ and $B \preceq C$, then $A \preceq C$. Transitivity

$\preceq$2. If $\{A\} \vdash B$, then $A \preceq B$. Dominance

$\preceq$3. $A \preceq A \wedge B$ or $B \preceq A \wedge B$. Conjunctivity

$\preceq$4. Suppose $\mathcal{B} \nvdash \bot$. Then $A \notin \mathcal{B}$ if, and only if, for all $B \in \mathcal{L}$: $A \preceq B$. Minimality

$\preceq$5. If $A \preceq B$ for all $A \in \mathcal{L}$, then $\emptyset \vdash B$. Maximality

The work that is done by the entrenchment ordering in a revision of the agent's beliefs can also be described differently in terms of expansion, revision, and contraction, which turn belief sets and new information into belief sets (see Caie, this volume). Formally they are functions from $\wp(\mathcal{L}) \times \mathcal{L}$ into $\wp(\mathcal{L})$.

Expansion $\dot{+}$ turns each old belief set $\mathcal{B} \subseteq \mathcal{L}$ and each sentence A from $\mathcal{L}$ into a new belief set $\mathcal{B}\dot{+}A = Cn(\mathcal{B} \cup \{A\})$. This is the easy case described earlier about which there is little more to be said.

The difficult and more interesting case is revision $*$, which turns each old belief set $\mathcal{B} \subseteq \mathcal{L}$ and each sentence A from $\mathcal{L}$ into a new belief set $\mathcal{B} * A$. The operator $*$ is required to satisfy a number of postulates.

Closure requires revised belief sets to be closed under the logical consequence relation: after the revision is completed, the agent ought to believe all (and only) the logical consequences of the revised belief set. Congruence is similar in spirit to Closure and requires that it is the content of the new information received, and not its particular formulation, that determines what is added, and what is removed, from the agent's belief set in a revision. Success requires that revising a belief set by new information succeeds in adding the new information to the agent's belief set—and, given Closure, all sentences it logically implies. Consistency requires the revised belief set to be consistent as long as the new information is consistent. The remaining postulates all formulate different aspects of the idea that, when revising her belief set by new information, the agent should add and remove as few beliefs as possible from her belief set, subject to the constraints that the resulting belief set is consistent and that the new information has been added successfully.

Inclusion requires that revising a belief set does not create any new beliefs that are not also created by simply adding the new information. In

a sense it says that expansion is a special case of revision. Preservation requires that revising a belief set by new information that does not contradict the agent's old belief set does not lead to the loss of any beliefs. Conjunction 1 requires that, when revising her belief set by a conjunction, the agent adds *only* beliefs that she also adds when first revising her belief set by one of the two conjuncts, and then adding the second conjunct. Conjunction 2 requires that, when revising her belief set by a conjunction, the agent adds *all* beliefs that she adds when first revising her belief set by one of the two conjuncts, and then adding the second conjunct—provided the second conjunct is consistent with the result of revising her belief set by the first conjunct. More precisely, a revision function has to satisfy the following postulates. For all sets of sentences $\mathcal{B} \subseteq \mathcal{L}$ and all sentences A and B from $\mathcal{L}$:

$*1$. $\mathcal{B} * A = Cn(\mathcal{B} * A)$. Closure

$*2$. $A \in \mathcal{B} * A$. Success

$*3$. $\mathcal{B} * A \subseteq Cn(\mathcal{B} \cup \{A\})$. Inclusion

$*4$. If $\mathcal{B} \nvdash \neg A$, then $\mathcal{B} \subseteq \mathcal{B} * A$. Preservation

$*5$. If $\{A\} \vdash B$ and $\{B\} \vdash A$, then $\mathcal{B} * A = \mathcal{B} * B$. Congruence

$*6$. If $\emptyset \nvdash \neg A$, then $\bot \notin \mathcal{B} * A$. Consistency

$*7$. $\mathcal{B} * (A \wedge B) \subseteq Cn\left((\mathcal{B} * A) \cup \{B\}\right)$. Conjunction 1

$*8$. If $\neg B \notin \mathcal{B} * A$, then $Cn\left((\mathcal{B} * A) \cup \{B\}\right) \subseteq \mathcal{B} * (A \wedge B)$. Conjunction 2

The two-step view of revision described previously is known as the *Levi identity* (Levi, 1977). It has the ideal doxastic agent first contract $\dot{-}$ her old belief set $\mathcal{B}$ by the negation of the new information, $\neg A$, thus making it consistent with the new information (as well as everything logically implied by the new information). Then it has her expand the result $\mathcal{B} \dot{-} \neg A$ by adding the new information A:

$$\mathcal{B} * A = Cn\left((\mathcal{B} \dot{-} \neg A) \cup \{A\}\right).$$

The Levi identity puts contraction center stage in the revision process. Contraction $\dot{-}$ turns each old belief set $\mathcal{B} \subseteq \mathcal{L}$ and each sentence A from $\mathcal{L}$ into a "reduced" belief set $\mathcal{B} \dot{-} A$ that is cleared of A as well as everything logically implying A. It is required to satisfy the following postulates. For all sets of sentences $\mathcal{B} \subseteq \mathcal{L}$ and all sentences A and B from $\mathcal{L}$:

$\dot{-}1$. $\mathcal{B} \dot{-} A = Cn(\mathcal{B} \dot{-} A)$. Closure

$\dot{-}2$. If $\emptyset \nvdash A$, then $A \notin Cn(\mathcal{B} \dot{-} A)$. Success

$\dot{-}3$. $\mathcal{B} \dot{-} A \subseteq Cn(\mathcal{B})$. Inclusion

$\dot{-}4$. If $\mathcal{B} \nvdash A$, then $\mathcal{B} \dot{-} A = \mathcal{B}$. Vacuity

$\dot{-}5$. If $\{A\} \vdash B$ and $\{B\} \vdash A$, then $\mathcal{B} \dot{-} A = \mathcal{B} \dot{-} B$. Congruence

$\dot{-}6$. $Cn(\mathcal{B}) \subseteq Cn\big((\mathcal{B} \dot{-} A) \cup \{A\}\big)$. Recovery

$\dot{-}7$. $(\mathcal{B} \dot{-} A) \cap (\mathcal{B} \dot{-} B) \subseteq \mathcal{B} \dot{-} (A \wedge B)$. Conjunction 1

$\dot{-}8$. If $A \notin \mathcal{B} \dot{-} (A \wedge B)$, then $\mathcal{B} \dot{-} (A \wedge B) \subseteq \mathcal{B} \dot{-} A$. Conjunction 2

Closure requires contracted belief sets to be closed under the logical consequence relation: after the contraction is completed, the agent ought to believe all (and only) the logical consequences of the contracted belief set. Congruence is similar in spirit to Closure and requires that it is the content of the sentence to be removed, and not its particular formulation, that determines what is removed from the agent's belief set in a contraction. Success requires that contracting a belief set by a sentence that is not tautological succeeds in removing this sentence from a belief set—and, given Closure, all sentences logically implying it. Inclusion requires that contracting a belief set does not add any beliefs to the belief set. The remaining postulates all formulate different aspects of the idea that, when contracting her belief set by a sentence, the agent should remove as few beliefs as possible from her belief set, subject to the constraints that the resulting belief set is consistent and that the sentence to be removed, together with all sentences logically implying it, is removed successfully.

Vacuity requires that contracting a belief set by a sentence leaves the belief set unchanged if the sentence that was to be removed was not even part of the belief set to begin with. Recovery requires that contracting a belief set by a sentence removes as few beliefs as possible so that adding the removed sentence again afterwards allows the agent to recover all her previously removed beliefs. Conjunction 1 requires that, when contracting her belief set by a conjunction, the agent does not remove any beliefs that she does not also remove when contracting by one or the other of the two conjuncts alone. Finally, Conjunction 2 requires the following: if a conjunct is removed in contracting a belief set by a conjunction, then no belief gets removed in contracting the belief set by this conjunct that does not also get removed in contracting this belief set by the entire conjunction. The idea behind the last two postulates is that giving up one of its conjuncts is all the ideal doxastic agent needs to do in order to give up an entire conjunction.

The Levi identity turns each contraction operator $\dot{-}$ satisfying $\dot{-}1 - \dot{-}8$ into a revision operator $*$ satisfying $*1 - *8$. The converse is true of the

Harper identity (Harper, 1976). The latter has the ideal doxastic agent first revise the old belief set $\mathcal{B}$ by the negation of the new information, $\neg A$. Then it has her remove everything from the result $\mathcal{B} * \neg A$ that was not already also a logical consequence of the old belief set $\mathcal{B}$:

$$\mathcal{B} \dot{-} A = (\mathcal{B} * \neg A) \cap Cn(\mathcal{B}).$$

If we have a belief set $\mathcal{B} \subseteq \mathcal{L}$ we can use an entrenchment ordering $\preceq$ for $\mathcal{B}$ to define a revision operator $*$ for $\mathcal{L}$ as follows. For every sentence A from $\mathcal{L}$:

$$\mathcal{B} * A = Cn(\{B \in \mathcal{L} : \neg A \prec B\} \cup \{A\}),$$

where $A \prec B$ holds if, and only if, $A \preceq B$ and $B \not\preceq A$.

The idea behind this equation is the following. When the ideal doxastic agent revises $*$ her old belief set $\mathcal{B}$ by the new information A she first has to clear $\mathcal{B}$ of $\neg A$ as well as everything else that is as entrenched as, or less entrenched than, $\neg A$. For instance, $\mathcal{B}$ also has to be cleared of everything that logically implies $\neg A$. However, it follows from the definition of an entrenchment ordering that all sentences B from the ideal doxastic agent's old belief set $\mathcal{B}$ that are more entrenched than $\neg A$ can be preserved. This gives us the "preserved" belief set $\{B \in \mathcal{L} : \neg A \prec B\}$. Then the ideal doxastic agent adds the new information A to obtain $\{B \in \mathcal{L} : \neg A \prec B\} \cup \{A\}$. Finally she adds all sentences that are logically implied by the preserved belief set together with the new information. As shown by Gärdenfors (1988) and Gärdenfors and Makinson (1988) one can then prove

Theorem 6 *Let $\mathcal{L}$ be a formal language. For each set of sentences $\mathcal{B} \subseteq \mathcal{L}$ and each entrenchment ordering $\preceq$ for $\mathcal{B}$ satisfying $\preceq 1 - \preceq 5$ there is a revision operator $*$ from $\{\mathcal{B}\} \times \mathcal{L}$ into $\wp(\mathcal{L})$ satisfying $*1 - *8$ restricted to $\mathcal{B}$ such that for all $A \in \mathcal{L}$:*

$$\mathcal{B} * A = Cn(\{B \in \mathcal{L} : \neg A \prec B\} \cup \{A\}).$$

For each revision operator $*$ from $\wp(\mathcal{L}) \times \mathcal{L}$ into $\wp(\mathcal{L})$ satisfying $*1 - *8$ and each set of sentences $\mathcal{B} \subseteq \mathcal{L}$ there is an entrenchment ordering $\preceq$ for $\mathcal{B}$ satisfying $\preceq 1 - \preceq 5$ such that for all $A \in \mathcal{L}$:

$$\mathcal{B} * A = Cn(\{B \in \mathcal{L} : \neg A \prec B\} \cup \{A\}).$$

This theorem states that the postulates for entrenchment orderings translate into the postulates for revision functions, and conversely. Caie (this volume, section 2.3) states the analogous theorem regarding the relationship between the postulates for entrenchment orderings and the postulates for contraction functions.

There is a different way of representing postulates $*1-*8$ for revision operators $*$ due to Grove (1988). Similar to Lewis' (1973) theory of counterfactuals it uses systems of spheres defined on a set of possible worlds instead of entrenchment orderings defined on a formal language (for more on counterfactuals see Briggs, this volume). A set of possible worlds can be thought of as a set of complete, or maximally specific, descriptions of the way the world could be. One approach, used by Grove (1988), is to identify possible worlds with maximally consistent sets of sentences from $\mathcal{L}$, i.e. sets of sentences that are consistent, but that become inconsistent as soon as a single new sentence is added. Another approach is to take possible worlds as primitive. For present purposes we do not have to take a stance on this issue and can assume that we are given a set of possible worlds $w_{\mathcal{L}}$ relative to which we interpret the sentences from $\mathcal{L}$.

In order to state Grove's (1988) approach it will be useful to have the following notation. $[\![A]\!] = \{\omega \in w_{\mathcal{L}} : \omega \models A\}$ is the proposition expressed by the sentence A from $\mathcal{L}$, i.e. the set of possible worlds in which the sentence A is true. $[\![\mathcal{B}]\!] = \{\omega \in w_{\mathcal{L}} : \omega \models A \text{ for all } A \in \mathcal{B}\}$ is the proposition expressed by the set of sentences $\mathcal{B} \subseteq \mathcal{L}$. In addition we need to assume that our language $\mathcal{L}$ is sufficiently rich in expressive power so that for each proposition $p \subseteq w_{\mathcal{L}}$ there is a set of sentences from $\mathcal{L}$, a "theory," $T(p)$ that expresses or means p, i.e. $[\![T(p)]\!] = p$.

Let $p \subseteq w_{\mathcal{L}}$ be a proposition and let $\mathbf{s} \subseteq \wp(w_{\mathcal{L}})$ be a set of propositions. The set $\mathbf{s}$ is a *system of spheres* in $w_{\mathcal{L}}$ that is *centered on* p if, and only if, for all propositions $q, r \subseteq w_{\mathcal{L}}$ and all sentences A from $\mathcal{L}$:

s1. If $q, r \in \mathbf{s}$, then $q \subseteq r$ or $r \subseteq q$. — $\mathbf{s}$ is nested

s2. $p \in \mathbf{s}$; and: if $q \in \mathbf{s}$, then $p \subseteq q$. — $\mathbf{s}$ is centered on p

s3. $w_{\mathcal{L}} \in \mathbf{s}$.

s4. If $[\![A]\!] \cap u \neq \emptyset$ for some $u \in \mathbf{s}$, then there is $u^* \in \mathbf{s}$ such that: $[\![A]\!] \cap u^* \neq \emptyset$, and $u^* \subseteq v$ for all $v \in \mathbf{s}$ with $[\![A]\!] \cap v \neq \emptyset$.

Requirement **s1** says that systems of spheres are *nested*: any two spheres are such that one is contained in the other, or they are the same sphere. Requirement **s2** says that the center of a system of spheres must itself be a sphere in this system, and that every other sphere in the system contains the center as a sub-sphere. Requirement **s3** says that the set of all possible worlds must be a sphere in every system of spheres. This implies that the set of all possible worlds contains every other sphere in any given system of spheres as a sub-sphere. Finally, in combination with **s3** requirement **s4** says that for each logically consistent sentence A there is a smallest sphere $u^* \in \mathbf{s}$ that properly overlaps (has a non-empty intersection) with the proposition expressed by A, $[\![A]\!]$.

Let $c_{\mathbf{s}}(A) = [\![A]\!] \cap u^*$ and define $c_{\mathbf{s}}(A) = \emptyset$ if A is logically inconsistent. Then $c_{\mathbf{s}}(A)$ is the set of possible worlds in $[\![A]\!]$ that are "closest" to the center p, where the meaning of 'closeness' is determined by the system of spheres $\mathbf{s}$. If A is logically consistent with (a set of sentences expressing) the center p, then $c_{\mathbf{s}}(A)$ is just the intersection of the center p with the set of possible worlds $[\![A]\!]$, $[\![A]\!] \cap p$. This is the easy case of expansion. The difficult case of revision arises when A is not logically consistent with (a set of sentences expressing) the center p. In this case the ideal doxastic agent has to leave the center and move to the first sphere u^* that properly overlaps with the proposition expressed by A and adopt their intersection, $[\![A]\!] \cap u^*$, as $c_{\mathbf{s}}(A)$. Figure 1 represents this situation.

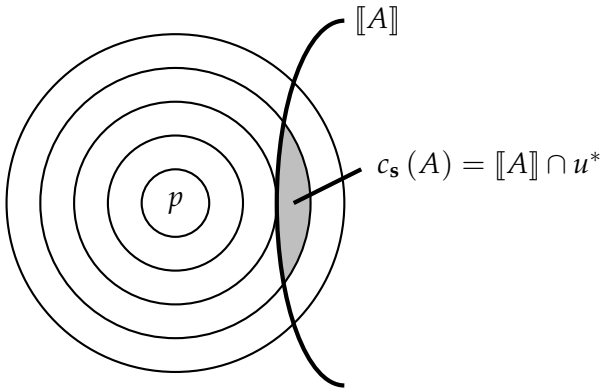


Figure 1: The possible worlds "closest" to the center p

If we have a belief set $\mathcal{B} \subseteq \mathcal{L}$ we can use a system of spheres $\mathbf{s}$ in $w_{\mathcal{L}}$ that is centered on $[\![\mathcal{B}]\!] \subseteq w_{\mathcal{L}}$ to define a revision operator $*$ restricted to $\mathcal{B}$ as follows. For every sentence A from $\mathcal{L}$:

$$\mathcal{B} * A = T\left(c_{\mathbf{s}}(A)\right).$$

The idea is that what the ideal doxastic agent ends up believing after revising $*$ her old belief set $\mathcal{B}$ with the new information A is (a set of sentences expressing) the proposition $c_{\mathbf{s}}(A)$ that contains the possible worlds in $[\![A]\!]$ that are closest when the proposition expressed by her old belief set, $[\![\mathcal{B}]\!]$, is the center. Expansion is the special case where the proposition expressed by the new information properly overlaps with the proposition expressed by the old belief set, $[\![A]\!] \cap [\![\mathcal{B}]\!] \neq \emptyset$. In this special case the ideal doxastic agent does not have to leave the old center $[\![\mathcal{B}]\!]$ of her doxastic state; it suffices if she narrows it down to the possible worlds also contained in $[\![A]\!]$. However, in the general case of revision this intersection may be empty. In this general case the ideal doxastic agent

may have to leave the innermost sphere $[\![\mathcal{B}]\!]$ and move to the smallest sphere u^* that properly overlaps with $[\![A]\!]$ and adopt their intersection, $u^* \cap [\![A]\!]$, as the new center of her doxastic state.

As before we can picture the system of spheres centered on $[\![\mathcal{B}]\!]$ as an "onion" around $[\![\mathcal{B}]\!]$. The grey area $[\![\mathcal{B} * A]\!] = [\![T\left(c_{\mathbf{s}}\left(A\right)\right)]\!] = u^* \cap [\![A]\!]$ is the logically strongest proposition the ideal doxastic agent believes after revising her old belief set $\mathcal{B}$ by the new information A; it is the new center of her doxastic state (Figure 2).

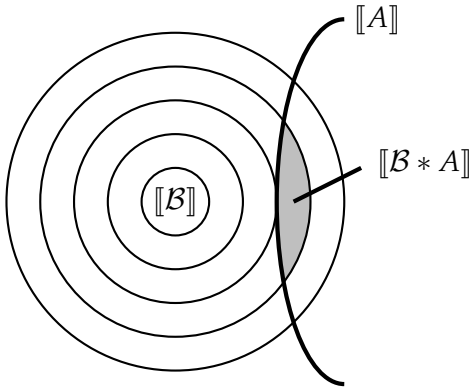


Figure 2: The strongest proposition believed after revising $\mathcal{B}$ by A

Grove (1988) proves the following theorem which states that an ideal doxastic agent can be represented as revising her beliefs by relying on a system of spheres satisfying $\mathbf{s}1-\mathbf{s}4$ if, and only if, she can be represented as revising her beliefs by employing a revision function satisfying postulates $*1-*8$.

Theorem 7 *Let $\mathcal{L}$ be a formal language, and let $w_{\mathcal{L}}$ be a set of possible worlds relative to which the sentences from $\mathcal{L}$ are interpreted and relative to which $\mathcal{L}$ is sufficiently rich. For each set of sentences $\mathcal{B} \subseteq \mathcal{L}$ and each system of spheres $\boldsymbol{s}$ in $w_{\mathcal{L}}$ that is centered on $[\![\mathcal{B}]\!]$ and satisfies $\boldsymbol{s}1-\boldsymbol{s}4$ there is a revision operator $*$ from $\{\mathcal{B}\} \times \mathcal{L}$ into $\wp(\mathcal{L})$ satisfying $*1-*8$ restricted to $\mathcal{B}$ such that for all $A \in \mathcal{L}$:*

$$\mathcal{B} * A = T\left(c_{\boldsymbol{s}}\left(A\right)\right).$$

For each revision operator $*$ from $\wp(\mathcal{L}) \times \mathcal{L}$ into $\wp(\mathcal{L})$ satisfying $*1-*8$ and each set of sentences $\mathcal{B} \subseteq \mathcal{L}$ there is a system of spheres $\boldsymbol{s}$ in $w_{\mathcal{L}}$ that is centered on $[\![\mathcal{B}]\!]$ and satisfies $\boldsymbol{s}1-\boldsymbol{s}4$ such that for all $A \in \mathcal{L}$:

$$\mathcal{B} * A = T\left(c_{\boldsymbol{s}}\left(A\right)\right).$$

The two representations of belief revision in terms of systems of spheres and in terms of belief revision functions are thus equivalent. Combined with Theorem 6 this implies that the representation of belief revision in terms of systems of spheres and in terms of entrenchment orderings are also equivalent.

As an aside let me note that Grove's (1988) notion of a system of spheres is more general than Lewis's (1973) notion in the following respect. Grove (1988) allows $\mathbf{s}$ to be centered on arbitrary propositions $p \subseteq w_{\mathcal{L}}$, whereas Lewis (1973, 14f) requires the center p to contain the actual world, and nothing but the actual world. These last two requirements are known as the principles of weak centering and of strong centering, respectively (see Briggs, this volume). In another respect Grove's (1988) notion is less general than Lewis's (1973). This is so, because requirement $\mathbf{s}4$ makes a doxastic version of the limit assumption, which Lewis (1973, 19f) famously rejects and which Herzberger (1979) shows to be equivalent to the condition that the set of counterfactual consequences $\{C \in \mathcal{L} : A \boxdot\!\!\rightarrow C\}$ of any consistent sentence A be consistent. Ranking theory also makes a doxastic version of the limit assumption.

In the AGM theory of belief revision the ideal agent's old doxastic state is represented by her belief set $\mathcal{B}$ together with her entrenchment ordering $\preceq$ for $\mathcal{B}$. The latter ordering guides the revision process in that it specifies which elements of the old belief set are given up, and which are kept, when new information D is received. The result of revising the old belief set by the new information D is a new belief set $\mathcal{B} * D$. Sophia's old belief set $\mathcal{B}$ includes the beliefs that it will rain on Tuesday, that it will be sunny on Wednesday, and that weather forecasts are always right. Suppose her belief A that it will be sunny on Wednesday is less entrenched than her belief B that it will rain on Tuesday, which in turn is less entrenched than her belief C that weather forecasts are always right, $A \prec B \prec C$.

On Monday Sophia comes to believe that the weather forecast for Tuesday and Wednesday predicts rain, D. Consequently she has to give up her belief A that it will be sunny on Wednesday or her belief C that weather forecasts are always right. The reason is that it follows from D that at least one of those two beliefs is false, i.e. $\{D\} \vdash \neg A \vee \neg C$. This implies that $A \wedge C \preceq \neg D$. Since A is less entrenched than C, i.e. $A \prec C$, A has to go. Furthermore, since $\{C, D\} \nvdash \neg B$ Sophia need not give up her belief B that it will rain on Tuesday if she holds onto her belief C that weather forecasts are always right, and adds the belief D that the weather forecast for Tuesday and Wednesday predicts rain. In addition let us assume that $\neg D \prec B$ so that Sophia's entrenchment ordering looks as follows: where $X \sim Y$ is short for $X \preceq Y$ and $Y \preceq X$,

$$\bot \sim \neg A \prec A \sim A \wedge C \preceq \neg D \prec B \prec C \prec A \vee \neg A.$$

Thus Sophia's new belief set is

$$\mathcal{B} * D = Cn\left(\{X : \neg D \prec X\} \cup \{D\}\right) = Cn\left(\{B, C, D, \neg A\}\right).$$

To Sophia's surprise it is sunny on Tuesday after all. Therefore Sophia wants to revise her newly acquired belief set $\mathcal{B} * D$ a second time by $\neg B$ to correct her belief B that it will rain on Tuesday. In addition, Sophia has to give up her belief D that the weather forecast for Tuesday and Wednesday predicts rain (this might be because she has misheard the weather forecast) or her belief C that weather forecasts are always right (this might be because she has been too gullible). The reason is that it follows from $\neg B$ that at least one of those two beliefs is false, i.e. $\{\neg B\} \vdash \neg D \vee \neg C$. Unfortunately AGM belief revision theory is of no help here. While Sophia could use her entrenchment ordering to revise her old belief set $\mathcal{B}$ to a new belief set $\mathcal{B} * D$, the entrenchment ordering itself has not been revised. Sophia's new doxastic state is silent as to whether D is now more entrenched than C (this might be because she was too gullible) or C is now more entrenched than D (this might be because she misheard the weather forecast) or C is now as entrenched as D (this might be because she was too gullible and misheard the weather forecast). However, the latter is exactly the kind of information that Sophia needs in order to revise her beliefs a second time.

3 ITERATED BELIEF REVISION

More generally, the problem is that Sophia's doxastic state is represented as a belief set plus an entrenchment ordering before the revision process, but as a belief set without an entrenchment ordering after the revision process. To handle *iterated* belief revisions the ideal agent's doxastic state has to be represented in the same way before and after the revision process. Gärdenfors and Rott (1995, p. 37) call this the "principle of categorical matching."

Nayak (1994), Boutilier (1996), Darwiche and Pearl (1997), Segerberg (1998), Fermé (2000), Rott (2003), Rott (2006), and others do exactly this (see also Caie, this volume, section 2.4). They augment the AGM postulates by additional postulates specifying how the ideal doxastic agent should revise her entrenchment ordering in addition to her belief set when she receives new information. On their accounts the ideal agent's doxastic state is represented as a belief set plus an entrenchment ordering both before and after the revision process, and both of these two elements are revised when new information is received.

Let us have a closer look at the proposal by Darwiche and Pearl (1997) (Caie, this volume, section 2.4 also discusses Boutilier 1996's proposal). In addition to postulates $*1 - *8$ they propose four more postulates for

iterated belief revision. The first of these, $*9$, says that revising an old belief set by new information (say, a conjunction) should result in the same new belief set as first revising the old belief set by a logical consequence of the new information (say, one of the two conjuncts) and then revising the resulting belief set by the new information in its entirety. That is, revision by a more specific piece of information such as that Sophia had red wine should override all changes that result from first revising the old belief set by a less specific piece of information such as that Sophia had wine.

The second of these new postulates, $*10$, says that revising an old belief set consecutively by two pieces of information that are logically inconsistent should result in the same new belief set as revising the old belief set by the second piece of information alone. That is, revision by the second piece of information—say, that Sophia had red wine—should override all changes that result from first revising the old belief by the first piece of information that is logically incompatible with the second piece of information—say, that Sophia had no wine.

Next suppose the ideal doxastic agent holds a belief after revising her old belief set by a piece of information. This may, but need not be a new belief, i.e. a belief not held previously. The third new postulate, $*11$, says that the ideal doxastic agent should also hold this belief if she first revises her old belief set by this very belief and then revises the resulting belief set by said piece of information.

Finally, suppose there is a sentence that is logically compatible with the result of revising the ideal doxastic agent's old belief set by a piece of information. The fourth new postulate, $*12$, says that this sentence should also be logically compatible with what the ideal doxastic agent ends up believing if she first revises her old belief set by this very sentence and then revises the resulting belief set by said piece of information.

More precisely, Darwiche and Pearl (1997) require the following of all sets of sentences $\mathcal{B} \subseteq \mathcal{L}$ and all sentences A and B from $\mathcal{L}$:

$*9$. If $\{A\} \vdash B$, then $(\mathcal{B} * B) * A = \mathcal{B} * A$.

$*10$. If $\{A\} \vdash \neg B$, then $(\mathcal{B} * B) * A = \mathcal{B} * A$.

$*11$. If $B \in \mathcal{B} * A$, then $B \in (\mathcal{B} * B) * A$.

$*12$. If $\neg B \notin \mathcal{B} * A$, then $\neg B \notin (\mathcal{B} * B) * A$.

In order to represent these four new postulates more perspicuously it will be helpful to consider the following reformulation of a system of spheres $\mathbf{s}$ in $w_{\mathcal{L}}$ centered on some proposition p.

Let $p \subseteq w_{\mathcal{L}}$ be a proposition and let $\leq$ be a binary relation on $w_{\mathcal{L}}$. The relation $\leq$ is an *implausibility ordering* on $w_{\mathcal{L}}$ with center p just in case the following holds for all possible worlds ω, ω', and ω'' from $w_{\mathcal{L}}$ and all propositions $q \subseteq w_{\mathcal{L}}$:

$\leq$1. $\omega \leq \omega'$ or $\omega' \leq \omega$. $\leq$ is connected

$\leq$2. If $\omega \leq \omega'$ and $\omega' \leq \omega''$, then $\omega \leq \omega''$. $\leq$ is transitive

$\leq$3. $\omega \in p$ if, and only if, for all $\omega^* \in w_{\mathcal{L}} : \omega \leq \omega^*$.

$\leq$4. If $q \neq \emptyset$, then $\{\omega \in q : \omega \leq \omega^* \text{ for all } \omega^* \in q\} \neq \emptyset$.

As suggested by its name, an implausibility ordering on $w_{\mathcal{L}}$ with center p orders the possible worlds from $w_{\mathcal{L}}$ according to their implausibility. Among other things, it is required that any two possible worlds can be compared with respect to their implausibility: either the first possible world is at least implausible as the second possible world, or the second possible world is at least as implausible as the first possible world, or both. It is also required that the ordering is transitive: if one possible world is at least as implausible as a second possible world, and the second possible world is at least as implausible as a third possible world, then the first possible world is at least as implausible as the third possible world.

Furthermore it is required that the possible worlds in the center are no more implausible than all possible worlds. That is, the center is the proposition that contains all and only the least implausible possible worlds. Finally it is required that each proposition that contains a possible world also contains a possible world that is no more implausible than all possible worlds in this proposition. The latter feature allows us to identify the implausibility of a non-empty or logically consistent proposition with the implausibility of the least implausible possible world(s) comprised by this proposition.

A system of spheres centered on p can be understood as an implausibility ordering with the center p containing the least implausible possible worlds. In terms of such an implausibility ordering the problem with the original AGM approach is the following. Before the revision process the ideal agent's doxastic state is represented as a belief set $\mathcal{B}$ plus an implausibility ordering $\leq_{\mathcal{B}}$ with center $[\![\mathcal{B}]\!]$. After revision by the new information A the ideal agent's doxastic state is represented as a belief set $\mathcal{B} * A$, but without a corresponding implausibility ordering $\leq_{\mathcal{B}*A}$. Gärdenfors and Rott's (1995) principal of categorical matching urges us to represent the ideal agent's doxastic state as a belief set plus an implausibility ordering both before and after the revision process. In these terms Darwiche and Pearl's (1997) postulates $*9 - *12$ become the following simple requirements.

First, the implausibility ordering among the possible worlds within the proposition expressed by the new information should be the same before and after a revision by the new information. Second, the implausibility ordering among the possible worlds outside of the proposition expressed by the new information should also be the same before and after a revision

by the new information. Third, if a possible world within the proposition expressed by the new information is less implausible than a possible world outside of this proposition before revision by the new information, then this should remain so after a revision by the new information. Fourth, if a possible world within the proposition expressed by the new information is at least as implausible as a possible world outside of this proposition before revision by the new information, then this should also remain so after a revision by the new information. That is, where $\omega < \omega'$ holds for arbitrary possible worlds ω and ω' from $w_{\mathcal{L}}$ if, and only if, $\omega \leq \omega'$ and $\omega' \not\leq \omega$, the following is required of all possible worlds ω and ω' from $w_{\mathcal{L}}$ and all sentences A from $\mathcal{L}$:

$\leq$5. If $\omega, \omega' \in [\![A]\!]$, then: $\omega \leq_{\mathcal{B}} \omega'$ just in case $\omega \leq_{\mathcal{B}*A} \omega'$.

$\leq$6. If $\omega, \omega' \notin [\![A]\!]$, then: $\omega \leq_{\mathcal{B}} \omega'$ just in case $\omega \leq_{\mathcal{B}*A} \omega'$.

$\leq$7. If $\omega \in [\![A]\!]$ and $\omega' \notin [\![A]\!]$ and if $\omega <_{\mathcal{B}} \omega'$, then $\omega <_{\mathcal{B}*A} \omega'$.

$\leq$8. If $\omega \in [\![A]\!]$ and $\omega' \notin [\![A]\!]$ and $\omega \leq_{\mathcal{B}} \omega'$, then $\omega \leq_{\mathcal{B}*A} \omega'$.

Before we turn to a representation theorem for iterated belief revision let us consider a third representation theorem for belief revision. Theorems 6 and 7 tell us that the representation of belief revision in terms of entrenchment orderings, in terms of belief revision functions, and in terms of systems of spheres are all equivalent. According to the following theorem due to Grove (1988) this equivalence extends to the representation of belief revision in terms of implausibility orderings.

Theorem 8 *Let $\mathcal{L}$ be a formal language, and let $w_{\mathcal{L}}$ be a set of possible worlds relative to which the sentences from $\mathcal{L}$ are interpreted and relative to which $\mathcal{L}$ is sufficiently rich. For each set of sentences $\mathcal{B} \subseteq \mathcal{L}$ and each implausibility ordering $\leq$ on $w_{\mathcal{L}}$ with center $[\![\mathcal{B}]\!]$ that satisfies $\leq$1$-\leq$4 there is a revision operator $*$ from $\{\mathcal{B}\} \times \mathcal{L}$ into $\wp(\mathcal{L})$ satisfying $*$1$-*$8 restricted to $\mathcal{B}$ such that for all $A \in \mathcal{L}$:*

$$\mathcal{B} * A = T\left(\{\omega \in [\![A]\!] : \omega \leq \omega^* \text{ for all } \omega^* \in [\![A]\!]\}\right).$$

*For each revision operator $*$ from $\wp(\mathcal{L}) \times \mathcal{L}$ into $\wp(\mathcal{L})$ that satisfies $*$1$-*$8 and each set of sentences $\mathcal{B} \subseteq \mathcal{L}$ there is an implausibility ordering $\leq$ on $w_{\mathcal{L}}$ with center $[\![\mathcal{B}]\!]$ satisfying $\leq$1$-\leq$4 such that for all $A \in \mathcal{L}$:*

$$\mathcal{B} * A = T\left(\{\omega \in [\![A]\!] : \omega \leq \omega^* \text{ for all } \omega^* \in [\![A]\!]\}\right).$$

The complicated looking proposition $\{\omega \in [\![A]\!] : \omega \leq \omega^* \text{ for all } \omega^* \in [\![A]\!]\}$ is simply the set of the least implausible possible worlds in which the new information A is true. This means that the belief set $\mathcal{B} * A$ that results from

revising $*$ the ideal doxastic agent's old belief set $\mathcal{B}$ by new information A expresses the proposition that is comprised by the least implausible A-worlds.

Against this background we can now state the following representation theorem for iterated belief revision due to Darwiche and Pearl (1997). According to it the representation of iterated belief revision in terms of belief revision functions à la postulates $*9-*12$ is equivalent to the simple representation of iterated belief revision in terms of implausibility orderings à la $\leq 5-\leq 8$.

Theorem 9 *Let $\mathcal{L}$ and $w_{\mathcal{L}}$ be as in Theorem 8. Suppose $*$ is a revision operator from $\wp(\mathcal{L}) \times \mathcal{L}$ into $\wp(\mathcal{L})$ that satisfies $*1-*8$. According to Theorem 8, there exists a family of implausibility orderings $(\leq_{\mathcal{B}})_{\mathcal{B}\subseteq\mathcal{L}}$ on $w_{\mathcal{L}}$ such that for each set of sentences $\mathcal{B} \subseteq \mathcal{L}$: $\leq_{\mathcal{B}}$ satisfies $\leq 1-\leq 4$ and is such that, for all sentences $A \in \mathcal{L}$, $\mathcal{B} * A = T\left(\{\omega \in [\![A]\!] : \omega \leq_{\mathcal{B}} \omega^* \text{ for all } \omega^* \in [\![A]\!]\}\right)$. For this $*$ and any one of these families $(\leq_{\mathcal{B}})_{\mathcal{B}\subseteq\mathcal{L}}$: $*$ satisfies $*9-*12$ if, and only if, for every set of sentences $\mathcal{B} \subseteq \mathcal{L}$, $\leq_{\mathcal{B}}$ satisfies $\leq 5-\leq 8$.*

The approaches to iterated belief revision mentioned above all have in common that the ideal agent's doxastic state is represented as a belief set plus an entrenchment ordering/system of spheres/implausibility ordering both before and after the revision process. Furthermore these approaches have in common that the new information is represented as a single sentence (or a single proposition). The latter is also true for the approach by Jin and Thielscher (2007) discussed below, but not for what Rott (2009) calls "two-dimensional" belief revision operators (see also Cantwell 1997; Fermé and Rott 2004; Rott 2007).

In one-dimensional belief revision, as we may call it, the new information comes as a "naked" (Rott, 2007) sentence or proposition. It is the job of the various belief revision methods, as opposed to the new information itself, to say exactly where in the new entrenchment ordering/system of spheres/implausibility ordering the new sentence or proposition should be placed. These belief revision methods include lexicographic revision (Nayak, 1994), natural revision (Boutilier, 1996), irrevocable revision (Segerberg, 1998; Fermé, 2000), irrefutable revision (Rott, 2006), and still others. In two-dimensional belief revision it is the new information itself that carries at least part of this information. Here the new information does not say *that* the input sentence A is *true* (so should be accepted according to the Success postulate). Instead it specifies, at least to some extent, *how firmly* A is *accepted* or *believed* by specifying that, in the new entrenchment ordering $\preceq^*$, A is at least as entrenched as some "reference sentence" B. Thus the new information is now of the form: $A \preceq^* B$. (For the purposes of this entry we may ignore "non-prioritized" belief revision, where the

new information need not be accepted. See Hansson, Fermé, Cantwell, and Falappa, 2001.)

Let us return to our example. On Monday Sophia comes to believe that the weather forecast for Tuesday and Wednesday predicts rain, D. In one-dimensional belief revision she picks one of the iterated belief revision methods mentioned above. Then she revises her old belief set $\mathcal{B}$ and entrenchment ordering $\preceq_{\mathcal{B}}$ to obtain a new belief set $\mathcal{B} * D$ and a new entrenchment ordering $\preceq_{\mathcal{B}*D}$. Different methods return different outputs, but on all of them Sophia ends up believing that it will rain on Tuesday, B. On Tuesday Sophia sees that it is sunny and so receives the new information that it does not rain after all, $\neg B$. In one-dimensional belief revision Sophia proceeds as before.

In two-dimensional belief revision Sophia does not receive the qualitative information $\neg B$ about Tuesday's weather. Instead she receives the comparative information $C \preceq^* \neg B$ about her new doxastic state. This piece of new information says that, in her new entrenchment ordering $\preceq^*$, the claim that it does not rain on Tuesday is at least as entrenched as the claim that weather forecasts are always right, indicating that she trusts her sight at least as much as the weatherperson (we could take a reference sentence other than C).

Now, there still are several belief revision methods to choose from (see Rott 2009). Among others, this reflects the fact that Sophia can respect the constraint $C \preceq^* \neg B$ by lowering the doxastic status of C, or by raising the doxastic status of $\neg B$. However, the new information now is more specific and leaves less room to be filled by the revision method. It is then only a small, but crucial step to equip Sophia with the quantitative, numerical information that $\neg B$ is entrenched to a specific degree. In this case the new information completely determines exactly where $\neg B$ is located in the new entrenchment ordering on its own, without the help of the revision method. The latter merely has to incorporate this new information into Sophia's old doxastic state in a consistent way. Ranking theory does exactly this.

Before presenting ranking theory let us return to the qualitative approaches to iterated belief revision. Postulates $*1 - *12$ are still compatible with many conflicting belief revision methods. Jin and Thielscher (2007) attempt to remedy this situation by additionally requiring the ideal doxastic agent to consider the new information B to be *independent* of a sentence A after revision by B if she considered B to be independent of A before revision by B. In other words, revision should preserve independencies. While the idea behind Jin and Thielscher (2007)'s proposal seems to be correct, their actual requirement turns out to be too strong. The reason is that their notion of dependence is too strong in the sense that too many sentences are rendered independent of other sentences.

According to Jin and Thielscher (2007) a believed sentence A is independent of another sentence B if the believed sentence A is still believed after revision by the negation of the other sentence, $\neg B$. However, I can receive new information $\neg B$—say, that the captain of my favorite soccer team will not be fit for the match—whose negation $\neg\neg B$ is positively relevant to, and so *not* independent of, a belief of mine A—say, that my favorite soccer team will win the match—without making me give up this belief of mine altogether. More generally, the ways two sentences can depend on each other are many and varied, and the qualitative and comparative notions of AGM belief revision theory and its refinements seem to be too coarse-grained to capture these dependencies. Hild and Spohn (2008) argue axiomatically, and we will see in the next section, that, in order to adequately represent all dependencies, and to handle iterated belief revisions, one has to go all the way from qualitative belief sets and comparative entrenchment orderings/systems of spheres/implausibility orderings to quantitative, numerical ranking functions.

4 RANKING THEORY

Ranking functions are introduced by Spohn (1988, 1990) to represent qualitative conditional belief. A comprehensive overview can be found in Spohn (2012). The theory is quantitative or numerical in the sense that ranking functions assign numbers, so-called *ranks*, to sentences or propositions. These numbers are needed for the definition of conditional ranking functions which represent conditional beliefs. As we will see, once conditional ranking functions are defined we can interpret everything in purely qualitative, albeit conditional terms. The numbers assigned by conditional ranking functions are called *conditional ranks*. They are defined as differences of non-conditional ranks.

Instead of taking the objects of belief to be sentences of a formal language it is both more general and more convenient to take them to be propositions of some field or algebra over a set of possible worlds W. Here is the relevant definition. A set of subsets of W, $\mathcal{A} \subseteq \wp(W)$, is an *algebra over* W if, and only if,

(i) the empty or contradictory set $\emptyset$ is a proposition in $\mathcal{A}$,

(ii) if A is a proposition in $\mathcal{A}$, then the complement or negation of A, $W \setminus A = \overline{A}$, is also a proposition in $\mathcal{A}$, and

(iii) if both A and B are propositions in $\mathcal{A}$, then the union or disjunction of A and B, $A \cup B$, is also a proposition in $\mathcal{A}$.

An algebra $\mathcal{A}$ over W is a *σ-algebra* if, and only if, the following holds for every countable set $\mathcal{B} \subseteq \wp(W)$: if all the members or elements of

$\mathcal{B}$ are propositions in $\mathcal{A}$, i.e. if $\mathcal{B} \subseteq \mathcal{A}$, then the union or disjunction of the elements of $\mathcal{B}$, $\bigcup \mathcal{B}$, is also a proposition in $\mathcal{A}$. Finally, an algebra $\mathcal{A}$ over W is *complete* if, and only if, the following holds for every (countable or uncountable) set $\mathcal{B} \subseteq \wp(W)$: if all the members or elements of $\mathcal{B}$ are propositions in $\mathcal{A}$, i.e. if $\mathcal{B} \subseteq \mathcal{A}$, then the union or disjunction of the elements of $\mathcal{B}$, $\bigcup \mathcal{B}$, is also a proposition in $\mathcal{A}$. The power-set of a set of possible worlds W, $\wp(W)$, is a complete algebra over W.

A function $\varrho : \mathcal{A} \to \mathbb{N} \cup \{\infty\}$ from an algebra of propositions $\mathcal{A}$ over a non-empty set of possible worlds W into the set of natural numbers $\mathbb{N}$ extended by infinity ∞, $\mathbb{N} \cup \{\infty\}$, is a *ranking function* on $\mathcal{A}$ just in case for all propositions A and B from $\mathcal{A}$:

$$\begin{aligned} \varrho(W) &= 0, && (1)\\ \varrho(\emptyset) &= \infty, && (2)\\ \varrho(A \cup B) &= \min\{\varrho(A), \varrho(B)\}. && (3) \end{aligned}$$

As in probability theory, if $\mathcal{A}$ is a σ-algebra, axiom (3) can be strengthened to countable unions. The resulting ranking function is called "countably minimitive." In contrast to probability theory, if $\mathcal{A}$ is a complete algebra, axiom (3) can even be strengthened to arbitrary unions. The resulting ranking function is called "completely minimitive."

For a non-empty or consistent proposition $A \neq \emptyset$ from $\mathcal{A}$ the conditional ranking function $\varrho(\cdot \mid A) : \mathcal{A} \setminus \{\emptyset\} \to \mathbb{N} \cup \{\infty\}$ based on the (non-conditional) ranking function $\varrho(\cdot) : \mathcal{A} \to \mathbb{N} \cup \{\infty\}$ is defined as

$$\varrho(\cdot \mid A) = \begin{cases} \varrho(\cdot \cap A) - \varrho(A), & \text{if} \quad \varrho(A) < \infty, \\ \infty \text{ or } 0, & \text{if} \quad \varrho(A) = \infty. \end{cases}$$

For the case where $\varrho(A) = \infty$ Goldszmidt and Pearl (1996, p. 63) suggest ∞ as the value for $\varrho(B \mid A)$ for all propositions B from $\mathcal{A}$. For this case Huber (2006, p. 464) suggests 0 as the value for $\varrho(B \mid A)$ for all non-empty or consistent propositions B from $\mathcal{A}$ and additionally stipulates $\varrho(\emptyset \mid A) = \infty$ to ensure that every conditional ranking function on $\mathcal{A}$ is a ranking function on $\mathcal{A}$.

A ranking function ϱ is *regular* if, and only if,

$$\varrho(A) < \varrho(\emptyset) = \infty,$$

for all non-empty or consistent propositions A from $\mathcal{A}$. In contrast to probability theory it is always possible to define a regular ranking function, no matter how rich or fine-grained the underlying algebra of propositions (see Hájek, manuscript).

Ranks are interpreted doxastically as grades of disbelief. A proposition A is disbelieved just in case A is assigned a positive rank, $\varrho(A) > 0$. A

proposition that is not disbelieved is assigned rank 0, but this does not mean that it is believed. Instead, belief in a proposition is characterized as disbelief in its negation: a proposition A is believed just in case the negation of A, $\overline{A}$, is disbelieved, $\varrho\left(\overline{A}\right) > 0$. An agent suspends judgment with respect to a proposition (and its negation) if, and only if, both the proposition and its negation are assigned rank 0.

A proposition A is disbelieved conditional on a proposition C just in case A is assigned a positive rank conditional on C, $\varrho\left(A \mid C\right) > 0$. A proposition A is believed conditional on a proposition C just in case the negation of A, $\overline{A}$, is disbelieved conditional on C, $\varrho\left(\overline{A} \mid C\right) > 0$. It takes getting used to read positive numbers in this "negative" way, but mathematically this is the simplest way to axiomatize ranking functions.

Note that it follows from Huber's (2006) definition of a conditional ranking function that the ideal doxastic agent should not disbelieve a proposition A conditional on itself, $\varrho\left(A \mid A\right) = 0$, if, and only if, A is non-empty or consistent.

In doxastic terms the first axiom says that the ideal doxastic agent should not disbelieve the tautological proposition W. The second axiom says that she should disbelieve the empty or contradictory proposition $\emptyset$ with maximal strength ∞. Given the definition of conditional ranks, the second axiom can also be read in purely qualitative, albeit conditional terms: in these terms it says that the ideal doxastic agent should disbelieve the empty or contradictory proposition conditional on any non-empty or consistent proposition. It follows that the ideal doxastic agent should believe the tautological proposition with maximal strength, or conditional on any non-empty or consistent proposition.

Part of what the third axiom says is that the ideal doxastic agent should disbelieve a disjunction $A \cup B$ just in case she disbelieves both its disjuncts A and B. Given the definition of conditional ranks, the third axiom extends this requirement to conditional beliefs. As noted above, the ideal doxastic agent should not disbelieve a non-empty or consistent proposition conditional on itself. Given this consequence of the definition of a conditional ranking function, the third axiom says—in purely qualitative, albeit conditional terms—the following. For all non-empty or consistent propositions C, the ideal doxastic agent should disbelieve a disjunction $A \cup B$ conditional on C just in case she disbelieves A conditional on C and she disbelieves B conditional on C. Countably and completely minimitive ranking functions extend this "conditional consistency" requirement to countable and arbitrary unions, respectively. For any non-empty or consistent proposition C, the ideal doxastic agent should disbelieve $\bigcup \mathcal{B}$ conditional on C just in case she disbelieves each disjunct B from $\mathcal{B}$ conditional on C. We thus see that all that axioms (1)–(3) of ranking theory ask

of the ideal doxastic agent is that her beliefs be consistent, and that her conditional beliefs be conditionally consistent.

Ranks are numerical, but unlike probabilities, which are measured on an absolute scale, ranks do not utilize all the information carried by these numbers. Instead, ranks are at best measured on a ratio scale (Hild & Spohn, 2008)—at best, because even the choice of 0 as threshold for disbelief is somewhat arbitrary, as Spohn (2015, p. 9) notes (but see Raidl, 2018, for subtle differences for conditional belief). Some positive, but finite natural number would do just as well. This is perhaps most perspicuous by considering what Spohn (2012) calls the *two-sided* ranking function $\beta : \mathcal{A} \to \mathbb{Z} \cup \{\infty\} \cup \{-\infty\}$ whose range is the set of integers $\mathbb{Z}$ extended by plus infinity ∞ and minus infinity $-\infty$, $\mathbb{Z} \cup \{\infty\} \cup \{-\infty\}$. β is defined in terms of ϱ as follows: for all propositions A in $\mathcal{A}$, $\beta(A) = \varrho\left(\overline{A}\right) - \varrho(A)$. Ranking functions and two-sided ranking functions are interdefinable. The latter are more difficult to axiomatize, but they may be more intuitive, because they characterize belief in positive terms as follows.

A proposition A is believed if, and only if, its two-sided rank is positive, $\beta(A) > 0$. A proposition A is believed conditional on a proposition C if, and only if, its two-sided conditional rank is positive, $\beta(A \mid C) > 0$. Interestingly, any other finite threshold equally gives rise to a notion of belief (that is consistent and deductively closed as explained below): a proposition is believed if, and only if, its rank is greater than some finite, non-negative threshold n, $\beta(A) > n$. Hence ranking theory validates the Lockean thesis (Foley, 2009; Hawthorne, 2009). Furthermore, while it may appear unfair to reserve infinitely many numbers for belief and for disbelief, and only the number 0 for suspension of judgment, we now see that this is not essential to the theory and can be fixed by adopting a threshold other than 0 (there are still only finitely many levels for suspension of judgment, though).

Doxastically interpreted, axioms (1)–(3) are synchronic norms for how an ideal doxastic agent should organize her beliefs and conditional beliefs at a given moment in time. These axioms are supplemented by diachronic norms for how she should update her beliefs and conditional beliefs over time if new information of various formats is received. The first update rule is defined for the case where the new information comes in the form a certainty. It mirrors the update rule of strict conditionalization from probability theory (Vineberg, 2000).

Update Rule 1 (Plain Conditionalization, Spohn 1988) *If* $\varrho(\cdot) : \mathcal{A} \to \mathbb{N} \cup \{\infty\}$ *is the ideal doxastic agent's ranking function at time* t*, and between* t *and the later time* t' *her ranks for* E *and* $\overline{E}$ *from* $\mathcal{A}$ *are directly affected and she becomes certain of* E*, but no logically stronger proposition (i.e. her rank for* E *at* t *is finite, and* E *is the logically strongest proposition for whose negation* $\overline{E}$ *her*

rank at t' is ∞), and her ranks are not directly affected in any other way such as forgetting etc., then her ranking function at t' should be $\varrho_E(\cdot) = \varrho(\cdot \mid E)$.

Plain conditionalization asks the ideal doxastic agent to revise her beliefs and conditional beliefs by holding onto those conditional beliefs whose condition is the most specific, i.e. logically strongest, proposition she became certain of, subject to the constraint that the beliefs and conditional beliefs in the resulting new belief set are consistent and conditionally consistent, respectively. In slightly different terminology we can say that plain conditionalization has the ideal agent revise her doxastic state by holding onto those inferential beliefs whose premise is the logically strongest proposition she became certain of as a result of some experiential event that is not under her doxastic control.

The second update rule is defined for the case where the new information comes in the form of new ranks for the elements of a partition. It mirrors the update rule of Jeffrey conditionalization from probability theory (Jeffrey, 1983).

Update Rule 2 (Spohn Conditionalization, Spohn 1988) *If $\varrho(\cdot) : \mathcal{A} \to \mathbb{N} \cup \{\infty\}$ is the ideal doxastic agent's ranking function at time t, and between t and the later time t' her ranks on the experiential partition $\{E_i \in \mathcal{A} : i \in I\}$ are directly affected and change **to** $n_i \in \mathbb{N} \cup \{\infty\}$ with $\min\{n_i : i \in I\} = 0$, and $n_i = \infty$ if $\varrho(E_i) = \infty$, and her ranks are not directly affected on any finer partition or in any other way such as forgetting etc., then her ranking function at t' should be $\varrho_{E_i \to n_i}(\cdot)$,*

$$\varrho_{E_i \to n_i}(\cdot) = \min_{i \in I} \left\{ \varrho(\cdot \mid E_i) + n_i \right\}.$$

Spohn conditionalization asks the ideal doxastic agent to revise her beliefs and conditional beliefs by holding onto those conditional beliefs whose conditions are the most specific propositions whose doxastic standing has changed as a result of some experiential event that is not under her doxastic control, subject to the constraint that the beliefs and conditional beliefs in the resulting new belief set are consistent and conditionally consistent, respectively. The restriction to hold fixed only those conditional beliefs whose conditions are the most specific propositions whose doxastic standing has been directly affected is important.

Suppose you hold the conditional beliefs that Sophia will have white wine tonight if there is wine left, and that she will have red wine tonight if there is red wine left, but no white wine—say, because you believe that Sophia prefers having white wine to having red wine to having no wine. Suppose further you subsequently come to believe, as a result of being told so by a source you deem reliable, that there is red wine left, but no white wine. Since your beliefs are deductively closed you also come to

believe that there is wine left. In this case you should *not* hold onto your conditional belief that Sophia will have white wine tonight if there is wine left. Instead, you should only hold onto your conditional belief that Sophia will have red wine tonight if there is red wine left, but no white wine. The same is true if you subsequently do not merely come to believe, but become certain in this way that there is red wine left, but no white wine. This is the reason for the restriction in plain conditionalization to hold fixed only those conditional beliefs whose condition is the logically strongest proposition the ideal doxastic agent becomes certain of. Furthermore, this illustrates that plain conditionalization is the special case of Spohn conditionalization where the experiential partition is $\left\{E, \overline{E}\right\}$ and where the new ranks are 0 and ∞, respectively.

The third update rule is defined for the case where the new information reports the differences between the old and the new ranks for the elements of a partition. It mirrors the update rule of Field conditionalization from probability theory (Field, 1978) and is developed further in Bewersdorf (2013) .

Update Rule 3 (Shenoy Conditionalization, Shenoy 1991) *If* $\varrho(\cdot) : \mathcal{A} \to \mathbb{N} \cup \{\infty\}$ *is the ideal doxastic agent's ranking function at time* t*, and between* t *and the later time* t' *her ranks on the experiential partition* $\{E_i \in \mathcal{A} : i \in I\}$ *are directly affected and change* ***by*** $z_i \in \mathbb{N}$*, where* $\min\{z_i : i \in I\} = 0$*, and her ranks are not directly affected on any finer partition or in any other way such as forgetting etc., then her ranking function at* t' *should be* $\varrho_{E_i \uparrow z_i}(\cdot)$*,*

$$\varrho_{E_i \uparrow z_i}(\cdot) = \min_{i \in I}\left\{\varrho(\cdot \cap E_i) + z_i - m\right\},$$

where $m = \min_{i \in I}\left\{z_i + \varrho(E_i)\right\}$.

Spohn conditionalizing E and $\overline{E}$ to 0 and n, respectively, keeps the relative positions of all possible worlds in E and all possible worlds in $\overline{E}$ fixed. It improves the rank of E to 0 and changes the rank of $\overline{E}$ to n. Shenoy conditionalizing E and $\overline{E}$ by 0 and n, respectively, improves the possibilities within E by n, as compared to the possibilities in $\overline{E}$. The value m is a normalization parameter ensuring that at least one possible world is assigned rank zero so that the result is a ranking function.

Spohn and Shenoy conditionalization can be defined in terms of each other. Their difference lies in the interpretation of the input parameters. Spohn conditionalization is *result-oriented* in the sense that the numbers n_i characterize the result of the experiential event on the agent's ranks for the propositions E_i. The latter depend in part on the agent's initial ranks, which is why the numbers n_i do not characterize the impact of the experiential event as such, independently of the agent's initial beliefs. In contrast to this the numbers z_i in Shenoy conditionalization do characterize

the impact of the experiential event as such, independently of the agent's initial beliefs. They do so in the sense that the rank of E_i is deteriorated by z_i relative to the rank of the "best" cell. Note that, when there are more than two cells, the latter need not be the cell with the lowest initial rank.

In the case of both Spohn and Shenoy conditionalization the new information consists of a (partition of) proposition(s) together with a (list of) number(s). This reflects the fact that the quality of new information varies with how reliable or trustworthy the agent deems its source: it makes a difference if the weatherperson who Sophia does not know predicts that it will rain, if a friend Sophia trusts tells her so, or if Sophia sees for herself that it is raining. In each case the proposition Sophia comes to believe is that it is raining, but the effect of the new information on her old beliefs will be a different one in each case. The difference in how reliable or trustworthy Sophia deems the sources of the new information is reflected in the numbers accompanying this proposition.

All that axioms (1)–(3) ask of the ideal doxastic agent is that her beliefs be consistent, and that her conditional beliefs be conditionally consistent. We will see below that all that update rules (1)–(3) ask of her is that her beliefs remain consistent, and that her conditional beliefs remain conditionally consistent.

Sophia's ranking function r will assign a positive rank to the proposition $\overline{[\![A]\!]}$ that it will not be sunny on Wednesday. Her ranking function r will assign a greater rank to the proposition $\overline{[\![B]\!]}$ that it will not rain on Tuesday. Her ranking function r will assign an even greater rank to the proposition $\overline{[\![C]\!]}$ that weather forecasts are not always right so that

$$0 < r\left(\overline{[\![A]\!]}\right) < r\left(\overline{[\![B]\!]}\right) < r\left(\overline{[\![C]\!]}\right).$$

More generally, for regular ranking functions r, the ordering $A \preceq_r B$ on $\mathcal{L}$ just in case

$$r\left(\overline{[\![A]\!]}\right) \leq r\left(\overline{[\![B]\!]}\right)$$

is an entrenchment ordering for

$$\mathcal{B} = \left\{C \in \mathcal{L} : r\left(\overline{[\![C]\!]}\right) > 0\right\}.$$

In what follows I will assume that r is regular.

In other words, the set of propositions

$$\mathbf{s}_r = \left\{r^{-1}(n) \subseteq W : n \in \mathbb{N}\right\}$$

is a system of spheres in W centered on $r^{-1}(0)$, where

$$r^{-1}(n) = \{\omega \in W : r(\{\omega\}) = n\}$$

is the set of possible worlds that are assigned rank n. In still other words, the ordering $\omega \leq_r \omega'$ on W just in case $r\left(\{\omega\}\right) \leq r\left(\{\omega'\}\right)$ is an implausibility ordering on W with the center being the conjunction or intersection of all beliefs,

$$\bigcap \left\{ [\![C]\!] \subseteq W : r\left(\overline{[\![C]\!]}\right) > 0 \right\} = \{\omega \in W : r\left(\{\omega\}\right) = 0\}.$$

(I make the simplifying assumption that the algebra of propositions $\mathcal{A}$ is the power set of W. If this assumption is not made, these definitions are slightly more complicated.) Therefore ranking theory satisfies the postulates of AGM belief revision theory. It also satisfies the four additional postulates $*9 - *12$ for iterated belief revision proposed by Darwiche and Pearl (1997). This can easily be verified by checking that the four postulates $\leq 5 - \leq 8$ hold for $\leq_r$ (see also Spohn 2012, chapter 5.6). In what follows I will suppress '$[\![]\!]$' and denote propositions by capital letters.

When Sophia comes to believe on Monday that the weather forecast for Tuesday and Wednesday predicts rain, she has to tell us how strongly she now disbelieves the proposition $\overline{D}$ that the weather forecast for Tuesday and Wednesday does not predict rain in order for Spohn conditionalization to tell her how to revise her beliefs. As an approximation it suffices if she tells us how many information sources saying $\overline{D}$ it would now take for her to give up her disbelief $\overline{D}$, as compared to how many information sources saying X it would then have taken for her to give up her disbelief that X for $X = A, B, C, D, \overline{A}, \overline{B}, \overline{C}, \overline{D}$. Suppose Sophia's old ranks are $r(\overline{A}) = 1$, $r(D) = 2$, $r(\overline{B}) = 5$, and $r(\overline{C}) = 7$, and her new rank is $r^*(\overline{D}) = 13$. According to Spohn conditionalization Sophia's new ranks are:

$$r^*\left(X\right) = \min\left\{ r\left(X \mid D\right) + 0, r\left(X \mid \overline{D}\right) + 13 \right\}.$$

In order to calculate Sophia's new ranks $r^*(X)$ we thus need her old conditional ranks $r(X \mid D)$ and $r(X \mid \overline{D})$ as well as her new ranks for the conditions D and $\overline{D}$. This in turn requires us to determine her old ranks for various conjunctions. Suppose the numbers are as in Figure 3. Then Sophia's new ranks are $r^*(\overline{C}) = 6$, $r^*(\overline{B}) = 7$, $r^*(A) = 7$, $r^*(\overline{D}) = 13$.

Note that C is a proposition Sophia believes both before and after revision by D, $r(\overline{C}) > 0$ and $r^*(\overline{C}) > 0$, although $\overline{D}$ is positively relevant to, and so not independent of, C in the sense that $r(\overline{C} \mid \overline{D}) = 7 > 6 = r(\overline{C} \mid D)$. In other words, Sophia receives new information D whose negation is positively relevant to, and so not independent of, her belief that C without making her give up her belief that C. On the other hand, if Sophia considers $\overline{D}$ independent of a proposition X before revision by D, then she also does so after revision by D. More generally, suppose two propositions A and B are independent according to a ranking function r, $r(A \mid B) = r(A \mid \overline{B})$

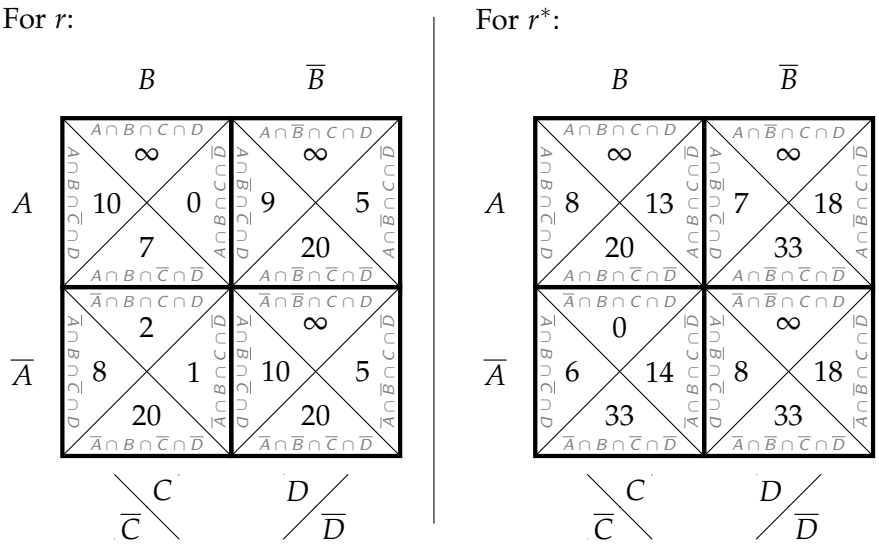


Figure 3: Sophia's old and new ranks for various conjunctions

and $r(\overline{A} \mid B) = r(\overline{A} \mid \overline{B})$. In this case A and B are independent according to any ranking function r^* that results from r by what we may call a "Spohn shift" on the partition $\{B, \overline{B}\}$, i.e. the result of Spohn conditionalization on this partition for an arbitrary pair of natural numbers.

This feature, which is known as *rigidity*, vindicates the idea behind Jin and Thielscher (2007)'s proposal that revision should preserve independencies. It does so by fixing their notion of independence. For more on the definition of rank-theoretic independence see Spohn (1999). As an aside let me note that, while rigidity is generally considered to be a desirable feature of an update rule, Weisberg (2009, 2015) uses rigidity to argue that neither Bayesianism nor Dempster-Shafer theory (Haenni, 2009) nor ranking theory can handle a phenomenon he terms *perceptual undermining*. Huber (2014a) defends these theories against Weisberg's charge.

Spohn conditionalization gives Sophia a complete new ranking function r^* that she can use to revise her newly acquired belief set

$$\mathcal{B}^* = \left\{ X \in \mathcal{A} : r^*\left(\overline{X}\right) > 0 \right\}$$

a second time when she learns on Tuesday that it is sunny after all. All she has to do is tell us how strongly she then disbelieves the proposition B that it will rain on Tuesday. If $r^{**}(B) = 13$, her newer ranks are $r^{**}(\overline{A}) = 1$, $r^{**}(C) = 11$, $r^{**}(\overline{D}) = 11$. See Figure 4.

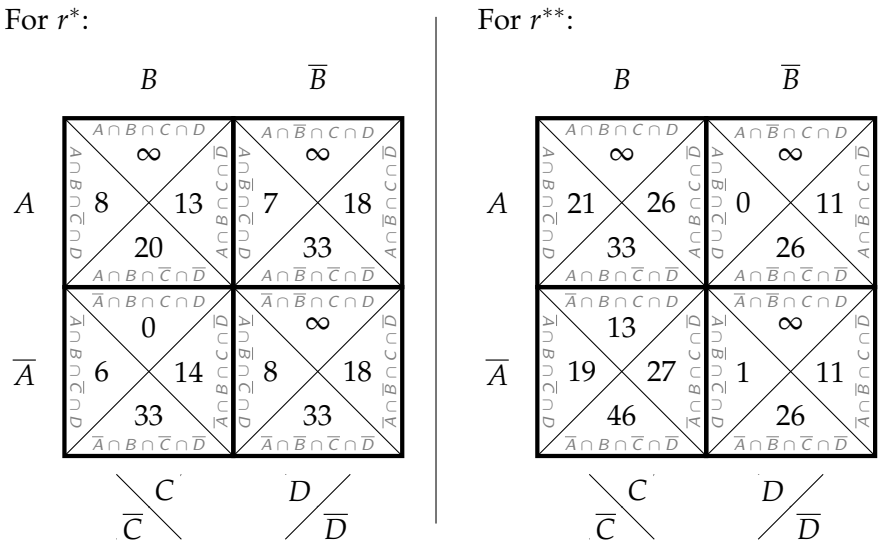


Figure 4: Sophia's new and newer ranks for various conjunctions

This means that Sophia did not mishear the weather forecast, but was too gullible (or so we assume for purposes of illustration), and so has to give up her belief C that weather forecasts are always right. In addition she also has to regain her belief A that it will be sunny on Wednesday.

At the end of this section Sophia's doxastic career is pictured as a sequence of "onions." The difference to the AGM theory is that, in ranking theory, the layers carry numbers which reflect how far apart they are from each other according to the ideal agent's doxastic state. A different way to picture the situation is to allow for *empty* layers and to have one, possibly empty, layer for each natural number.

We see that ranking theory handles indefinitely iterated belief revisions. It does so in contrast to the AGM theory of belief revision. However, it does so also in contrast to probability theory. As yet another aside, let me briefly explain why. In probability theory the ideal doxastic agent is sometimes forced to assign probability 0 to some non-empty or consistent proposition. In order to enable her to learn such propositions the ideal doxastic agent is usually represented by a Popper-Rényi measure which is more general than a classical probability (Popper, 1955; Rényi, 1955; Stalnaker, 1970; Spohn, 1986; Easwaran, this volume). However, as already noted by Harper (1976), Popper-Rényi measures violate the principal of categorical matching and so cannot handle iterated revisions of degrees of belief: the result of conditionalizing a Popper-Rényi measure is not another Popper-Rényi measure, but a classical probability; and as Boutilier (1995)

notes, there is no straightforward analogue of Jeffrey conditionalization for Popper-Rényi measures. Spohn (2006b) provides an even more general notion of probability, *ranked probability*, which results from making probabilities the objects of rank-theoretic belief. It handles iterated revisions of probabilistic degrees of belief and satisfies the principal of categorical matching: the result of conditionalizing a ranked probability is another ranked probability.

Ranking theory is a normative theory that addresses the question how an ideal doxastic agent should organize her beliefs and conditional beliefs at a given moment in time, and how she should revise these beliefs across time if she receives new information of various formats. Why should an ideal doxastic agent obey the norms of ranking theory? That is, why should an ideal doxastic agent organize her beliefs and conditional beliefs at a given moment in time according to axioms (1)–(3)? And why should she update her beliefs and conditional beliefs across time according to update rules (1)–(3) if she receives new information of the appropriate format? Who are we, Sophia asks, to tell her what—or rather: how—to believe? To answer these questions, and to respond to Sophia, we need a bit of terminology.

An ideal doxastic agent's *grade of entrenchment* for a proposition A is defined as the smallest number n such that she would give up her disbelief in A if she received the information A from n sources she deemed independent and minimally positively reliable, *mp-reliable*, about A, and this was all that directly affected her doxastic state. If the ideal doxastic agent does not disbelieve A to begin with, her grade of entrenchment for A is 0. Her grade of entrenchment for A is higher, the more information sources of the sort described it would take for her to give up her disbelief in A.

As mentioned previously, whereas probabilities are measured on an absolute scale, ranks are at best measured on a ratio scale. The same is true for grades of entrenchment. Therefore we need to fix a *unit* for these grades of entrenchment. We need to do the same when we want to report the amount of money in your bank account, which is measured on a ratio scale, or the temperature in Vienna on January 1, 2018, which is measured on an interval scale. To say that the amount of money in your bank account, or the temperature in Vienna on January 1, 2018, equals 17 is not saying anything if we do not also specify a unit such as Euros or degrees of Celsius. Information sources that are deemed mp-reliable are used to define the unit in which grades of entrenchment are reported. Furthermore, to guarantee that these units can be added and compared, just as we can add and compare sums of Euros and degrees of Celsius, we need to make sure that these information sources are not only deemed

to be mp-reliable by the ideal doxastic agent, but also independent in the relevant sense.

We non-ideal doxastic agents generally do not deem our sources of information independent or mp-reliable. One expert's saying A will sometimes make us stop disbelieving A immediately, while the sermons of a dozen others won't. And the last-born's telling a parent that there is no red wine left after the first-born has already confessed to drinking it up won't make much of a difference to the parent's grade of disbelief that there is red wine left. However, this is no argument against the usefulness of this notion. Information sources that are deemed independent and mp-reliable are a theoretical construct that are assumed or postulated to exist. They are the smallest units such that the reliability one deems any possible information source to possess can be expressed as a multiple of them.

Let ϱ be the ideal doxastic agent's entrenchment function, i.e. the function that summarizes her grades of entrenchment for all propositions from $\mathcal{A}$. Her *belief set* $\mathcal{B}_\varrho$ is the set of propositions with a positive grade of entrenchment,

$$\mathcal{B}_\varrho = \left\{ A \in \mathcal{A} : \varrho\left(\overline{A}\right) > 0 \right\}.$$

Her belief set conditional on the consistent proposition C is the set of propositions with a positive grade of entrenchment conditional on C,

$$\mathcal{B}_{\varrho(\cdot|C)} = \left\{ A \in \mathcal{A} : \varrho\left(\overline{A} \mid C\right) > 0 \right\}.$$

$\mathcal{B} \subseteq \mathcal{A}$ is *consistent in the finite/countable/complete sense* if, and only if, for every finite/countable/arbitrary $\mathcal{B}^- \subseteq \mathcal{B}$, $\bigcap \mathcal{B}^- \neq \emptyset$. It is *deductively closed in the finite/countable/complete sense* if, and only if, for every finite/countable/arbitrary $\mathcal{B}^- \subseteq \mathcal{B}$ and all $A \in \mathcal{A}$: if $\bigcap \mathcal{B}^- \subseteq A$, then $A \in \mathcal{B}$. Similarly, for a proposition C from $\mathcal{A}$, $\mathcal{B} \subseteq \mathcal{A}$ is *conditionally consistent given C in the finite/countable/complete sense* if, and only if, for every finite/countable/arbitrary $\mathcal{B}^- \subseteq \mathcal{B}$: $C \cap \bigcap \mathcal{B}^- \neq \emptyset$. It is *conditionally deductively closed given C in the finite/countable/complete sense* if, and only if, for every finite/countable/arbitrary $\mathcal{B}^- \subseteq \mathcal{B}$ and all $A \in \mathcal{A}$: if $C \cap \bigcap \mathcal{B}^- \subseteq A$, then $A \in \mathcal{B}$.

Now we can respond to Sophia as well as answer the question why an ideal doxastic agent should organize her beliefs and conditional beliefs at a given moment in time according to axioms (1)–(3), and why she should update her beliefs and conditional beliefs across time according to update rules (1)–(3) if she receives new information of the appropriate format. She should do so, because

Theorem 10 *An ideal doxastic agent's belief set $\mathcal{B}_\varrho$ and conditional belief sets $\mathcal{B}_{\varrho(\cdot|C)}$ for consistent conditions C are (conditionally) consistent and deductively closed in the finite / countable / complete sense (given C)—and would remain so*

in response to any finite sequence of experiences—if, and only if, ϱ is a finitely / countably / completely minimitive ranking function that would be revised according to update rules (1)–(3).

This theorem from Huber (2007) rests on several unstated assumptions which are spelled out in Huber (manuscript).

The argument based on this theorem is supposed to establish the thesis that an ideal doxastic agent's beliefs and conditional beliefs should obey the synchronic and diachronic rules of ranking theory. It provides a means-end justification for this thesis in the spirit of epistemic consequentialism (Percival, 2002; Stalnaker, 2002). The idea is that obeying the normative constraints of ranking theory is a (necessary and sufficient) means to attaining the end of being "eternally consistent and deductively closed." The latter end in turn is a (necessary, but insufficient) means to attaining the end of always having only true beliefs, and, subject to the constraint that *all* of them are true, as many thereof as possible. To the extent that the ideal doxastic agent has this end, she should obey the norms of ranking theory. It is not that we are telling Sophia what and how to believe. *She* is the one who is assumed to have these ends. We merely point out the obtaining means-end relationship. Of course, if Sophia does not desire to always hold only true beliefs, and, subject to the constraint that all of them are true, as many thereof as possible, our response will cut no ice. But that is beside the point: it is mistaking a hypothetical imperative for a categorical imperative.

Brössel, Eder, and Huber (2013) discuss the implications of this result as well as its Bayesian role-model, Joyce's (1998, 2009) "non-pragmatic vindication of probabilism" (see also Pettigrew 2011, 2013), for considering doxastic rationality a form of instrumental rationality, and for means-end epistemology in general. Alternatively one may use the representation result by Hild and Spohn (2008), or the rank-theoretic decision theory by Giang and Shenoy (2000), to obtain a justification of ranking theory that is deontological in spirit. For instance, the former result can be used to argue that all and only ranking functions obey the duties, or categorical imperatives, of iterated belief contraction, where these duties, or categorical imperatives, take the form of axioms for iterated contractions of beliefs.

Figure 5 depicts Sophia's ranking functions r and r^* as "numbered onions." Alternatively (Figure 6) Sophia's ranking function r can be pictured as an onion with one, possibly empty, layer $r^{-1}(n)$ for each natural number n. Sophia's old rank for D is 2, i.e. $r(D) = 2$, and her old rank for $\overline{D}$ is 0, i.e. $r(\overline{D}) = 0$. Sophia's new ranking function r^* results from her old ranking function r by first improving the possible worlds in which D is true by 2 ranks so that the new rank of D is 0, i.e. $r^*(D) = 0$. In a second step the possible worlds in which $\overline{D}$ is true are deteriorated by 13 ranks so that the new rank of $\overline{D}$ is 13, i.e. $r^*(\overline{D}) = 13$. The relative positions of the

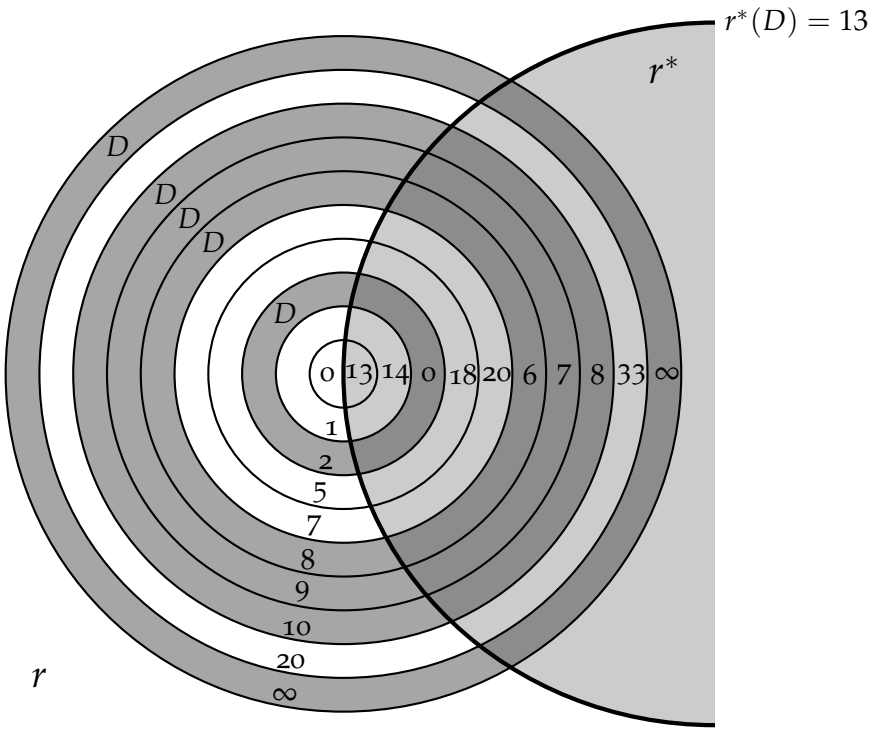


Figure 5: Sophia's ranking functions depicted as "numbered onions"

possible worlds in which D is true, and the possible worlds in which $\overline{D}$ is true, are expressed in the conditional ranking functions $r(\cdot \mid D) = r^*(\cdot \mid D)$ and $r(\cdot \mid \overline{D}) = r^*(\cdot \mid \overline{D})$. These relative positions or conditional ranks are kept fixed.

5 AREAS OF FUTURE RESEARCH

In epistemology ranking theory is a theory of belief and its revision. It studies how an ideal doxastic agent should organize her beliefs and conditional beliefs at a given moment in time, and how she should revise her beliefs and conditional beliefs across time when she receives new information.

In this entry we have distinguished between the following four cases of belief revision. The case where the new information comes in the qualitative form of a sentence or proposition of the agent's language or algebra, as in the AGM theory of belief revision. The case where the new information comes in the comparative form of the relative positions of an input sentence and a reference sentence, as in two-dimensional belief revision. The case where the new information comes in the quantitative form of new grades of disbelief for various propositions, as in the case

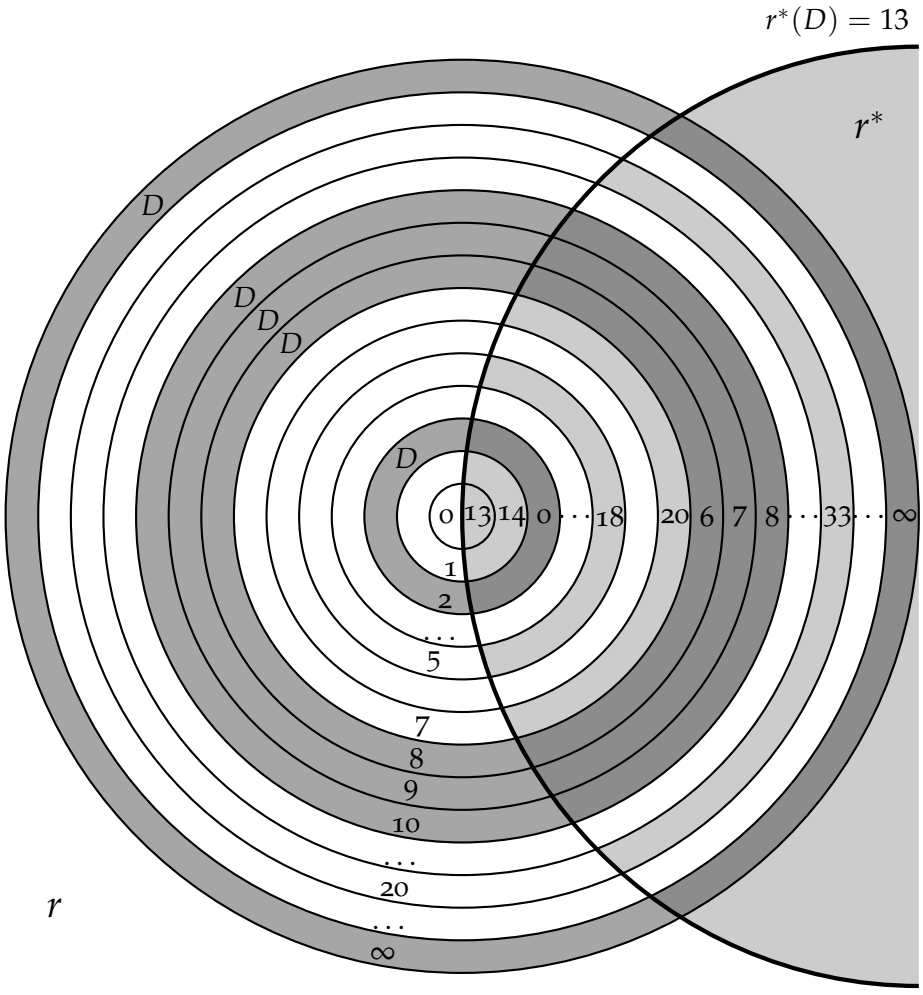


Figure 6: Sophia's ranking functions depicted as "numbered onions" with empty layers

of update rules 1 and 2 of ranking theory. And the case where the new information comes in the quantitative form of differences between the old and new grades of disbelief for such sentences or propositions, as in update rule 3 of ranking theory.

Let us call information that concerns only individual sentences or propositions of the agent's language or algebra *factual information*, and the corresponding changes in belief *factual belief changes*. In this entry we have only discussed factual information and factual belief changes. Besides these there are at least two other forms of information an ideal doxastic agent can receive and, corresponding to these, at least two forms of non-factual belief change. I will briefly mention these and then I will conclude by mentioning applications of ranking theory outside epistemology, in the philosophy of science, in metaphysics, and in the philosophy of language.

The first of these non-factual belief changes takes place when the ideal doxastic agent learns that her language or algebra was too poor or coarse-grained. For instance, Sophia may start out with a language that allows her to distinguish between red wine and white wine, and then may acquire the concept of rosé. Or she may learn that among the red wines one can distinguish between barriques and non-barriques. When the ideal doxastic agent receives such *conceptual information* she should perform a *conceptual belief change*. A prominent conceptual change is that of *logical learning*. In the syntactic AGM framework logical learning is normally studied in terms of *belief bases* (Hansson, 1999). Belief bases differ from belief sets by not being required to be closed under the logical consequence relation. Huber (2015a) shows how logical learning, and conceptual belief changes in general, can be dealt with in the semantic framework of ranking theory.

Another form of non-factual information is *meta-information*, and an ideal doxastic agent receiving meta-information should perform a *meta-belief change* (Stalnaker, 2009). Information about her own doxastic state, as well as about *(in-) dependencies* among propositions, as reported by indicative conditionals, causal claims, and counterfactual conditionals, may be a form of meta-information. In the syntactic AGM framework one might be able to study meta-changes with the help of *dynamic doxastic logic*, *DDL* (Segerberg 1995; Lindström and Rabinowicz 1999; Caie, this volume). DDL allows one to reason about one's own beliefs. In the semantic framework of ranking theory reasoning about one's own beliefs has been studied by Spohn (2012, chaper 9) based on Hild (1998). Huber (2015a) shows how indicative conditionals can be learned in ranking theory.

In the philosophy of language Spohn (2013, 2015) uses ranking theory to develop a unified theory of indicative, counterfactual, and many other conditionals. On this expressivist account most conditionals express conditional beliefs, but counterfactuals express propositions relative to the agent's conditional beliefs and a partition. Huber (2014b, 2017) introduces

so-called alethic ranking functions and defines counterfactuals in terms of them. Raidl (forthcoming) proves completeness results for these and other semantics, and corrects mistakes in Huber (2014b, 2015b, 2017). Alethic ranking functions are related to subjective ranking functions by "the royal rule." This is a normative principle that constrains a priori subjective ranks by alethic ranks much like Lewis (1980)'s principal principle constrains a priori subjective credences by objective chances. Huber (2017) show the royal rule to be the necessary and sufficient means to attaining a cognitive end that relates true beliefs in purely factual, non-modal propositions and true beliefs in purely modal propositions. The philosophical background for this is an idealism about alethic or metaphysical modality that contrasts with the projectivist account of the metaphysical modalities of chance and necessity developed by Spohn (2010a).

In metaphysics Spohn (1983, 2006a) uses ranking theory to develop a theory of causation. This theory works with subjective ranking functions, and so results in a subjective notion of causation, although there are attempts at objectification (Spohn, 1993, 2012, chapter 15). Huber (2011) uses the above-mentioned alethic ranking functions to arrive at a counterfactual notion of causation. The conditional nature of ranking functions and a precisification of Lewis' (1979, p. 472) "system of weights or priorities" allow Huber (2013c) to unify the two modalities of so-called "extended causal models" (Halpern 2008; Halpern and Hitchcock 2010) into the one modality of alethic ranking functions. Spohn (2010b) relates ranking theory and causal models in a very different way.

In the philosophy of science Spohn explicates ceteris paribus conditions (Spohn, 2002, 2014) and laws (Spohn, 2005) in terms of subjective ranking functions. Huber (2015b) shows how the statistical notion of modes can be used to empirically confirm the above-mentioned counterfactuals that are defined in terms of alethic ranking functions.

None of this compares to Spohn (2012), which is the most comprehensive treatment of ranking theory, and an invaluable resource for formal epistemology full of philosophical insights.

ACKNOWLEDGMENTS

I am very much indebted to Wolfgang Spohn. This entry would not have been possible without him. I think this is true—even trivially so, as he has invented ranking theory. He thinks this is not true, at least not in any straightforward sense (Spohn, 2015, section 8), but merely expresses my conditional belief. Therefore let me put it in terms we understand in the same way, except that we do not, because he understands them better than anyone else. He has taught me everything in this entry that remains after

iterated contractions by all falsehoods. I am very grateful to him for doing so.

I am also very grateful to the editors of *The Open Handbook of Formal Epistemology*, Richard Pettigrew and Jonathan Weisberg, for their extensive and helpful feedback, and for putting so much time and energy into editing this wonderful volume.

In preparing this entry I have relied on and, with permission of Wiley, reused the material from Huber (2013a, 2013b). My research was supported by the Canadian SSHRC through its Insight program and by the Connaught Foundation through its New Researcher program.

REFERENCES

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

Bewersdorf, B. (2013). *Belief and its revision* (Doctoral dissertation, Konstanz University).

Boutilier, C. (1995). On the revision of probabilistic belief states. *Notre Dame Journal of Formal Logic*, *36*(1), 158–183.

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

Brössel, P., Eder, A.-M., & Huber, F. (2013). Evidential support and instrumental rationality. *Philosophy and Phenomenological Research*, *87*, 279–300.

Cantwell, J. (1997). On the logic of small changes in hypertheories. *Theoria*, *63*, 54–89.

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

Easwaran, K. (2011a). Bayesianism I: Introduction and arguments in favor. *Philosophy Compass*, *6*, 312–320.

Easwaran, K. (2011b). Bayesianism II: Applications and criticisms. *Philosophy Compass*, *6*, 321–332.

Fermé, E. (2000). Irrevocable belief revision and epistemic entrenchment. *Logic Journal of the IGPL*, *8*, 645–652.

Fermé, E. & Rott, H. (2004). Revision by comparison. *Artificial Intelligence*, *157*, 5–47.

Field, H. (1978). A note on Jeffrey Conditionalization. *Philosophy of Science*, *45*, 361–367.

Foley, R. (2009). Belief, degrees of belief, and the Lockean thesis. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (pp. 37–47). Dordrecht: Springer.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states.* Cambridge, MA: 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). San Francisco: Morgan Kaufmann.

Gärdenfors, P. & Rott, H. (1995). Belief revision. In D. M. Gabbay, C. J. Hogger, & J. A. Robinson (Eds.), *Epistemic and temporal reasoning* (pp. 35–132). Oxford: Handbook of Logic in Artificial Intelligence and Logic Programming: Volume 4. : Clarendon Press.

Giang, P. H. & Shenoy, P. P. (2000). A qualitative linear utility theory for Spohn's theory of epistemic beliefs. In C. Boutilier & M. Goldszmidt (Eds.), *Uncertainty in artificial intelligence 16* (pp. 220–229). San Francisco: Morgan Kaufmann.

Goldszmidt, M. & Pearl, J. J. (1996). Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence*, *84*, 57–112.

Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic*, *17*, 157–170.

Haenni, R. (2009). Non-additive degrees of belief. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief synthese library 342.* (pp. 121–159). Dordrecht: Springer.

Hájek, A. (manuscript). Staying regular? Retrieved from http://hplms.berkeley.edu/HajekStayingRegular.pdf

Halpern, J. Y. (2008). Defaults and normality in causal structures. In *Proceedings of the eleventh international conference on principles of knowledge representation and reasoning (kr 2008* (pp. 198–208).

Halpern, J. Y. & Hitchcock, C. R. (2010). Actual causation and the art of modelling. In R. Dechter, H. Geffner, & J. Halpern (Eds.), *Heuristics, probability, and causality* (pp. 383–406). London: College Publications.

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

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

Harper, W. L. (1976). Rational conceptual change. *PSA*, *1976*(2), 462–494.

Hawthorne, J. (2009). The lockean thesis and the logic of belief. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (pp. 49–74). Dordrecht: Springer.

Herzberger, H. G. (1979). Counterfactuals and consistency. *Journal of Philosophy*, *76*, 83–88.

Hild, M. (1998). Auto-epistemology and updating. *Philosophical Studies*, *92*, 321–361.

Hild, M. & Spohn, W. (2008). The measurement of ranks and the laws of iterated contraction. *Artificial Intelligence*, *172*, 1195–1218.

Huber, F. (manuscript). *Belief and counterfactuals. a study in means-end philosophy*. Under contract with Oxford University Press.

Huber, F. (2006). Ranking functions and rankings on languages. *Artificial Intelligence*, *170*, 462–471.

Huber, F. (2007). The consistency argument for ranking functions. *Studia Logica*, *86*, 299–329.

Huber, F. (2011). Lewis causation is a special case of Spohn causation. *British Journal for the Philosophy of Science*, *62*, 207–210.

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

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

Huber, F. (2013c). Structural equations and beyond. *The Review of Symbolic Logic*, *6*, 709–732.

Huber, F. (2014a). For true conditionalizers Weisberg's Paradox is a false alarm. *Symposion*, *1*, 111–119.

Huber, F. (2014b). New foundations of counterfactuals. *Synthese*, *191*, 2167–2193.

Huber, F. (2015a). How to learn concepts, consequences, and conditionals. *Analytica*, *1*, 20–36.

Huber, F. (2015b). What should I believe about what would have been the case? *Journal of Philosophical Logic*, *44*, 81–110.

Huber, F. (2017). Why follow the royal rule? *Synthese*, *194*(5), 1565–1590.

Jeffrey, R. C. (1983). *The logic of decision.* (2nd). Chicago: University of Chicago Press.

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

Joyce, J. M. (1998). A non-pragmatic vindication of probabilism. *Philosophy of Science*, *65*, 575–603.

Joyce, J. M. (2009). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (pp. 263–297). Synthese Library 342. Dordrecht: Springer.

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

Lewis, D. K. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Lewis, D. K. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*, 455–476.

Lewis, D. K. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability* (pp. 263–293). Vol. II. Berkeley: University of Berkeley Press.

Lindström, S. & Rabinowicz, W. (1999). DDL unlimited: Dynamic doxastic logic for introspective agents. *Erkenntnis*, *50*, 353–385.

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

Percival, P. (2002). Epistemic consequentialism. In *Proceedings of the aristotelian society* (pp. 121–151). Supplementary Volume 76.

Pettigrew, R. (2011). Epistemic utility arguments for probabilism. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*.

Pettigrew, R. (2013). Epistemic utility and norms for credence. *Philosophy Compass*, *8*, 897–908.

Popper, K. R. (1955). Two autonomous axiom systems for the calculus of probabilities. *British Journal for the Philosophy of Science*, *6*, 51–57.

Raidl, E. (forthcoming). Completeness for counter-doxa conditionals—using ranking semantics. *The Review of Symbolic Logic*. doi:10.1017/S1755020318000199

Raidl, E. (2018). Ranking semantics for doxastic necessities and conditionals. In P. Arazim & T. Lávička (Eds.), *Logica yearbook 2017*. College Publications.

Rényi, A. (1955). On a new axiomatic system for probability. *Acta Mathematica Academiae Scientiarum Hungaricae*, *6*, 285–335.

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

Rott, H. (2003). Coherence and conservatism in the dynamics of belief. Part II: Iterated belief change without dispositional coherence. *Journal of Logic and Computation*, *13*, 111–145.

Rott, H. (2006). Revision by comparison as a unifying framework: Severe withdrawal, irrevocable revision and irrefutable revision. *Theoretical Computer Science*, *355*, 228–242.

Rott, H. (2007). Two-dimensional belief change: An advertisement. In G. Bonanno, J. Delgrande, J. Lang, & H. Rott (Eds.), *Formal models of belief change in rational agents*. Dagstuhl Seminar Proceedings 07351.

Rott, H. (2009). Shifting priorities: Simple representations for twenty-seven iterated theory change operators. In D. Makinson, J. Malinowski, & H. Wansing (Eds.), *Towards mathematical philosophy* (pp. 269–296). Trends in Logic 28. Dordrecht: Springer.

Segerberg, K. (1995). Belief revision from the point of view of doxastic logic. *Bulletin of the IGPL*, *3*, 535–553.

Segerberg, K. (1998). Irrevocable belief revision in dynamic doxastic logic. *Notre Dame Journal of Formal Logic*, *39*, 287–306.

Shenoy, P. P. (1991). On Spohn's rule for revision of beliefs. *International Journal of Approximate Reasoning*, *5*, 149–181.

Spohn, W. (1983). *Eine theorie der kausalität*. Unpublished Habilitation thesis. Munich: LMU Munich, 1983.

Spohn, W. (1986). On the representation of Popper Measures. *Topoi*, *5*, 69–74.

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 II* (pp. 105–134). Dordrecht: Kluwer.

Spohn, W. (1990). A general non-probabilistic theory of inductive reasoning. In R. D. Shachter, T. S. Levitt, J. Lemmer, & L. N. Kanal (Eds.), *Uncertainty in artificial intelligence* (pp. 149–158). Amsterdam, North-Holland.

Spohn, W. (1993). Causal laws are objectifications of inductive schemes. In J. Dubucs (Ed.), *Philosophy of probability* (pp. 223–252). Dordrecht: Kluwer.

Spohn, W. (1999). Ranking functions, AGM style. In S. H. B. Hansson, N.-e. Sahlin, & W. Rabinowicz (Eds.), *Internet festschrift for peter gärdenfors.* Lund. Retrieved from www.lucs.lu.se/spinning/categories/dynamics/Spohn/Spohn.pdf

Spohn, W. (2002). Laws, ceteris paribus conditions, and the dynamics of belief. *Erkenntnis*, *57*, 373–394.

Spohn, W. (2005). Enumerative induction and lawlikeness. *Philosophy of Science*, *72*, 164–187.

Spohn, W. (2006a). Causation: An alternative. *British Journal for the Philosophy of Science*, *57*, 93–119.

Spohn, W. (2006b). Isaac Levi's potentially surprising epistemological picture. In E. J. Olsson (Ed.), *Knowledge and inquiry: Essays on the pragmatism of Isaac Levi*. New York: Cambridge University Press.

Spohn, W. (2010a). Chance and necessity: From Humean supervenience to Humean projection. In E. Eells & J. Fetzer (Eds.), *The place of probability in science* (pp. 101–131). Boston Studies in the Philosophy of Science 284. Dordrecht: Springer.

Spohn, W. (2010b). The structural model and the ranking theoretic approach to causation: A comparison. In R. Dechter, H. Geffner, & J. Halpern (Eds.), *Heuristics, probability, and causality* (pp. 507–522). London: College Publications.

Spohn, W. (2012). *The laws of belief. ranking theory and its philosophical applications.* Oxford: Oxford University Press.

Spohn, W. (2013). A ranking-theoretic approach to conditionals. *Cognitive Science*, *37*, 1074–1106.

Spohn, W. (2014). The epistemic account of ceteris paribus conditions. *European Journal for the Philosophy of Science*, *4*, 385–408.

Spohn, W. (2015). Conditionals: A unifying ranking-theoretic perspective. *Philosophers' Imprint*, *15*(1), 1–30.

Stalnaker, R. C. (1970). Probability and conditionality. *Philosophy of Science*, *37*, 64–80.

Stalnaker, R. C. (2002). Epistemic consequentialism. In *Proceedings of the aristotelian society supplementary volume, 76* (pp. 153–168).

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

Vineberg, S. (2000). The logical status of conditionalization and its role in confirmation. In N. Shanks & R. B. Gardner (Eds.), *Logic, probability and science* (pp. 77–94). Poznan Studies in the Philosophy of the Science: Rodopi.

Weisberg, J. (2009). Commutativity or holism? A dilemma for conditionalizers. *British Journal for the Philosophy of Science*, *60*, 793–812.

Weisberg, J. (2011). Varieties of Bayesianism. In D. M. Gabbay, S. Hartmann, & J. Woods (Eds.), *Handbook of the history of logic, volume 10, inductive logic* (pp. 477–551). Oxford: Elsevier.

Weisberg, J. (2015). Updating, undermining, and independence. *British Journal for the Philosophy of Science*, *66*, 121–159.

9

FULL & PARTIAL BELIEF

Konstantin Genin

Philosophers and scientists in allied fields use the term 'belief' to refer roughly to the attitude taken toward a proposition regarded as true. That first approximation is unlikely to satisfy those in search of a non-circular definition. Early twentieth-century psychologists and philosophers of mind attempted to address that difficulty by reducing belief to some sort of behavioral disposition. Although the behaviorist project is usually taken to have been a failure, there is broad consensus that belief leaves a distinctive behavioral footprint: most philosophers would agree that an agent who believes P can be expected to accept P as a premise in reasoning, planning and deliberation. If I believe that the A train is running express, I will plan to take it if going to Harlem, but try to catch the local if going to the Museum of Natural History. Similarly banal examples can be easily multiplied.

It is commonly held that belief is not just an all-or-nothing matter, but admits of degrees. A veteran subway rider may have a higher degree of belief in the proposition that the A train will run local next weekend than in the proposition that the A train will run local next rush hour. Philosophers sufficiently impressed by examples of this sort orient their activity around the structure of "partial belief" rather than the all-or-nothing attitude denoted by "full belief," or belief *simpliciter.* Although it is easy to generate plausible examples of partial beliefs, it is harder to say exactly what is meant by a degree of belief. An agent's degree of belief in P may reflect their level of confidence in the truth of P, their willingness to assent to P in conversation, or perhaps how much evidence is required to convince them to abandon their belief in P. A venerable tradition, receiving classical expression in Ramsey (1931) and de Finetti (1937), holds that degrees of belief are most directly reflected in which bets regarding P an agent is willing to accept. At least since Pascal, mainstream philosophical opinion has held that degrees of belief are well-modeled by probabilities (see Hacking, 1975, for a readable history). To this day, subjective, or "epistemic," probability remains one of the dominant interpretations of the probability calculus.

A parallel tradition, though never as dominant, holds that degrees of belief are neither so precise, nor as definitely comparable as suggested by Pascal's probabilistic analysis. Keynes (1921) famously proposes that degrees of belief may enjoy only an *ordinal* structure, which admits of

qualitative, but not quantitative, comparison. Keynes even suggests that the strength of some pairs of partial beliefs cannot be compared at all.

Cohen (1980) traces another minority tradition to Francis Bacon's *Novum Organum*. On the usual probability scale a degree of belief of zero in some proposition implies maximal conviction in its negation. On the Baconian scale, a degree of belief of zero implies no conviction in either the proposition or its negation. Thus, the usual scale runs from "disproof to proof" whereas the Baconian runs from "no evidence, or non-proof to proof" (Cohen, 1980, p. 224). In the past few decades, Baconian probability has received increasing attention, resulting in theories approaching the maturity and sophistication of those in the Pascalian tradition (Spohn, 2012; Huber, 2019).

Formal epistemologists are traditionally interested in both full and partial belief, although most would probably take partial belief as the primary object of study. Moss (2018) even argues that there are instances of probabilistic *knowledge* that do not involve any full beliefs. On the other hand, traditional analytic epistemology and philosophy of mind routinely studied full belief and related all-or-nothing attitudes such as knowledge and desire, but only rarely show interest in their graded counterparts. The differential emphasis on partial beliefs, although often commented upon, may reflect sociological factors more than any essential difference between the fields. These differences will likely become less pronounced in the future.

What is less often remarked upon is traditional epistemology's focus on individual beliefs, rather than entire systems of belief, as is typical in formal epistemology. Traditional philosophers are interested in what it means for an agent S to believe a particular proposition P. Representationalist philosopher of mind wonder how intentional states, or states that involve "aboutness" arise at all, especially if the agents involved are correctly understood as purely physical systems. Formal epistemologists tend to take matters of mental representation for granted, rarely inquiring into how the trick is worked. Dispositionalist philosophers of mind are interested in analyzing an agent's belief that P into a disposition to reason or act, although they will disagree about how readily these dispositions will be observed in behavior. Their focus on individual beliefs gives rise to certain standard objections. A Muscovite who believes, in the 1930s, that the Stalinist terror is morally wrong, may not betray her beliefs in her behavior at all.

Formal epistemologists resolve such difficulties by insisting on a holism about belief: it is entire *systems* of belief (and perhaps utility) that are reflected in deliberation and action, otherwise underdetermined by individual beliefs. In general, formal epistemologists are interested in the norms governing the structure and dynamics of whole systems of full or

partial belief: how individual beliefs must systematically cohere in order to be rational; how they must be reflected in decision making; and how they ought to accommodate new evidence. Accordingly, those issues will be the focus of this article. For a good introduction to belief in the philosophy of mind, see Schwitzgebel (2015). See Hájek and Lin (2017) for a suggestive discussion of how mainstream and formal epistemology would benefit from increased sensitivity to each other's concerns.

Not everyone agrees that both partial and full beliefs exist—there are theorists who attempt to eliminate one or the other attitude. But anyone who admits the existence of both full and partial belief inherits a thorny problem: how are full beliefs related to partial beliefs? Two answers immediately suggest themselves. The first claims that full belief is just the maximal degree of partial belief. The second argues that full belief is just partial belief above a certain threshold. Both answers give rise to formidable problems. Other theorists claim that an agent's partial beliefs underdetermine their full beliefs in the absence of information about the agent's preferences.

In the last few years, the question of how partial and full belief are related has received considerable attention in formal epistemology, giving rise to several subtle, elegant and, unfortunately, incompatible solutions. The debate between these alternatives is the heart of this article and is presented in Section 5. The preceding sections develop the context and background necessary to understand and appreciate this debate. Readers who feel comfortable with these prerequisites, as well as those who are in a hurry, may skip to the final section and refer back to previous sections only as necessary.

1 THE OBJECTS OF BELIEF

In the following we will see several proposed models for the structure of belief. Most of these proposals take the objects of belief to be either *propositions*, or *sentences* in a formalized language. This section reviews the basic notions required to work with propositions and sentences. If the reader feels overwhelmed with the technicalities in this section, they should feel free to postpone them, and refer back to it on-the-fly. Readers who are accustomed to working with these objects may freely skip this section.

For our purposes, a *possible world* is a way the world, or some interesting aspect of the world, might be. We let W denote the set of *all* possible worlds, i.e. the set of all possible ways the world might be. It is not necessary to think of these as objective, metaphysical realities. More often, possible worlds are constrained by contextual presuppositions, and their granularity reflects our interests. Suffice it to say that knowing the *true*

possible world $w \in W$ would satisfy an agent's curiosity—she would thereby settle some interesting matter under discussion. A proposition $P \subseteq W$ is a *set* of possible worlds, i.e. it is a partial specification of the way the world is. To know that P is true is to know that the true world is among the set of worlds $\{w : w \in P\}$ since P is true in a possible world w iff $w \in P$.

Propositions enjoy a set-theoretic structure. The relative complement of P, $\neg P = W \setminus P$, is the set of all worlds in which P is false. If P, Q are arbitrary propositions, then their intersection $P \cap Q$ is the set of all worlds in which P and Q are both true. The disjunction $P \cup Q$ is the set of worlds in which at least one of P, Q is true. The material conditional $P \to Q$ is the set of worlds $\neg P \cup Q$, in which either P is false or Q is true. If $P \subseteq Q$ we say that P *entails* Q and also that P is *logically stronger* than Q. If $P \subseteq Q$ and $Q \subseteq P$ we write $P \equiv Q$ and say that P and Q are *logically equivalent*. The tautological proposition W is true in all worlds and the contradictory proposition, the empty set $\varnothing$, is not true in any world. A set of propositions $\mathcal{A}$ is *consistent* iff there is a world in which all the elements of $\mathcal{A}$ are true, i.e. if $\cap\mathcal{A} \neq \varnothing$. Otherwise, we say that $\mathcal{A}$ is *inconsistent*. A set of propositions $\mathcal{A}$ is *mutually exclusive* iff the truth of any one element implies the falsehood of all other elements. The set of logical consequences of $\mathcal{A}$, written $Cn(\mathcal{A})$, is the set $\{B \subseteq W : \cap\mathcal{A} \text{ entails } B\}$. Note that if $\mathcal{A}$ is inconsistent, then $Cn(\mathcal{A})$ is $\wp(W)$, the set of all propositions over W.

A set of propositions $\mathcal{F}$ is a *field* (sometimes *algebra*) iff $\mathcal{F}$ contains W and it is closed under intersection, union and complementation. That is to say that if A, B are both elements of $\mathcal{F}$ then $W, A \cup B, A \cap B$, and $\neg A$ are also elements of $\mathcal{F}$. A set of propositions $\mathcal{F}$ is a *σ-field* (sometimes *σ-algebra*) iff it is a field that is closed under *countable* intersections, i.e. if $\mathcal{S} \subseteq \mathcal{F}$ is a countable collection of propositions, then the intersection of all its elements $\cap\mathcal{S}$ is also an element of $\mathcal{F}$. That definition implies that a σ-field is also closed under countable unions. It is not difficult to prove that the intersection of σ-fields is also a σ-field. That implies that every collection of propositions $\mathcal{F}$ generates $\sigma(\mathcal{F})$, the least σ-field containing $\mathcal{F}$, by intersecting the set of all σ-fields containing $\mathcal{F}$.

Propositions, although usually expressed by sentences in a language, are not themselves sentences. That distinction is commonly drawn by saying that propositions are *semantic* objects, whereas sentences are *syntactic* objects. Semantic objects (like propositions) are meaningful, since they represent meaningful possibilities, whereas bits of syntax must be "interpreted" before they become meaningful. In a slogan: sentences are potentially meaningful, whereas propositions already are.

For our purposes, a *language* Λ is identified with the set of all grammatical sentences it contains. Sentences will be denoted by lowercase letters $p, q, \ldots$. The language Λ is assumed to contain a set of *atomic* sentences

$a, b, \ldots$ which are not built out of any other sentences, as well as all the sentences generated by combining the atomic sentences with truth-functional connectives from propositional logic. In other words: if p, q are sentences in Λ then $\neg p$, $p \vee q$, $p \wedge q$, $p \rightarrow q$, and $p \leftrightarrow q$ are also sentences in Λ. These are meant to be read respectively as "not p," "p or q," "p and q," "if p, then q," and "p if and only if q." The symbol $\bot$ (pronounced "falsum") denotes an arbitrarily chosen contradiction (e.g. $p \wedge \neg p$) and the symbol $\top$ denotes an arbitrary tautology. Some of the sentences in Λ follow "logically" from others. For example, under the intended interpretation of the truth-functional connectives, p follows from the sentence $p \wedge q$ and also from the set of sentences $\{q, q \rightarrow p\}$. To capture the essentials of logical consequence, we introduce a *consequence operator*, which maps any set of sentences Γ to its logical consequences $Cn(\Gamma)$. The consequence operator is assumed to satisfy the following properties, which abstract the characteristic features of deductive logic.

$$\Gamma \subseteq Cn(\Gamma). \qquad \text{(Inclusion)}$$

$$\text{If } \Gamma \subseteq \Delta, \text{ then } Cn(\Gamma) \subseteq Cn(\Delta). \qquad \text{(Monotony)}$$

$$Cn(\Gamma) = Cn(Cn(\Gamma)). \qquad \text{(Idempotence)}$$

Inclusion merely expresses the triviality that any sentence p is a deductive consequence of itself. *Monotony* expresses the fact that adding more premises to a deductive argument allows you to derive all the same conclusions as you could with fewer. *Idempotence* says that $Cn(\Delta)$ contains *all* the deductive consequences of Δ. We use $\Gamma \vdash p$ as an alternative notation for $p \in Cn(\Gamma)$ and $\Gamma \nvdash p$ for $p \notin Cn(\Gamma)$. We write $\vdash p$ for $p \in Cn(\varnothing)$. The set of theorems of propositional logic is denoted by $Cn(\varnothing)$ since these can be derived from the axioms alone, without any additional assumptions.

In the following, we will sometimes assume that the consequence operator satisfies the following additional property:

$$q \in Cn(\Delta \cup \{p\}) \text{ implies } (p \rightarrow q) \in Cn(\Delta). \qquad \text{(Deduction theorem)}$$

The deduction theorem expresses the fact that you can prove the conditional sentence $p \rightarrow q$ by assuming p and then deriving q. Unsurprisingly, it is possible to prove that this property holds for most deductive logics one would encounter, including both propositional and first-order logic.

There is, of course, a systematic way to map sentences in a language to propositions. A *valuation function* V maps every atomic sentence a in Λ to a proposition $V(a)$, the set of worlds in which a is true under that interpretation of the atoms. The valuation function also interprets the non-atomic sentences in a way that respects the intended meanings of the logical connectives, i.e. so that $V(\top) = W$, $V(\neg p) = W \setminus V(p)$, and $V(p \wedge q) = V(p) \cap V(q)$. In this fashion, each sentence in Λ is mapped to a

set of possible worlds. Each language Λ and valuation function V generate the field $\mathcal{F}_{\Lambda,V} = \{V(p) : p \in \Lambda\}$. In turn, $\mathcal{F}_{\Lambda,V}$ generates $\sigma(\mathcal{F}_{\Lambda,V})$, the least σ-field containing it.

We write $\Gamma \vDash p$ if for all valuations V,

$$\bigcap_{q \in \Gamma} V(q) \subseteq V(p).$$

Then, $\Gamma \vDash p$ expresses the fact that no matter how the non-logical vocabulary of Λ are interpreted, p is true in all the worlds in which all sentences in Γ are true. We say that p is *valid* iff $\{\top\} \vDash p$, i.e if $W \subseteq V(p)$ for all valuation functions. Then, p is valid iff p is true in all possible worlds, no matter how the non-logical vocabulary are interpreted. For example, the sentence $p \vee \neg p$ is valid.

We assume the following property of our deductive consequence relation.

If $\Gamma \vdash p$, then $\Gamma \vDash p$. (Soundness)

Soundness says that if the sentence p is a derivable consequence of the set of sentences Γ, then no matter how the non-logical vocabulary of Λ are interpreted, p is true in all the worlds in which all the sentences in Γ are true. That is to say that from true premises, our consequence relation always derives true conclusions. Soundness also implies that every theorem is valid. Soundness is a basic requirement of any *deductive* consequence relation, and illustrates the intended connection between deductive proof and semantic entailment.

Sentences are, in a sense, capable of expressing distinctions that propositions cannot. For example, the two sentences p and $\neg\neg p$ are obviously distinct. But if p and q are provably equivalent, i.e. if $\vdash p \leftrightarrow q$, then $\{p\} \vdash q$ and $\{q\} \vdash p$. By Soundness, $\{p\} \vDash q$ and $\{q\} \vDash p$. Therefore, for any valuation function, $V(p) = V(q)$. So p and q must express the same proposition. Of course, an agent who is unaware of the equivalence might believe p without believing q. What's worse, every sentence p such that $\vdash p$ must express the tautological proposition W. Of course, ordinary agents do not always recognize theorems of propositional logic. For this reason, some argue that it is sentences, rather than propositions, that are the appropriate objects of belief. However, most of the proposed models we will study require that rational agents adopt the same belief attitude toward logically equivalent sentences. So long as that is required, there is no significant difference between taking the objects of belief to be sentences or propositions. Still others are not satisfied with either sentences, or propositions. Perry (1979), Lewis (1979) and Stalnaker (1981) argue that in order to capture *essentially indexical* beliefs—beliefs that essentially involve indexicals such as *I*, *here*, or *now*—the objects of belief must be

centered propositions. We will not take up this helpful suggestion here, but see Liao (2012) for a discussion of the costs and benefits of centered propositions.

2 STRUCTURES FOR FULL BELIEF

2.1 *Non-monotonic Logic*

In Section 1 we introduced the notion of a deductive consequence relation. The characteristic feature of a deductive consequence relation is that conclusions are not retracted when premises are added.

$$\text{If } \Gamma \subseteq \Delta, \text{ then } Cn(\Gamma) \subseteq Cn(\Delta). \qquad \text{(Monotony)}$$

Of course, all sorts of seemingly rational everyday reasoning violates Monotony. Reasoning according to *typicality* seems justified in ordinary circumstances, but fails to satisfy Monotony. If you were told that Tweety is a bird, you would be justified in concluding that Tweety flies, since typical birds fly. You would retract your conclusion however, if you were to learn that Tweety is a penguin. That does not mean that your original inference was unreasonable or irrational. *Inductive* inference is also famously non-monotonic. After observing one hundred white swans, you might conclude that all swans are white. Of course, you would retract your conclusion if you ever came across a black swan. *Pace* Pyrrhonian skepticism, there must be at least *some* justified inductive inferences. Ethical reasoning is also shot through with non-monotonicities. Ross (1930) discusses *prima facie* duties, or defeasible obligations, that are binding unless superseded by more urgent, and competing obligations. Ullman-Margalit (1983) points out that legal reasoning routinely relies on presumptions—of innocence, good faith, sanity, etc.—that may be withdrawn in light of new evidence. Non-Monotony is simply unavoidable in ordinary human contexts.

Non-monotonic logic studies a defeasible consequence relation $\mid\!\sim$ between premises, on the left of the wavy turnstile, and conclusions on the right. One may think of the premises on the left as a set Γ of sentences expressing "hard evidence" that an agent may possesses, and the conclusions on the right to be the defeasible conclusions that are justified on the basis of Γ. Thus, the expression $\Gamma \mid\!\sim p$ may be read as "if I were to learn all and only the sentences in Γ, I would be justified in concluding that p."

Recall from Section 1 that a deductive consequence relation satisfies Soundness, i.e. $\Gamma \vdash p$ only if p is true in all the worlds in which all sentences in Γ are true. It is clear from the preceding examples that defeasible reasoning cannot satisfy Soundness. If $\Gamma \mid\!\sim p$ then perhaps p is true in "typical" worlds in which Γ is true, or in "most" worlds in which Γ is true, or perhaps p is a sharply testable possibility compatible with Γ.

We call a consequence relation *ampliative* if $\Gamma \mathrel{|\!\sim} p$, but there are worlds in which all sentences in Γ are true, but p is false. It is possible to construct consequence relations that are non-ampliative and non-monotonic, but ampliativity and non-monotonicity go hand in hand in all paradigmatic cases.

The field of artificial intelligence has, since its inception, been concerned with implementing some form of rational, ampliative, non-monotonic reasoning in artificial agents. For these purposes, deductive consequence relations are unhelpfully restrictive. That does not preclude the possibility that there is some other logic that governs good ampliative reasoning. The past forty years have seen the creation of many logics for non-monotonic inference, often developed to model a specific kind of defeasible reasoning. See Strasser and Antonelli (2018) for an excellent overview.

In view of this profusion of specialized logics, *non-monotonic logic* investigates which properties a logic of defeasible consequence must have in order to count as a *logic* at all.[1] Non-monotonic logic provides a crucial *lingua franca* for comparing different logics of defeasible inference. It is also extremely apt for the purposes of this article, because it allows us to compare different normative theories of how beliefs ought to be updated in light of new evidence, as well as theories of how full and partial beliefs ought to relate to each other.

Before we proceed to the technical development, it will be helpful to introduce an important early critique of nonmonotonic logic due to the philosopher John Pollock. Pollock (1987) identifies two sources of non-monotonicity in defeasible reasoning. An agent may believe p, because she believes q and takes q to be a defeasible *reason* for p. Pollock distinguishes two kinds of *defeaters* for this inference: a *rebutting defeater* is a defeasible reason to believe $\neg p$, whereas an *undercutting defeater* is a reason to believe $\neg q$. Either kind of defeater may induce an agent to retract her belief in p. Pollock's point is that since nonmonotonic logics typically do not represent the structure of an agent's reasons, they often fail to elegantly handle cases of undercutting defeat. We shall soon see several examples.

2.1.1 *Principles for Nonmonotonic Logic*

Let Λ be a formal language, and let $Cn(\cdot)$ be a deductive consequence relation, as discussed in Section 1. There are in fact two closely related approaches to the study of non-montonic consequence relations. The finitary approach studies a relation between individual sentences $p \mathrel{|\!\sim} q$. That approach is taken, for example, in the very influential Kraus, Lehmann, and Magidor (1990). The infinitary approach studies a relation $\Gamma \mathrel{|\!\sim} p$ between an arbitrary set of sentences on the left and individual

1 Gabbay (1985) was the first to suggest this abstract point of view.

sentences on the right. That approach is taken in the canonical reference work Makinson (1994) and cannot in general be simulated by the finitary approach. For the most part we will follow Makinson (1994). However, some results are known to hold only for the finitary settings. Furthermore, the more general infinitary principles are sometimes better appreciated by their finitary consequences. For that reason, we will sometimes switch back and forth between the infinitary and the finitary approach. We write $C(\Gamma)$ for the set $\{p : \Gamma \mathrel{|\!\sim} p\}$. shorthand for $\Gamma \cup \{p\} \mathrel{|\!\sim} q$.

If defeasible logics fail to satisfy Monotony, which principles ought they satisfy? Are there some logical principles which ought to be validated by all rational defeasible reasoning? Almost all consequence relations studied in the literature satisfy the following principle.

$$\Gamma \subseteq C(\Gamma). \qquad \text{(Inclusion)}$$

In its single-premise formulation Inclusion merely says that $p \mathrel{|\!\sim} p$, which is surely unexceptionable. The following two principles are also widely accepted in non-monotonic logic.

$$\Gamma \subseteq \Delta \subseteq C(\Gamma) \text{ implies } C(\Delta) \subseteq C(\Gamma). \qquad \text{(Cut)}$$

$$\Gamma \subseteq \Delta \subseteq C(\Gamma) \text{ implies } C(\Gamma) \subseteq C(\Delta). \qquad \text{(Cautious Monotony)}$$

As special cases, these two principles entail:

$$\Gamma \mathrel{|\!\sim} p \text{ and } \Gamma \cup \{p\} \mathrel{|\!\sim} q \text{ implies } \Gamma \mathrel{|\!\sim} q; \qquad \text{(Cut)}$$

$$\Gamma \mathrel{|\!\sim} p \text{ and } \Gamma \mathrel{|\!\sim} q \text{ implies } \Gamma \cup \{p\} \mathrel{|\!\sim} q. \qquad \text{(Cautious Monotony)}$$

Cut says that adding conclusions inferred from Γ to the set of premises does not *increase* inferential power. Cautious Monotony says that it does not *decrease* inferential power. If we think of the premises on the left of $\mathrel{|\!\sim}$ as my set of "hard" evidence, and the set $C(\Gamma)$ as a theory inductively inferred on the basis of Γ, then Cautious Monotony is an expression of hypothetico-deductivism: if I observe a consequence of my theory $C(\Gamma)$, I should not thereby retract any previous conclusions. Moreover, Cut says that I should not add any new conclusions. Taken together the two principles say that if you observe a consequence of your theory, you should not change it:

$$\Gamma \subseteq \Delta \subseteq C(\Gamma) \text{ implies } C(\Gamma) = C(\Delta). \qquad \text{(Cumulativity)}$$

Gabbay (1985) proposes that (finitary versions of) Inclusion, Cut and Cautious Monotony are the minimal properties that every interesting non-monotonic logic must satisfy. That remains the consensus view to this day. It is easy to show that Inclusion and Cut jointly imply a principle familiar from Section 1:

$$C(\Gamma) = C(C(\Gamma)). \qquad \text{(Idempotence)}$$

There is also the question of how a non-monotonic consequence relation $C(\cdot)$ should interact with a classical relation of deductive consequence $Cn(\cdot)$. The following principle says that defeasible reasoning allows you to make strictly more conclusions than classical deductive reasoning:

$$Cn(\Gamma) \subseteq C(\Gamma). \qquad \text{(Supraclassicality)}$$

That is perhaps unreasonable if we think of $C(\cdot)$ as modeling the defeasible reasoning of some bounded agent. It begins to sound better if we think of $C(\Gamma)$ as modeling the ampliative conclusions that are justified on the basis of Γ.

Makinson (1994) observes that any supraclassical $C(\cdot)$ that satisfies Idempotence and Cumulativity also satisfies the following pair of principles.

$$Cn(C(\Gamma)) = C(\Gamma). \qquad \text{(Left Absorption)}$$

$$C(\Gamma) = C(Cn(\Gamma)). \qquad \text{(Right Absorption)}$$

Left Absorption says that $C(\Gamma)$ is closed under deductive consequence. Right Absorption says that the conclusions that are justified on the basis of Γ depend only on the logical content of Γ, and not on its mode of presentation. The conjunction of Right and Left Absorption is called Full Absorption.

Makinson advocates for one more interaction principle:

$$C(\Gamma) \cap C(\Delta) \subseteq C(Cn(\Gamma) \cap Cn(\Delta)). \qquad \text{(Distribution)}$$

That condition is perhaps too complex to admit of an intuitive gloss. However, we can better understand its meaning from its finitary consequences. Any supraclassical consequence relation satisfying Distribution and Full Absorption also satisfies the following.

$$\Gamma \cup \{p\} \mathrel{|\!\sim} r \text{ and } \Gamma \cup \{q\} \mathrel{|\!\sim} r \text{ implies } \Gamma \cup \{p \vee q\} \mathrel{|\!\sim} r. \qquad \text{(Or)}$$

$$\Gamma \cup \{p\} \mathrel{|\!\sim} q \text{ and } \Gamma \cup \{\neg p\} \mathrel{|\!\sim} q \text{ implies } \Gamma \mathrel{|\!\sim} q. \qquad \text{(Case reasoning)}$$

These two principles seem to be very compelling. Any genuine consequence relation ought to enable reasoning by cases. If I would infer q irrespective of what I learned about p, I should be able to infer q before the matter of p has been decided. Similarly, if p follows defeasibly from both p and q, it ought to follow from their disjunction. Any consequence relation that satisfies Supraclassicality, Left Absorption and Case Reasoning must also satisfy the following principle:

$$\Gamma \cup \{p\} \mathrel{|\!\sim} q \text{ implies } \Gamma \mathrel{|\!\sim} p \to q. \qquad \text{(Conditionalization)}$$

To prove that entailment suppose that $\Gamma \cup \{p\} \mathrel{|\!\sim} q$. Since $p \to q$ is a deductive consequence of q, it follows by Left Absorption that $\Gamma \cup \{p\} \mathrel{|\!\sim} p \to q$. Furthermore, since $p \to q$ is a deductive consequence of $\neg p$ it follows by supraclassicality that $\Gamma \cup \{\neg p\} \mathrel{|\!\sim} p \to q$. By Case Reasoning, $\Gamma \mathrel{|\!\sim} p \to q$.

Conditionalization says that upon learning new evidence, you never "jump to conclusions" that are not entailed by the deductive closure of your old beliefs with the new evidence. That is not an obviously appealing principle. An agent that starts out with $\Gamma = Cn(\varnothing)$ will either fail to validate Conditionalization or never make any ampliative inferences at all. Suppose that after observing 100 black ravens an agent validating Conditionalization comes to believe that all ravens are black. Then, at the outset of inquiry, she must have believed that either all ravens are black, or she will see the first non-black raven among the first hundred. Such an agent seems strangely opinionated about when the first counterexample to the inductive generalization must appear.

For a more realistic example, consider the 1887 Michelson-Morely experiment. After a null result failing to detect any *significant* difference between the speed of light in the prevailing direction of the presumed aether wind, and the speed at right angles to the wind, physicists turned against the aether theory. If the physicists validated Conditionalization then, before the experiments, they must have believed that either there is no luminiferous aether, or the aether wind blows quickly enough to be detected by their equipment. But why should they have been so confident that the aether wind is not too slow to be detectable? Even if there is nothing objectionable about an agent who validates Conditionalization, there is something very anti-inductivist about the thesis that *all* justified defeasible inferences on the basis of new evidence can be reconstructed as deductive inferences from prior conclusions plus the new evidence. Schurz (2011) makes a similar criticism, in a slightly different context:

> Inductive generalizations as well as abductive conjectures accompany belief expansions by new observations, in science as well as in common sense cognitions. After observing several instances of a 'constant conjunction,' humans almost automatically form the corresponding inductive generalization; and after performing a new experimental result sufficiently many times, experimental scientists proclaim the discovery of a new empirical law ... [Conditioning]-type expansion is not at all creative but merely additive: it simply adds the new information and forms the deductive closure, but never generates new (non-logically entailed) hypotheses.

Schurz objects that, according to Conditionalization, dispositions to form inductive generalizations must be "programmed in" with material conditionals at the outset of inquiry. Anyone sympathetic to this view must reject either Supraclassicality, Left Absorption, or Case Reasoning. Finding such surprising consequences of seemingly unproblematic principles is one of the boons of studying non-monotonic logic.

We finish this section by introducing one more prominent and controversial principles of non-monotonic logic. The position one takes on this principle will determine how one feels about many of the theories which we turn to in the following. Kraus et al. (1990) claim that any rational reasoner should validate the following strengthening of Cautious Monotony.

$$\Gamma \mathrel{|\!\sim} p \text{ and } \Gamma \not\mathrel{|\!\sim} \neg q \text{ entails } \Gamma \cup \{q\} \mathrel{|\!\sim} p. \qquad \text{(Rational Monotony)}$$

Rational Monotony says that so long as new evidence q is logically compatible with your prior beliefs $C(\Gamma)$, you should not retract any beliefs from $C(\Gamma)$. Accepting both Rational Monotony and Conditionalization amounts to saying that when confronted with new evidence that is logically consistent with her beliefs, a rational agent responds by simply forming the deductive closure of her existing beliefs with the new evidence. On that view, deductive logic is the only necessary guide to reasoning, so long as you do not run into contradiction. Stalnaker (1994) gives the following well-known purported counterexample to Rational Monotony.

Suppose an agent initially believes the following about the three composers Verdi, Bizet, and Satie.

(Iv) Verdi is Italian;

(Fb) Bizet is French;

(Fs) Satie is French.

Let p be the sentence that Verdi and Bizet are compatriots, i.e. $(Fv \wedge Fb) \vee (Iv \wedge Ib)$. Let q be the sentence that Bizet and Satie are compatriots. Suppose that the agent receives the evidence p. As a result, she retracts her belief in $Iv \wedge Fb$ concluding that either Verdi and Bizet are both French or they are both Italian. She retains her belief that Satie is French. Notice that after updating on p, she believes it is possible that Bizet and Satie are compatriots, i.e. $p \not\mathrel{|\!\sim} \neg q$. Now suppose that she receives the evidence q. Since q is compatible with all her previous conclusions, Rational Monotony requires her to conclude that all three composers are French. However, it seems perfectly rational to suspend judgment and concludes that the three are either all Italian, or all French.

Kelly and Lin (forthcoming) give the following counterexample to Rational Monotony, based on Lehrer's (1965) no-false-lemma variant of Gettier's

famous (1963) scenario. There are just two people in your office, named Alice and Bob. You are interested in whether one of them owns a certain Ford. Let p be the sentence that Alice owns the Ford. Let q be the sentence that Bob has the Ford. You have inconclusive evidence that Alice owns the Ford—you saw her driving one just like it. You have weaker evidence that Bob owns the Ford—his brother owns a Ford dealership. Based on that evidence Γ you conclude $p \vee q$, i.e. that someone in the office owns the Ford, but do not go so far as inferring p, or q. You ask Alice and she tells you that the Ford she was driving was rented. That defeats your main *reason* for $p \vee q$, therefore you retract your belief that someone in the office has a Ford. But since $\Gamma \not\vdash \neg p$, Rational Monotony requires you to conclude that Bob owns the Ford. However, there does not seem to be anything irrational about how you have reasoned. This seems to be an illustration of Pollock's (1987) point: the logic is going wrong because it is ignoring the structure of the agent's reasons.

It is also possible to understand Rational Monotony as a thesis about what counts as a legitimate *input* to a belief revision, rather than as a restriction on how one can rationally assimilate such an input. Let d be a new datum that I have collected. Let q be a sentence expressing the total evidential import of d. Suppose that upon assimilating the total evidential import of the data d, I would give up belief in p. Then one might argue that the total evidential import of d is logically inconsistent with p. Moreover, if my initial belief set is consistent, I must have disbelieved q *ex ante*. Thus Rational Monotony can be understood as saying that the appropriate input to a belief revision is not merely a set of data, but rather their total evidential import. From this perspective it can be seen as an expression of Carnap's (1947) principle of total evidence.

We end this section on a terminological note. It is common in the literature to use *System P* (Preferential) to refer to the following set of single-premise principles, labeled so that the reader can identify their infinitary analogues. The terminology is due to Kraus et al. (1990).

$p \mathrel{|\!\sim} p$. (Reflexivity)

$\vdash p \leftrightarrow q$ and $p \mathrel{|\!\sim} r$ implies $q \mathrel{|\!\sim} r$. (Left equivalence)

$\vdash q \rightarrow r$ and $p \mathrel{|\!\sim} q$ implies $p \mathrel{|\!\sim} r$. (Right weakening)

$p \mathrel{|\!\sim} q$ and $p \mathrel{|\!\sim} r$ implies $p \mathrel{|\!\sim} q \wedge r$. (And)

$p \mathrel{|\!\sim} r$ and $q \mathrel{|\!\sim} r$ implies $p \vee q \mathrel{|\!\sim} r$. (Or)

$p \mathrel{|\!\sim} q$ and $p \mathrel{|\!\sim} r$ implies $p \wedge q \mathrel{|\!\sim} r$. (Cautious monotony)

System R (Rational) arises from System P by adding a single-premise version of Rational Monotony:

$p \mathrel{|\!\sim} r$ and $p \not\mathrel{|\!\sim} \neg q$ implies $p \wedge q \mathrel{|\!\sim} r$. (Rational monotony)

2.1.2 *Preferential Semantics*

So far we have considered a non-monotonic consequence relation merely as a relation between syntactic objects. We can rephrase properties of non-monotonic logic "semantically," i.e. in terms of the possible worlds in which the sentences are true or false. In some cases, this allows us to give a very perspicuous view on defeasible logic.

Recall from Section 1 that a deductive consequence relation satisfies Soundness, i.e. that $\Gamma \vdash p$ only if p is true in all the worlds in which all sentences in Γ are true. As we have discussed, non-monotonic logics are ampliative, and therefore must violate Soundness. Shoham (1987) inaugurated a semantics for non-monotonic logics in which $\Gamma \mathrel{|\!\sim} p$ only if p is true in a "preferred" set of worlds in which Γ is true. On a typical interpretation, these are the most typical, or most normal worlds in which all sentences in Γ are true. If Γ is a set of sentences in Λ and V is a valuation function, we write $V(\Gamma)$ as shorthand for $\cap_{q\in\Gamma} V(q)$. See Section 1 if you need a refresher on valuation functions. Kraus et al. (1990) first proved most of the results of this section for single-premise consequence relations. We follow Makinson (1994) in presenting their infinitary generalizations.

A *preferential model* is a triple $\langle W, V, <\rangle$ where W is a set of possible worlds, V is a valuation function and $<$ is an arbitrary relation on the elements of W. The relation $<$ is *transitive* iff $x < y$ and $y < z$ implies $x < z$. The relation $<$ is *irreflexive* iff for all $w \in W$ it is not the case that $w < w$. A transitive, irreflexive relation is called a *strict order*. We write $w \leq v$ iff $w < v$ or $w = v$. The strict order $<$ is *total* iff for $w, v \in W$ either $w \leq v$ or $v \leq w$.

If Γ is a set of sentences, we say that $w \in \mathrm{Min}_<(\Gamma)$ iff $w \in V(\Gamma)$ and there is no $v \in V(\Gamma)$ such that $v < w$. In other words, $w \in \mathrm{Min}_<(\Gamma)$ iff w is a $<$-minimal element of $V(\Gamma)$. Every preferential model gives rise to a consequence relation by letting

$$\Gamma \mathrel{|\!\sim_<} p \text{ iff } \mathrm{Min}_<(\Gamma) \subseteq V(p),$$

i.e. $\Gamma \mathrel{|\!\sim_<} p$ iff p is true in all the *minimal* worlds in which all sentences in Γ are true. Write $C_<(\Gamma)$ for the set $\{p : \Gamma \mathrel{|\!\sim_<} p\}$.

We say that a preferential model is *stoppered* iff for every set of sentences Γ, if $w \in V(\Gamma)$ then there is $v \leq w$ such that $v \in \mathrm{Min}_<(\Gamma)$. (Note that Kraus et al., 1990, called stoppered models *smooth* models.) Makinson (1994) proves the following.

Theorem 11 *Suppose that $\mathfrak{M} = \langle W, V, <\rangle$ is a preferential model. Then $C_<(\cdot)$ satisfies Inclusion, Cut, Supraclassicality, and Distribution. If $\mathfrak{M}$ is stoppered, then $C_<(\cdot)$ also satisfies Cautious Monotony.*

Makinson (1994) also gives the following two partial converses. The latter essentially reports a result from Kraus et al. (1990).

Theorem 12 *If $C(\cdot)$ satisfies Inclusion, Cut, and Cautious Monotony, there is a stoppered preferential model $\mathfrak{M} = \langle W, V, < \rangle$ such that $C(\cdot) = C_<(\cdot)$.*

Theorem 13 *If $C(\cdot)$ satisfies Inclusion, Cut, Cautious Monotony, Supraclassicality, and Distribution, then there is a stoppered preferential model $\mathfrak{M} = \langle W, V, < \rangle$ such that for all finite $\Delta \subseteq \Lambda$, $C(\Delta) = C_<(\Delta)$. Moreover, $\mathfrak{M}$ may be constructed such that $<$ is a strict order.*

Taken together, Theorem 11 and Theorem 13 say that, at least for finitary consequences, the consequence relations generated by preferential models are exactly the consequence relations satisfying Inclusion, Cut, Cautious Monotonicity, Supraclassicality, and Distribution. In fact, one can always think of these preferential models as generated by a strict (partial) order. The question remains whether there are any natural conditions on preferential models that ensure that Rational Monotony is also satisfied. It turns out that Rational Monotony follows from the requirement that the preference relation $<$ is a total order.

Say that a preferential model $\mathfrak{M} = \langle W, V, < \rangle$ is *modular* iff for all $w, u, v \in W$, if $w < u \approx v$ then $w < v$. Here $u \approx v$ means that u, v are unordered, i.e. it is not the case that $u < v$ and it is not the case that $v < u$. If $<$ is a strict order, modularity is equivalent to the intuitive property of *rankedness*: there is a totally ordered set T and a function $\rho : W \to T$ such that for all $u, v \in W$, $u < v$ iff $\rho(u) \ll \rho(v)$, where $\ll$ is the total ordering of T. Makinson proves the following.

Theorem 14 *Suppose that $\mathfrak{M} = \langle W, V, < \rangle$ is a preferential model. If $\mathfrak{M}$ is modular, then $C_<(\cdot)$ satisfies Rational Monotony.*

Kraus et al. (1990) prove the following partial converse.

Theorem 15 *If $C(\cdot)$* finitarily *satisfies Inclusion, Cut, Cautious Monotony, Supraclassicality, Distribution, and Rational Monotony, then there is a ranked, stoppered preferential model $\mathfrak{M} = \langle W, V, < \rangle$ such that for all finite $\Delta \subseteq \Lambda$, $C(\Delta) = C_<(\Delta)$. Moreover, $\mathfrak{M}$ may be constructed such that $<$ is a strict order.*

The essential difference between preferential models that satisfy Rational Monotony and those that do not is that the former correspond to those generated by a *ranked* partial order. This result is helpful to keep in mind because in the following we will see several models of belief that can be understood as arising from a *total* plausibility order, and some that arise from a merely *partial* plausibility order. In light of Theorem 13 and Theorem 15, we can expect the former to satisfy System R and the latter to satisfy only the weaker System P.

2.2 *AGM Belief Revision Theory*

The theory of belief revision is concerned with how to update one's beliefs in light of new evidence, especially when new evidence is inconsistent with prior beliefs. It is especially occupied with the following sort of scenario, borrowed from Gärdenfors (1992). Suppose that you believe all the following sentences:

(a) All European swans are white;

(b) The bird in the pond is a swan;

(c) The bird in the pond comes from Sweden;

(d) Sweden is in Europe.

Now suppose that you were to learn the sentence e that the bird in the pond is black. Clearly, e is inconsistent with your beliefs a, b, c, d. If you want to incorporate the new information e and remain consistent, you will have to retract some of your original beliefs. The problem of belief revision is that deductive logic alone cannot tell you which of your beliefs to give up—this has to be decided by some other means. Considering a similar problem, Quine and Ullian (1970) enunciated the principle of "conservatism," counseling that our new beliefs "may have to conflict with some of our previous beliefs; but the fewer the better." In his (1990), Quine dubs this the "maxim of minimal mutilation." Inspired by these suggestive principles, Alchourrón, Gärdenfors, and Makinson (1985) develop a highly influential theory of belief revision, known thereafter as AGM theory, after its three originators.

In AGM theory, beliefs held by an agent are represented by a set B of sentences. The set B is called the *belief state* of the agent. This set is usually assumed to be closed under logical consequence. Of course, this is an unrealistic idealization, since it means that the agent believes all logical consequences of her beliefs. Levi (1991) defends this idealization by changing the interpretation of the set B—these are the sentences that the agent is *committed* to believe, not those that she actually believes. Although we may never live up to our commitments, Levi argues that we are committed to the logical consequences of our beliefs. That may rescue the principle, but only by changing the interpretation of the theory.

AGM theory studies three different types of belief change. *Contraction* occurs when the belief state B is replaced by $B \div p$, a logically closed subset of B no longer containing p. *Expansion* occurs when the belief state B is replaced with $B + p = Cn(B \cup \{p\})$, the result of simply adding p to the set of beliefs and closing under logical consequence. *Revision* occurs when the belief state B is replaced by $B * p$, the result of adding p to B and

removing whatever is necessary to ensure that the resulting belief state $B * p$ is logically consistent.

Contraction is the fundamental form of belief change studied by AGM. There is no mystery in how to define expansion, and revision is usually defined derivatively via the *Levi identity* (1977): $B * p = (B \div \neg p) + p$. Alchourrón et al. (1985) and Gärdenfors and Makinson (1988) proceed axiomatically: they postulate several principles that every rational contraction operation must satisfy. Fundamental to AGM theory are several representation theorems showing that certain intuitive constructions give rise to contraction operations satisfying the basic postulates and conversely, that every operation satisfying the basic postulates can be seen as the outcome of such a construction. See Lin (2019) for an introduction to these results.

AGM theory is unique in focusing on belief contraction. For someone concerned with maintaining a database, contraction is a fairly natural operation. Medical researchers might want to publish a data set, but make sure that it cannot be used to identify their patients. Privacy regulations may force data collectors to "forget" certain facts about you and, naturally, they would want to do this as conservatively as possible. However, a plausible argument holds that all forms of rational belief change occurring "in the wild" involve learning new information, rather than conservatively removing an old belief. All the other formalisms covered in the article focus on this form of belief change. For this reason, we focus on the AGM theory of revision and neglect contraction.

Before delving into some of the technical development, we mention some important objections and alternatives to the AGM framework. As we have mentioned, the belief state of an agent is represented by the (deductively closed) set B of sentences the agent is committed to believe. The structure of the agent's *reasons* is not represented: you cannot tell of any two $p, q \in B$ whether one is a reason for the other. Gärdenfors (1992) distinguishes between *foundations* theories, that keep track of which beliefs justify which others, and *coherence* theories, which ignore the structure of justification and focus instead on whether beliefs are consistent with one another. Arguing for the coherence approach, Gärdenfors (1992) draws a stark distinction between the two:

> According to the foundations theory, belief revision should consist, first, in giving up all beliefs that no longer have a *satisfactory justification* and, second, in adding new beliefs that have become justified. On the other hand, according to the coherence theory, the objectives are, first, to maintain *consistency* in the revised epistemic state, and, second, to make *minimal changes* of the old state that guarantee overall coherence.

Implicit in this passage is the idea that foundations theory are fundamentally out of sympathy with the principle of minimal mutilation. Elsewhere (1988), Gärdenfors is more apologetic, suggesting that some hybrid theory is possible and perhaps even preferable:

> I admit that the postulates for contractions and revisions that have been introduced here are quite simpleminded, but they seem to capture what can be formulated for the meager structure of belief sets. In richer models of epistemic states, admitting, for example, reasons to be formulated, the corresponding conservativity postulates must be formulated much more cautiously (p. 67).

Previously, we have seen Pollock (1987) advocating for foundationalism. In artificial intelligence, Doyle's Doyle's (1979) *reason maintenance system* is taken to exemplify the foundations approach. Horty (2012) argues that default logic aptly represents the structure of reasons. For a defense of foundationalism, as well as a useful comparison of the two approaches, see Doyle (1992).

Another dissenting tradition advocates for *belief bases* instead of belief states. A belief base is a set of sentences that is typically not closed under logical consequence. Its elements represent "basic" beliefs that are not derived from other beliefs. This allows us to distinguish between sentences that are explicit beliefs, like "Shakespeare wrote Hamlet" and never thought-of consequences like "Either Shakespeare wrote Hamlet or Alan Turing was born on a Monday." Revision and contraction are then redefined to operate on belief bases, rather than belief sets. That allows for increased expressive power, since belief bases which have the same logical closure are not treated interchangeably. For an introduction to belief bases see Hansson (2017). For a book-length treatment, see Hansson (1999).

Finally, one of the most common criticisms of AGM theory is that it does not illuminate *iterated* belief change. In the following, we shall see that the canonical revision operation takes as input an entrenchment ordering on a belief state, but outputs a belief state without an entrenchment order. That severely underdetermines the result of a subsequent revision. For more on the problem of iterated belief revision, see Huber (2013a).

The treatment in this article is necessarily rather compressed. There are several excellent survey articles on belief revision. See Hansson (2017), Huber (2013a, 2013b), and Lin (2019).

2.2.1 *Revision*

Alchourrón et al. (1985) propose the following postulates for rational belief revision.

$B * p = Cn(B * p)$. (Closure)

$p \in B * p$. (Success)

$B * p \subseteq Cn(B \cup \{p\})$. (Inclusion)

If $\neg p \notin Cn(B)$, then $B \subseteq B * p$. (Preservation)

$B * p$ is consistent if p is consistent. (Consistency)

If $(p \leftrightarrow q) \in Cn(\varnothing)$, then $B * p = B * q$. (Extensionality)

By now, Closure, Success, Consistency, and Extensionality should be straightforward to interpret. These postulates impose synchronic constraints on $B * p$. Preservation and Inclusion are the only norms that are really about *revision*—they capture the diachronic spirit of AGM revision. Inclusion says that revision by p should yield *no more* new beliefs than expansion by p. In other words, any sentence q that you come to believe after revising by p is a deductive consequence of p and your prior beliefs. Consider the following principle:

If $q \in B * p$, then $(p \rightarrow q) \in B$. (Conditionalization)

In Section 2.1.1, we considered an analogue of Conditionalization for nonmonotonic logic. All the same objections apply equally well in the context of belief revision. Recall from Section 1 that a deductive consequence relation admits a deduction theorem iff $\Delta \cup \{p\} \vdash q$ implies that $\Delta \vdash p \rightarrow q$. So long as a deduction theorem is provable for $Cn(\cdot)$, Inclusion and Conditionalization are equivalent. To see this, suppose that the revision operation $*$ satisfies Inclusion. Then, if $q \in B * p$, it follows that $B \cup \{p\} \vdash q$. By the deduction theorem, $B \vdash p \rightarrow q$. For the converse, suppose that the revision operation $*$ satisfies Conditionalization. Then, if $q \in B * p$, it follows that $p \rightarrow q \in B$ and $q \in Cn(B \cup \{p\})$. If you found any of the arguments against Conditionalization convincing, you ought to be skeptical of Inclusion.

Preservation says that, so long as the new information p is logically consistent with your prior beliefs, all of your prior beliefs survive revision by p. In the setting of non-monotonic logic, we called this principle Rational Monotony. All objections and counterexamples to Rational Monotony from Section 2.1.1 apply equally well in belief revision. As we have seen, Preservation rules out any kind of *undercutting* defeat of previously successful defeasible inferences. Accepting both Preservation (Rational Monotonicity) and Inclusion (Conditionalization) amounts to saying that when confronted with new evidence that is logically consistent with her beliefs, a rational agent responds by simply forming the deductive closure of her existing beliefs with the new evidence. On that view, deductive logic is

the only necessary guide to reasoning, so long as you do not run into contradiction.

Alchourrón et al. (1985) also propose the following supplementary revision postulates, closely related to Inclusion and Preservation.

$$B * (p \wedge q) \subseteq (B * p) + q. \qquad \text{(Conjunctive Inclusion)}$$

$$\text{If } \neg q \notin Cn(B * p), \text{ then } (B * p) + q \subseteq B * (p \wedge q). \qquad \text{(Conjunctive Preservation)}$$

It is possible to make the connection between belief revision and nonmonotonic logic precise. Given a belief set B and a revision operation $*$, we can define a single-premise consequence relation by setting

$$p \mathrel{|\!\sim} q \text{ iff } q \in B * p.$$

Similarly, given a single-premise consequence relation $\mathrel{|\!\sim}$ we can define

$$B = \{p : \top \mathrel{|\!\sim} p\} \text{ and } B * p = \{q : p \mathrel{|\!\sim} q\}.$$

Then it is possible to prove the following correspondences between AGM belief revision and the set of single-premise principles we called System R in Section 2.1.1. It follows, by Theorem 15, that AGM revision can be represented in terms of a ranked, stoppered preferential model over possible worlds.

Theorem 16 *Suppose that $*$ is a revision operation for B satisfying all eight revision postulates. Then, the nonmonotonic consequence relation given by $p \mathrel{|\!\sim} q$ iff $q \in B * p$ satisfies all the principles of System R.*

Theorem 17 *Suppose that $\mathrel{|\!\sim}$ is a consequence relation that satisfies all the principles of System R and such that $p \mathrel{|\!\sim} \bot$ only if $\vdash \neg p$. Then, the revision operation $*$ defined by letting $B = \{p : \top \mathrel{|\!\sim} p\}$ and $B * p = \{q : p \mathrel{|\!\sim} q\}$ satisfies all eight revision postulates.*

2.2.2 *Entrenchment*

Gärdenfors and Makinson (1988) introduce the notion of an *entrenchment relation* on sentences.

> Even if all sentences in a . . . set are accepted or considered as facts . . . , this does not mean that all sentences are of equal value for planning or problem-solving purposes. Certain . . . beliefs about the world are more important than others when planning future actions, conducting scientific investigations, or reasoning in general. We will say that some sentences . . . have a higher degree of *epistemic entrenchment* than others. The degree

> of entrenchment will, intuitively, have a bearing on what is abandoned ..., and what is retained, when a contraction or revision is carried out.

To model the degree of entrenchment, Gärdenfors and Makinson (1988) introduce a relation $\leq$ holding between sentences of the language Λ. The notation $p \leq q$ is pronounced "p is at most as entrenched as q." Gärdenfors and Makinson (1988) propose that the entrenchment relation $\leq$ satisfy the following postulates.

If $p \leq q$ and $q \leq r$, then $p \leq r$. (Transitivity)

If $p \vdash q$, then $p \leq q$. (Dominance)

Either $p \leq (p \wedge q)$, or $q \leq (p \wedge q)$. (Conjunctiveness)

If B is consistent, then $p \notin B$ iff $p \leq q$ for all q. (Minimality)

If $q \leq p$ for all q, then $p \in Cn(\varnothing)$. (Maximality)

Note that, in light of Minimality, an entrenchment relation is defined for a particular belief set B. It follows from the first three of these postulates that an entrenchment order is *total*, i.e. for all p, q either $p \leq q$ or $q \leq p$.

Given a belief set B and an entrenchment relation $\leq$, it is possible to define a revision operation directly by setting:

$$B * p = Cn(\{q \in \Lambda : \neg p < q\} \cup \{p\}). \qquad (C*)$$

The idea behind this equation is that the agent revises by p by first clearing from her belief set anything less entrenched than $\neg p$, (by dominance, this includes everything entailing $\neg p$) adding p, and then closing under logical consequence. This illustrates why AGM theory is not a theory of *iterated* revision: the revision operation takes as input an entrenchment order and belief state, but outputs only a belief state. That severely underdetermines the results of subsequent revisions. Gärdenfors (1988) proves the following.

Theorem 18 *If a relation $\leq$ satisfies the five entrenchment postulates, then the revision function $*$ determined via $(C*)$ satisfies the six basic and the two supplementary revision postulates.*

Finally, given a belief set B, an entrenchment relation can be recovered from a revision operation by setting:

$$p \leq q \text{ iff } p \notin B * \neg(p \wedge q) \text{ or } \vdash q. \qquad (C^*_{\leq})$$

The idea is that p is no more entrenched than q if p does not survive a revision by $\neg(p \wedge q)$ or if q is a tautology. Rott (2003) proves the following.

Theorem 19 *If a revision operation $*$ satisfies the six basic and the two supplementary contraction postulates, then the entrenchment relation determined via $(C^*_{\leq})$ satisfies the five entrenchment postulates.*

2.2.3 *Sphere Semantics*

So far we have thought of belief revision syntactically: a revision operation $*$ takes in a set B of syntactic objects and a sentence p and outputs another set of sentences $B * p$. Grove (1988) gives a perspicuous way to represent the revision postulates semantically, i.e. in terms of the possible worlds in which the sentences are true or false.

As before, let W be a set of possible worlds and let $V : \Lambda \to \wp(W)$ be a valuation function.[2] If Γ is a set of sentences in Λ, we write $V(\Gamma)$ as shorthand for $\cap_{q \in \Gamma} V(q)$. If E is a proposition, we write $T(E)$ as shorthand for $\{p \in \Lambda : E \subseteq V(p)\}$, i.e. the set of all sentences p such that E entails $V(p)$.

A set of propositions $\mathcal{S}$ is a *system of spheres* centered on $V(B) \subseteq W$ iff for all $E, F \subseteq W$ and all $p \in \Lambda$, the following conditions hold.

If $E, F \in \mathcal{S}$, then $E \subseteq F$ or $F \subseteq E$. (Nested)

$V(B) \in \mathcal{S}$ and if $E \in \mathcal{S}$ then $V(B) \subseteq E$. (Centered)

$W \in \mathcal{S}$. (Maximum)

If $p \nvdash \bot$, then there is $E \in \mathcal{S}$ such that
$E \cap V(p) \neq \varnothing$ and if $F \cap V(p) \neq \varnothing$ then $E \subseteq F$. (Well order)

In other words, a system of spheres centered on $V(B)$ is a nested set of propositions, all entailed by $V(B)$, with the following property: if p is a consistent sentence, then there is a logically strongest element of $\mathcal{S}$ consistent with $V(p)$. If p is a consistent sentence, let $\mathcal{S}(p)$ be $E \cap V(p)$, where E is the logically strongest element of $\mathcal{S}$ consistent with $V(p)$. Otherwise, let $V(p) = \varnothing$. In other words: $\mathcal{S}(p)$ is the set of worlds compatible with $V(p)$ that is "closest" to $V(B)$ according to the sphere system. Note that if $V(p) \cap V(B) \neq \varnothing$, then $\mathcal{S}(p) = V(B) \cap V(p)$. If $V(p) \cap V(B) = \varnothing$ we find the closest sphere compatible with $V(p)$ and intersect the two. Given a belief set B and a system of spheres $\mathcal{S}$ centered on $V(B)$ we can define a revision operator by setting:

$$B * p = T(\mathcal{S}(p)).$$

The idea is this: when you revise on a sentence p compatible with your previous beliefs, then the strongest proposition you believe is $V(p) \cap V(B)$. If p is incompatible with your beliefs, you fall back to $E \cap V(p)$, where E is the p-compatible proposition closest to your old belief $V(B)$. Thus, the system of spheres $\mathcal{S}$ can be seen as a set of "fallback positions" for updating on incompatible propositions. See Figure 1.

Grove (1988) proves the following.

2 See Section 1 if you need a refresher on valuation functions.

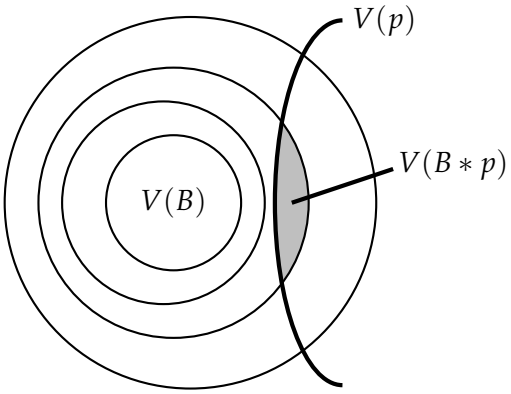


Figure 1: A system of spheres centered on $V(B)$. The shaded region is $V(B * p)$.

Theorem 20 *Let B be a belief set. For each system of spheres $\mathcal{S}$ centered on $V(B)$, there is an operation $*$ satisfying the six basic and the two supplementary revision postulates such that $B * p = T(\mathcal{S}(p))$. Moreover, for every revision operation $*$ satisfying the six basic and the two supplementary revision postulates, there is a sphere system $\mathcal{S}$ centered on $V(B)$ such that $B * p = T(\mathcal{S}(p))$.*

Finally, the sphere semantics gives a dramatic illustration of the critique that AGM does not illuminate iterated belief change. A revision maps a set of spheres $\mathcal{S}$ and a sentence p to a *set of sentences* $T(\mathcal{S}(p))$. It does not output a new set of spheres centered on the new belief state. That severely underdetermines the result of future revisions.

2.3 *The Paradox of the Preface*

The models of full belief that we have seen so far require that beliefs be consistent and closed under deductive consequence. While it is admitted that this requirement is not psychologically realistic, or perhaps even feasible for bounded agents, it is proffered as a normative principle that we should strive to approximate. After all, consistency and closure are *necessary* conditions for achieving the following two related ends: believing only true sentences (consistency) and believing as many true sentences as possible without risking error in any more possible worlds (closure).

Nevertheless, the Paradox of the Preface, due to Makinson (1965), challenges even the *normativity* of deductive consistency. The story goes like this. A famous theorist has just finished her latest book. As is customary for such works, she includes a passage in the preface thanking her colleagues and students for their help in editing and proofreading, but

accepting sole responsibility for the mistakes that inevitably remain. She seems to be saying that, despite her best efforts, she believes that not everything that she asserts in the book is true. Let $s_1, \ldots, s_n$ be the claims she asserts in the book. Presumably, she believes each of the s_i or else she would not have asserted them. Yet in the preface she claims to believe $\neg(s_1 \wedge \ldots \wedge s_n)$, the claim that at least one of the s_i is false. The theorist seems to be behaving perfectly rationally, yet on this reconstruction there is no way that she can be both consistent and deductively closed.

It is tempting to say that inconsistency in the service of intellectual humility is no sin. Yet this creates a further difficulty: surely some inconsistencies are vicious and should be eliminated. Others, it seems, are virtuous and ought to be tolerated. But how do we know which inconsistencies are which? If we point out an inconsistency in someone's beliefs, we tend to think that they are under some pressure to resolve it. But why can't everyone respond to such a challenge by claiming that their inconsistency is a virtue? The Preface Paradox seems to challenge the very normativity of logical consistency.

There are several ways to respond to this challenge. The first, and perhaps the most common route, is to claim that belief is fundamentally a matter of degree. The theorist merely has a high *degree* of belief in each of the statements of her book. And there is nothing surprising about having a high degree of belief in each of the s_i but not in their conjunction. In fact, if the structure of partial belief is probabilistic, this would emerge as a simple consequence of the probability calculus: it is to be expected that the probability of each of the s_i exceeds the probability of their conjunction, so long as the probability of the s_i falls short of unity. This analysis also entails something about the relationship between full and partial belief: it is rationally admissible to fully believe statements that have a high, but not maximal, degree of belief. These themes will be taken up in subsequent sections.

A second set of responses to the paradox calls our attention to the variety of cognitive attitudes that are involved in the story. For example, Cohen (1992) attributes many confusions and apparent paradoxes to the erroneous conflation of two related cognitive attitudes: belief and *acceptance*. Belief in p, according to Cohen, is a disposition to feel it true that p, whenever attending to issues raised by the proposition that p. This disposition may or may not be reflected in speech and action, and is not under direct volitional control. But to *accept* that p is to adopt a policy of deeming, positing, or postulating that p—i.e. of including it among one's premises for deciding what to do or think in a particular context, whether or not one feels it to be true. Acceptance is a volitional matter, and is sensitive to our cognitive context.

Belief is sometimes not even a *prima facie* reason for acceptance: in scientific contexts many of our most cherished beliefs are not accepted, as it is our duty to subject them to criticism and not argue from them as premises. Cohen claims that a person who accepts nothing that she believes is intellectually paralyzed, but someone who accepts everything she believes is recklessly uncritical. Furthermore, acceptance may sometimes promote belief, at least in the long run, but often has no effect: for example, a defense lawyer may accept that her client is innocent, but believe otherwise.

Cohen claims that acceptance ought to be closed, at least under accepted deductive consequences. Consistency is also, presumably, a norm of acceptance. Belief, however, is different:

> . . . you are not intellectually pledged by a set of *beliefs*, however strong, to each deductive consequence of that set of beliefs, even if you recognize it to be such. That is because belief that p is a disposition to feel that p, and feelings that arise in you . . . through involuntary processes . . . no more impose their logical consequences on you than do the electoral campaign posters that people stick on your walls without your consent. (Cohen, 1992, p. 31)

Armed with the distinction between belief and acceptance, we can attempt a redescription of the preface paradox. In the context of her theoretical work, the theorist accepts $s_1, \ldots, s_n$ and is bound to maintain consistency and accept their (accepted) deductive consequences. In fact, she would be in dereliction of her duty as theorist if she accepted the preface sentence in the body of the book. However, the context of the preface is different: here it is customary to drop the professional exigencies of the theorist and acknowledge broader features of the author's cognitive life. She has fulfilled her duty as theorist and done the utmost to accept only those claims that are justified by her evidence and arguments. However, some of these conclusions may not yet be attended with the inner glow of belief. Perhaps, if the work meets with no devastating objections, she may eventually cease to believe the humble claim in the preface. Thus the distinction between belief and acceptance explains why we are not alarmed by the sentence in the context of the preface, but we would be shocked if we saw it used as a premise in the body of the text. For a similar resolution of the paradox, see Chapter 5 of Stalnaker (1984).

It is easy to underestimate the consequences of accepting Cohen's arguments. For one, we would have to reinterpret all of the theories of rational belief that we have discussed as theories of rational acceptance. In fact, there may be no theory of rational belief, but only psychological tricks and heuristics for coming to believe, similar to those Pascal recommends for arriving at faith in Christ. Longstanding dogmas about the relation

between belief and knowledge would have to be revisited. Moreover, excessive appeal to the distinction threatens the unity and cohesiveness of our cognitive lives. For a discussion of these kinds of objections see Kvanvig (2016). For an overview of the distinction between belief and acceptance, see Weirich (2004).

3 STRUCTURES FOR PARTIAL BELIEF

3.1 *Bayesianism*

Bayesianism, or subjective probability theory, is by far the dominant paradigm for modeling partial belief. The literature on the subject is by now very large and includes many approachable introductions. The summary provided here will, of necessity, be rather brief. For an article-length introduction see Huber (2016), Easwaran (2011a, 2011b), or Weisberg (2011). For a book-length introduction see Earman (1992) or Howson and Urbach (2006). For an article-length introduction to Bayesian models of rational action, see Briggs (2017). For an approachable book-length introduction to the theory of rational choice see Resnik (1987).

The heart of the Bayesian theory is roughly the following:

1. There is a fundamental psychological attitude called *degree of belief* (sometimes called *confidence* or *credence*) that can be represented by numbers in the $[0, 1]$ interval.

2. The degrees of belief of rational agents satisfy the axioms of probability theory.

3. The degrees of belief of rational agents are *updated* by some flavor of probabilistic conditioning.

The first two principles are the synchronic requirements of Bayesian theory; the third principle concerns diachronic updating behavior. Most Bayesians would also agree to some version of the following principles, which link subjective probabilities with deliberation and action:

4. Possible states of the world (sometimes *outcomes*) are assigned a *utility*: a positive or negative real number that reflects the desirability or undesirability of that outcome.

5. Rational agents perform only those actions that maximize *expected* utility, which is calculated by weighing the utility of outcomes by their subjective probability.

What makes Bayesianism so formidable is that, in addition to providing an account of rational belief and its updating, it also provides an account

of rational action and deliberation. No other theory can claim a developed, fine-grained account of all three of these aspects of belief. In the following we will briefly spell out some of the technical details of the Bayesian picture.

3.1.1 *Probabilism*

In this section we flesh out the details of the synchronic component of the Bayesian theory. For the purposes of this section we will take propositions to be the objects of (partial) belief. It is also possible to take a syntactic approach and assign degrees of belief to sentences in a formal language. For the most part, nothing hinges on which approach we choose. For arguments in favor of the syntactic approach, see Weisberg (2011).

As usual, let W be a set of possible worlds. Let $\mathcal{F}$ be a σ-field over W.[3] A *credence* function p assigns a degree of belief to every proposition in $\mathcal{F}$. *Probabilism* requires that the credence function satisfies the axioms of probability. For every $E, F \in \mathcal{F}$:

$$p(E) \text{ is a positive, real number;} \qquad \text{(Positivity)}$$

$$p(W) = 1; \qquad \text{(Unitarity)}$$

$$\text{if } E \cap F = \varnothing, \text{ then } p(E \cup F) = p(E) + p(F). \qquad \text{(Additivity)}$$

From these principles it is possible to derive several illuminating theorems. For example, the degree of belief assigned to the contradictory proposition is equal to zero. Furthermore, if E entails F, then $p(E) \leq p(F)$. Finally, for any proposition $E \in \mathcal{F}$ we have that $0 \leq p(E) \leq 1$.

In the standard axiomatization of probability theory due to Kolmogorov (1950), additivity is strengthened to Countable Additivity.

$$\text{If } E_1, E_2, \ldots \text{ are mutually exclusive,}$$
$$\text{then } p(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} p(E_i). \qquad \text{(Countable Additivity)}$$

This requirement is not as innocent as it looks: it rules out the possibility that any agent is indifferent over a countably infinite set of mutually exclusive possibilities. De Finetti (1970, 1972) famously argued that we ought to reject countable additivity since it is conceivable that God could pick out a natural number "at random" and with equal (zero) probability. For another example, suppose you assign 50% credence to the proposition $\neg B$ that not all ravens that will ever be observed are black. Let $\neg B_i$ be the proposition that the i^{th} observed raven is the first non-black raven to appear. Then $\neg B = \bigcup_{i=1}^{\infty} \neg B_i$. Countable additivity entails that for all $\epsilon > 0$ there is a finite n such that $p(\bigcup_{i=1}^{n} \neg B_i) = 1/2 - \epsilon$. So you must be nearly certain that if all ravens are not black, the first non-black raven will

3 Refer to Section 1 if you need to refresh yourself on the definition of a σ-field.

appear among the first n ravens. The only way to assign equal probability to all $\neg B_i$ is to violate countable additivity by setting $p(\neg B_i) = 0$ for all i. This solution has its own drawbacks. On all standard models of Bayesian update it will be impossible to become convinced that the i^{th} raven is indeed non-black, even if you are looking at a white one. For more on countable additivity, see Chapter 13 in Kelly (1996).

Now that we have defined probabilism, it is natural to ask how to justify it: why *should* a rational agent's degrees of belief obey the probability axioms? There are roughly two kinds of answers to this question current in the Bayesian canon.

The traditional answer is that an agent that violates the axioms of probability opens herself up to systems of bets that, although fair from the agent's perspective, guarantee a sure loss. Answers of this flavor are called *Dutch book* arguments and they require positing some connection between degrees of belief and fair betting quotients. Some epistemologists find Dutch book arguments to be unconvincing either because they disavow any tight connection between degrees of belief and betting quotients, or they deny that any facts about something so pragmatic as betting could have normative epistemic force. These epistemologists tend to prefer *accuracy* arguments, which purport to show that any agent violating the probability axioms will have beliefs which are less accurate, or "further from the truth," than agents that satisfy the axioms. We will briefly review the traditional Dutch book-style arguments. For the original articulation of the accuracy perspective see Joyce (1998). For an article-length overview of accuracy-style arguments see Pettigrew (2016b). For a book-length treatment see Pettigrew (2016a).

Dutch book arguments require specifying some connection between degrees of belief and fair betting quotients. For de Finetti (1937) the connection was definitional: an agent's degree of belief in a proposition A simply is her fair odds ratio for a bet that pays \$1 if A is true and nothing otherwise. If you are willing to pay at most \$.50 for a bet that pays \$1 if A is true and nothing otherwise, then it must be that your degree of confidence in A is 50%. It is easy to see what is wrong with this kind of definition: there may be factors other than the subject's degree of belief which affect her fair betting quotient. She may be risk averse, or risk-loving; she may abhor gambling, or love showing off. Ramsey (1931) avoids some of these problems by pricing bets in utility, rather than money, and appealing to the existence of an "ethically neutral" proposition that is considered equally likely to be true and false. For more on the connection between degrees of belief and betting ratios see Eriksson and Hájek (2007).

Supposing that a suitable connection between degrees of belief and fair betting quotients exists, it is possible to construct a "Dutch book" against an agent violating the axioms of probability. To get such an argument

going we suppose that if the agent's degree of belief in A is $p(A)$, then she considers fair a bet that costs $\$p(A) \cdot Y$ and pays $\$Y$ if A is true and \$0 otherwise. Note that we allow Y to take positive and negative values. This means that the agent is willing to assume the role of the bookie and sell a bet that "costs" $-\$p(A) \cdot Y$ and "pays" $-\$Y$ if A is true and \$0 otherwise. Now suppose that such an agent violates finite additivity. One way this may happen is if for A, B such that $A \cap B = \varnothing$, we have that $p(A \cup B) > p(A) + p(B)$. Then, the agent considers fair

1. a bet that costs $-\$p(A)$ and pays $-\$1$ if A is true and \$0 otherwise;
2. a bet that costs $-\$p(B)$ and pays $-\$1$ if B is true and \$0 otherwise;
3. a bet that costs $\$p(A \cup B)$ and pays \$1 if $A \cup B$ is true and \$0 otherwise.

There are three possible scenarios: either A and B are both false or exactly one of them is true. The reader should confirm that in any of these scenarios the agent is left with exactly $\$p(A) + \$p(B) - \$p(A \cup B) < 0$. By reversing which bets the agent buys and sells, we can construct a Dutch book against an agent that violates additivity by having $p(A \cup B) < p(A) + p(B)$. Similar strategies work to construct Dutch books against agents that violate Positivity, Unitarity, and Countable Additivity. Furthermore, it is possible to show that if your degrees of belief validate the probability axioms, then no Dutch book can be made against you (Kemeny, 1955). For more on Dutch book arguments see Section 3.3 in Hájek (2012).

3.1.2 *Updating by Conditioning*

We have discussed the synchronic content of the Bayesian theory, but we still need to talk about how degrees of belief are updated upon receiving new information. There are two standard models of partial belief update: strict conditionalization and Jeffrey conditionalization. Strict conditionalization assumes that the information received acquires the maximal degree of belief. Jeffrey conditionalization allows for the situation in which no proposition is upgraded to full certainty when new information is acquired.

For all propositions $A, B \in \mathcal{F}$ such that $p(A) > 0$, the conditional probability of B given A is defined as:

$$p(B \mid A) := \frac{p(A \cap B)}{p(A)}.$$

Conditionalization by A restricts all possibilities to those compatible with A and then renormalizes by the probability of A to ensure that unitarity holds. By far the most standard modeling of partial belief update holds

that degrees of belief ought to be updated by conditionalization. In other words, if p_t is your credence function at time t and A is a proposition expressing the total new information acquired by $t' > t$, then $p_{t'}$ ought to equal $p_t(\cdot \mid A)$, whenever $p_t(A) > 0$.

What if new information does not render any proposition certain, but merely changes the subjective probability of some propositions? Jeffrey (1983) proposes the following update rule. Suppose that p_t is your credence function at time t. Suppose that the total evidence that comes in by time t' updates your degrees of belief in partition $\{A_i\}_{1 \leq i \leq n}$ (and in no finer partition) setting each respectively to a_i with $\sum_i a_i = 1$. Then your new credence function $p_{t'}$ ought to be $\sum_i p(\cdot \mid A_i) a_i$.

Why should a rational agent update by strict or Jeffrey conditionalization? Dutch-book style arguments for strict conditionalization are given in Teller (1973) and Lewis (1999) and extended to Jeffrey conditionalization in Armendt (1980). For more see Skyrms (2009). For an accuracy-style argument in favor of strict conditionalization and against Jeffrey conditionalization, see Leitgeb and Pettigrew (2010).

For our purposes it is important to point out that conditional probability is always a lower bound for the probability of the material conditional. In other words, for all $E, H \in \mathcal{F}$,

$$p(H \mid E) \leq p(E \to H),$$

whenever $p(E) > 0$. We can see this as a quantitative version of the qualitative principle of Conditionalization we discussed in Section 2.1.1: however confident a Bayesian agent becomes in H after updating on E, she must have been at least as confident that H is a material consequence of E. Popper and Miller (1983) took this observation to be "completely devastating to the inductive interpretation of the calculus of probability." For the history of the Popper-Miller debate see Chapter 4 in Earman (1992). A similar property can be demonstrated for Jeffrey conditioning (Genin, 2017).

Both strict and Jeffrey conditionalization are defined in terms of conditional probability. The probability of B conditional on A is standardly defined as the ratio of the unconditional probabilities $p(A \cap B)$ and $p(A)$. Clearly, this ratio is undefined when $p(A) = 0$. Some theorists would like conditional probability to be defined even when conditioning on propositions of probability zero. The standard approach in mathematical statistics, due to Kolmogorov (1950), is via the *conditional expectation*. On that approach, conditional probability remains dependent on unconditional probability. An alternative approach, adopted by Popper (1955) and Renyi (1955), takes conditional probability as a primitive, rather than a derivative, notion. For a defense of the conditional expectation, see Gyenis, Hofer-Szabó, and Rédei (2017). For an introduction to primitive conditional

probabilities, see Easwaran (2015b). For a critique of the standard notion of conditional probability, see Hájek (2003).

3.1.3 *Deliberation and Action*

One of the signal advantages of the Bayesian model of partial belief is that it is ready-made to plug into a prominent model of practical deliberation. Decision theory, or rational choice theory, is too large and sprawling a subject to be effectively covered here, although it will be presented in cursory outline. For an excellent introduction, see Briggs (2017) or Thoma (2019). For our purposes, it is enough to note that a well-developed theory exists and that no comparable theory exists for alternative models of belief.[4]

Suppose you would like to make a six egg omelet. You've broken 5 fresh eggs into a mixing bowl. Rooting around your fridge, you find a loose egg of uncertain provenance. If you are feeling lucky you can break the suspect egg directly into the mixing bowl; if you are wary of the egg, you might break it into a saucer first and incur more dishwashing.

There are four essential ingredients to this sort of decision-theoretic situation. There are *outcomes*, over which we have defined *utilities* measuring the desirability of the outcome. In the case of the omelet the outcomes are a ruined omelet or a 5–6 egg omelet, with or without extra washing. There are *states*—usually unknown to and out of the control of the actor—which influence the outcome of the decision. In our case the states are exhausted by the possible states of the suspect egg: either good or rotten. Finally, there are *acts* which are under the control of the decision maker. In our case the acts include breaking the egg into the bowl or the saucer. Of course there are other conceivable acts: you might throw the suspect egg away and make do with a 5-egg omelet; you might even flip a coin to decide what to do. We omit these for the sake of simplicity. These four elements are usually summarized in a payoff table (see Table 1). To fit this into the framework of partial belief we assume that the set of acts $A_1, A_2, \ldots, A_n$ partition W. We also assume the set of states $S_1, S_2, \ldots, S_m$ partition W. We assume that the credence function assigns a probability to every outcome. We assume that acts and states are logically independent, so that no state rules out the performance of any act. Finally, we assume that given a state of the world S_j and an act A_i there is exactly one outcome O_{ij}, which is assigned a utility $U(O_{ij})$. The ultimate counsel of rational choice theory is

4 However, recent work such as Lin (2013) and Spohn (2017, 2019) may remedy that inadequacy in the case of qualitative belief.

	States	
	Good	Rotten
Acts	**Outcomes**	
Break into Bowl	6-Egg Omelet +10	No Omelet 0
Break into Saucer	6-Egg Omelet, extra washing +8	5-Egg Omelet, extra washing +4

Table 1: A payoff table for the morning chef

that agents ought to perform only those acts that maximize *expected* utility. The expected utility of an act is defined as:

$$EU(A_i) = \sum_{j=1}^{m} p_{A_i}(S_j)U(O_{ij}),$$

where $p_{A_i}(S_j)$ is roughly how likely the agent considers S_j given that she has performed act A_i. Difficulties about how this quantity should be defined give rise to the schism between evidential and causal decision theory (see Section 3.3 in Thoma, 2019). However, in many situations, including the dilemma of the omelet, the act chosen does not affect the probabilities with which states obtain. This is called "act-state independence" in the jargon of rational choice theory. In cases of act-state independence there is broad consensus that $p_{A_i}(S_j)$ should be equal to the unconditional degree of belief $p(S_j)$.

Central to the literature on decision theory are a number of *representation theorems* showing that every agent with qualitative preferences satisfying a set of rationality postulates can be represented as an expected utility maximizer (von Neumann & Morgenstern, 1944; Savage, 1954). These axioms are controversial, and are subject to intuitive counterexamples. Allais (1953) and Ellsberg (1961) give examples in which seemingly rational agents violate the rationality postulates and therefore cannot, even in principle, be represented as expected utility maximizers. For more on this subject, see Sections 2 and 3 in Briggs (2017).

3.1.4 *Modifications and Alternatives*

Dissatisfaction with various aspects of the Bayesian theory has spawned a number of formal projects. Many epistemologists reject the notion that rational agents must have *precise* credences in every proposition that they can entertain; instead they claim that rational agents may have *imprecise*

credences representable by *intervals* of real numbers. For an introduction to imprecise probability, see Mahtani (2019). The theory of Dempster-Shafer belief functions (Dempster, 1968; Shafer, 1976) rejects the tight connection between degrees of belief and fair betting ratios. Fair betting ratios ought indeed satisfy the axioms of probability, but degrees of belief need not. Nevertheless, it should be possible to calculate fair betting ratios from degrees of belief when these are necessary. For this purpose, degrees of belief may satisfy a weaker set of axioms than those of the probability calculus. For an introduction to Dempster-Shafer belief functions see Section 3.1 in Huber (2016).

Many epistemologists have held that degrees of belief are not so definitely comparable as suggested by the probabilistic analysis. Keynes (1921) famously proposes that degrees of belief may enjoy only an *ordinal* structure, which admits of qualitative, but not quantitative, comparison. Keynes even suggests that the strength of some pairs of partial beliefs cannot be compared at all. Koopman (1940) and Fine (1973) pursue Keynes' suggestions, developing axiomatic theories of qualitative probability. See Konek (2019) for an introduction to qualitative probability comparisons.

3.2 *Ranking Theory*

Cohen (1977, 1980) distinguishes between two rival probabilistic traditions. Pascalian probability finds its latest expression in contemporary Bayesianism. But Cohen traces a rival tradition back to Francis Bacon. Roughly, these two can be distinguished by the *scale* they select for the strength of belief. On the Pascalian scale a degree of belief of zero in some proposition implies maximal conviction in its negation. On the Baconian scale, a degree of belief of zero implies no conviction in either the proposition or its negation. Thus, the Pascalian scale runs from "disproof to proof" whereas the Baconian runs from "no evidence, or non-proof to proof" (Cohen, 1980, p. 224). Cohen (1977) argues that despite the conspicuous successes of Pascalian probability, the Baconian scale is more appropriate in other settings, including legal proceedings.

Ranking theory, first developed in Spohn (1988), is a sophisticated contemporary theory of Baconian probability. For an article-length introduction to ranking theory see Huber (2013b, 2019). For an extensive book-length treatment, with applications to many subjects in epistemology and philosophy of science, see Spohn (2012). We mention some of its basic features, as it provides a useful counterpoint to the models of belief we have already discussed.

As before, let W be a set of possible worlds. Let $\mathcal{F}$ be an algebra over W.[5] A function $\beta : \mathcal{F} \to \mathbb{N} \cup \{\infty\}$ from $\mathcal{F}$ into the set of natural numbers $\mathbb{N}$ extended by ∞, is a *positive ranking function* on $\mathcal{F}$ just in case for any $A, B \in \mathcal{F}$:

$$\beta(\varnothing) = 0; \tag{Consistency}$$

$$\beta(W) = \infty; \tag{Infinitivity}$$

$$\beta(A \cap B) = \min\{\beta(A), \beta(B)\}. \tag{Minimitivity}$$

A positive ranking function expresses degrees of belief. If $\beta(A) > 0$, then we may say that A is (fully) believed and $\neg A$ is disbelieved. If $\beta(A) = 0$ then A is not believed and $\neg A$ may not be believed either. Thus, ranking theory can be seen as satisfying the "Lockean thesis," the intuitive proposal that a degree of belief above some threshold is necessary and sufficient for full belief (see Section 5.2). Note however that nothing in ranking theory requires us to say that the threshold is exactly zero: we could have chosen any positive number n.

Let β be a positive ranking function and $A \in \mathcal{F}$ with $\beta(\neg A) < \infty$. Then for any $B \in \mathcal{F}$ the *conditional positive rank of B given A* is defined as

$$\beta(B \mid A) = \beta(\neg A \cup B) - \beta(\neg A).$$

The function $\beta_A : B \mapsto \beta(B \mid A)$ is called the *conditionalization of β by A* and is itself a positive ranking function. This definition is used to articulate an update rule for ranking theory: if β is your positive ranking function at time t and between t and t' you become certain of $E \in \mathcal{F}$ and no logically stronger proposition, then β_E should be your new ranking function at time t'. Spohn (1988) also defines ranking-theoretic analogues of Jeffrey conditioning.

It is clear from the definition of conditioning that, as in the Bayesian case, the rank of the material conditional is a lower bound for the conditional rank: $\beta(A \to B) \leq \beta(B \mid A)$. It also satisfies a version of Rational Monotony: if $\beta(\neg A) = 0$ and $\beta(B) > 0$, then $\beta(B \mid A) > 0$.[6] Therefore, ranking-theoretic update satisfies the "spirit" of AGM update. Note however, that ranking theory has no trouble with iterated belief revision: a revision takes as input a ranking function and an evidential proposition and outputs a new ranking function.

Ranking theory lies somewhat awkwardly between a theory of full and partial belief. On the one hand, all propositions of positive rank are fully believed. On the other hand, the rank of a proposition measures something

5 Refer to Section 1 if you need to refresh yourself on the definition of an algebra

6 Rational Monotony is not satisfied if we set the threshold for full belief at some number greater than zero.

about the strength of that belief. But how should we interpret these ranks? Huber (2019) investigates the relation between ranking-theoretic degrees of belief, and AGM-style degrees of entrenchment. The *degree of entrenchment* for a proposition A is defined as the number of independent and reliable information sources testifying *against* A that it requires for the agent to give up full belief in A. Degrees of entrenchment may be used to *measure* ranking-theoretic degrees of belief; alternatively, it is possible to *identify* ranking-theoretic degrees of belief with degrees of entrenchment. Huber (manuscript) proves that if an agent defines her full beliefs from an entrenchment function, her beliefs will be consistent and deductively closed iff the entrenchment function is a ranking function.

One of the advantages of ranking theory over AGM is that it allows *reasons* to be defined (Spohn, 2012). Say that A *is a reason for* B with respect to the positive ranking function β iff $\beta(B \mid A) > \beta(B \mid \neg A)$. Say that an agent *has* A *as a reason for* B iff A is a reason for B according to her positive ranking function β and $\beta(A) > 0$. Note that it is not possible to make such a definition in the AGM theory since the conditional degree of entrenchment is not defined. Thus ranking theory may provide an answer to Pollock's criticism of belief revision by allowing various kinds of defeat of reasons to be represented (Spohn, 2012, Section 11.5).

4 ELIMINATIONISMS

There are those who deny that there are any interesting principles bridging full and partial belief. Theorists of this persuasion often want either to eliminate one of these attitudes or reduce it to a special case of the other. Jeffrey (1970) suggests that talk of full belief is vestigial and will be entirely superseded by talk of partial belief and utility:

> . . . nor am I disturbed by the fact that our ordinary notion of *belief* is only vestigially present in the notion of degree of belief. I am inclined to think Ramsey sucked the marrow out of the ordinary notion, and used it to nourish a more adequate view. But maybe there is more there, of value. I hope so. Show me; I have not seen it at all clearly, but it may be there for all that (p. 172).

Theorists such as Kaplan (1996) also suggests that talk of full belief is superfluous once the mechanisms of Bayesian decision theory are in place. After all, only partial beliefs (or *confidence* in Kaplan's terminology) and utilities play any role in the Bayesian framework of rational deliberation, whereas full belief need not be mentioned at all. Those committed to full beliefs have the burden of showing what difference they make to our lives:

> Making the case that talk of investing confidence leaves out something important—something we have in mind when we talk of belief—is going to require honest toil. One has to say . . . exactly how an account of rational human activity will be the poorer if it has no recourse to talk of belief. In short, one has to meet *the Bayesian Challenge*. (p. 100)

Stalnaker (1984) is much more sympathetic to a qualitative notion of belief (or acceptance) but acknowledges the force of the Bayesian Challenge.

> Bayesian decision theory gives a complete account of how probability values . . . ought to guide behavior . . . So what could be the point of selecting an interval near the top of the probability scale and conferring on the propositions whose probability falls in that interval the honorific title "accepted"? Unless acceptance . . . makes a difference to how the agent behaves, or ought to behave it is difficult to see how the concept of acceptance can have the interest and importance for inquiry that it seems to have. (p. 91)

It is true that there is no canonical qualitative analogue to the Bayesian theory of practical deliberation. However, the fact that it is the theorist of full belief that feels the challenge, and not *vice versa*, may be an accident of history: if a qualitative theory of practical deliberation had been developed first, the shoe would now be on the other foot. The situation would be even more severe if qualitative decision making, which we seem to implement as a matter of course, were less cognitively demanding than its Bayesian counterpart. Of course, this anticipates a robust theory of rational qualitative deliberation that is not immediately forthcoming. However, recent work such as Lin (2013) and Spohn (2017, 2019) may remedy that inadequacy. For example, Lin (2013) proves a Savage-style representation theorem characterizing the relationship between full beliefs, desires over possible outcomes, and preferences over acts. By developing a theory of rational action in terms of qualitative belief, Lin demonstrates how one might answer the Bayesian challenge.

On the other hand there are partisans of full belief that are deeply skeptical about partial beliefs.[7] Many of these object that partial beliefs have no psychological reality and would be too difficult to reason with if they did. Horgan (2017) goes so far as to say that typically "there is no such psychological state as the agent's credence in p" and that Bayesian epistemology is "like alchemy and phlogiston theory: it is not about any real phenomena, and thus it also is not about any genuine norms that

7 See Harman (1986), Pollock (2006), Moon (2017), and Horgan (2017). See also the "bad cop" in Hájek and Lin (2017).

govern real phenomena" (p. 7). Harman (1986) argues that we have very few explicit partial beliefs. A theory of reasoning, according to Harman, can concern only explicit attitudes, since these are the only ones that can figure in a reasoning process. Therefore, Bayesian epistemology, while perhaps an account of dispositions to act, is not a guide to reasoning. Nevertheless, partial beliefs may be implicit in our system of full beliefs in that they can be reconstructed from our dispositions to revise them.

> How should we account for the varying strengths of explicit beliefs? I am inclined to suppose that these varying strengths are implicit in a system of beliefs one accepts in a yes/no fashion. My guess is that they are to be explained as a kind of epiphenomenon resulting from the operation of rules of revision. For example, it may be that P is believed more strongly than Q if it would be harder to stop believing P than to stop believing Q, perhaps because it would require more of a revision of one's view... (Harman, 1986, p. 22)

On this picture, almost all of our explicit beliefs are qualitative. Partial beliefs are not graded *belief attitudes* toward propositions, but rather dispositions to revise our *full* beliefs. The correct theory of partial belief, according to Harman, has more to do with entrenchment orders (see Section 2.2.2) or ranking-theoretic degrees of belief (see Section 3.2) than with probabilities. Other apparently partial belief attitudes are explained as full beliefs *about* objective probabilities. So, in the case of a fair lottery with ten thousand tickets, the agent does not believe to a *high degree* that the n^{th} ticket will not win, but rather fully believes that it is objectively improbable that it will win.

Frankish (2009) objects that Harman's view requires that an agent have a full belief in any proposition that we have a degree of belief in: "And this is surely wrong. I have some degree of confidence (less than 50%) in the proposition that it will rain tomorrow, but I do not believe flat-out that it will rain—not, at least, by the everyday standards for flat-out belief" (p. 4). Harman might reply that Frankish merely has a full belief in the objective probability of rain tomorrow. Frankish claims that this escape route is closed to Harman because single events "do not have objective probabilities," but this matter is hardly settled.

Staffel (2013) gives an example in which a proposition with a higher degree of belief is apparently less entrenched than one with a lower degree of belief. Suppose that you will draw a sequence of two million marbles from a big jar full of red and black marbles. You do not know what proportion of the marbles are red. Consider the following cases.

Scenario 1. You have drawn twenty marbles, 19 black and one red. Your degree of belief that the last marble you will draw is black is .95.

SCENARIO 2. You have drawn a million marbles, 900,000 of which have been black. Your degree of belief that the last marble you will draw is black is $19/20 = .90$.

Staffel argues that your degree of belief in the first case is higher than in the second, but much more entrenched in the second than in the first. Therefore, degree of belief cannot be reduced to degree of entrenchment. Nevertheless, the same gambit is open to Harman in the case of the marbles—he can claim that in both scenarios you merely have a full belief in a proposition about objective chance. See Staffel (2013) for a much more extensive engagement with Harman (1986).

5 BRIDGE PRINCIPLES FOR FULL AND PARTIAL BELIEF

Anyone who allows for the existence of both full and partial belief inherits a thorny problem: how are full beliefs related to partial beliefs? That seemingly innocent question leads to a treacherous search for *bridge principles* connecting a rational agent's partial beliefs with her full beliefs. Theorists engaged in the search for bridge principles usually take for granted some set of rationality principles governing full belief and its revision e.g. AGM theory, or a rival system of non-monotonic reasoning. Theorists usually also take for granted that partial belief ought to be representable by probability functions obeying some flavor of Bayesian rationality. The challenge is to propose additional rationality postulates governing how a rational agent's partial beliefs cohere with her full beliefs. In this section, we will for the most part accept received wisdom and assume that orthodox Bayesianism is the correct model of partial belief and its updating. We will be more open-minded about the modeling of full belief and its rational revision.

In this section, we will once again take propositions to be the objects of belief. In the background, there will be a (usually finite) set W of possible worlds. As before, the reader is invited to think of W as a set of coarse-grained, mutually exclusive, possible ways the actual world might be. The actual world is assumed to instantiate one of these coarse-grained possibilities. We write $\mathcal{B}$ to denote the set of *propositions* that the agent believes and use $\mathcal{B}(A)$ as shorthand for $A \in \mathcal{B}$. We will also require some notation for qualitative propositional belief change. For all $E \subseteq W$, write $\mathcal{B}_E$ for the set of propositions the agent would believe upon learning E and no stronger proposition. We will also write $\mathcal{B}(A \mid E)$ as shorthand for $A \in \mathcal{B}_E$. By convention, $\mathcal{B} = \mathcal{B}_W$. If $\mathcal{F}$ is a set of propositions, we let $\mathcal{B}_{\mathcal{F}}$ be the set $\{\mathcal{B}_E : E \in \mathcal{F}\}$. The set $\mathcal{B}_{\mathcal{F}}$ represents an agent's *dispositions* to update her qualitative beliefs given information from $\mathcal{F}$.

The following normative constraint on the set of full beliefs $\mathcal{B}$ plays a large role in what follows.

For all propositions $A, B \subseteq W$: (Deductive Cogency)

1. $\mathcal{B}(W)$;
2. not $\mathcal{B}(\varnothing)$;
3. if $\mathcal{B}(A)$ and $A \subseteq B$, then $\mathcal{B}(B)$;
4. if $\mathcal{B}(A)$ and $\mathcal{B}(B)$ then $\mathcal{B}(A \cap B)$.

The first two clauses say that the agent believes the true world to be among the worlds in W and that she does not believe the empty set to contain the true world. The third clause says that belief is closed under single-premise entailment, i.e. if the agent believes A and A logically entails B, then she believes B. The final clause says that the agent's beliefs are closed under conjunction, i.e. if she believes A and she believes B, then she believes $A \cap B$. Together, clauses 3 and 4 say that the agent's beliefs are closed under entailment by finitely many premises. When W is finite, the set $\mathcal{B}$ must be finite as well, implying that Deductive Cogency is equivalent to the following formulation:

$\mathcal{B}$ is consistent and $\mathcal{B}(B)$ iff $\cap\mathcal{B} \subseteq B$. (Deductive Cogency)

In other words, Deductive Cogency means that there is a single, non-empty proposition, which is the logically strongest proposition that the agent believes, entailing all her other beliefs. When the two formulations of Deductive Cogency come apart, we will always mean the latter one.Deductive Cogency only mentions the set of full beliefs $\mathcal{B}$, and is therefore not a bridge principle at all. Bridge principles are expressed as constraints holding for pairs $\langle \mathcal{B}, p \rangle$.

All of the rationality norms that we have seen for updating qualitative beliefs have propositional analogues. The following are propositional analogues for the six basic AGM principles. Here E, F are arbitrary subsets of W.

$\mathcal{B}_E = Cn(\mathcal{B}_E)$. (Closure)

$E \in \mathcal{B}_E$. (Success)

$\mathcal{B}_E \subseteq Cn(\mathcal{B} \cup \{E\})$. (Inclusion)

If $\neg E \notin Cn(\mathcal{B})$ then $\mathcal{B} \subseteq \mathcal{B}_E$. (Preservation)

$\mathcal{B}_E$ is consistent if $E \neq \varnothing$. (Consistency)

If $E \equiv F$, then $\mathcal{B}_E = \mathcal{B}_F$. (Extensionality)

$\mathcal{B}_{E\cap F} \subseteq Cn(\mathcal{B}_E \cup \{F\})$. (Conjunctive inclusion)

If $\neg F \notin Cn(\mathcal{B}_E)$, then
$Cn(B_E \cup \{F\}) \subseteq B_{E\cap F}$. (Conjunctive preservation)

Supposing that for all $E \subseteq W$, $\mathcal{B}_E$ satisfies Deductive Cogency, the first six postulates reduce to the following three, for arbitrary $E \subseteq W$.

$\cap\mathcal{B}_E \subseteq E$. (Success)

$\cap\mathcal{B} \cap E \subseteq \cap\mathcal{B}_E$. (Inclusion)

If $\cap\mathcal{B} \nsubseteq \neg E$, then $\cap\mathcal{B}_E \subseteq \cap\mathcal{B} \cap E$. (Preservation)

Together, Inclusion and Preservation say that whenever information E is consistent with current belief $\cap\mathcal{B}$,

$$\cap\mathcal{B}_E = \cap\mathcal{B} \cap E.$$

If $\mathcal{F}$ is a collection of propositions and for all $E \in \mathcal{F}$, the belief sets $\mathcal{B}, \mathcal{B}_E$ satisfy the AGM principles, we say that $\mathcal{B}_\mathcal{F}$, the agent's disposition to update her qualitative beliefs given information from $\mathcal{F}$, satisfies the basic AGM principles.

We will use $p(\cdot)$ to denote the probability function representing the agent's partial beliefs. Of course, $p(\cdot)$ is defined on a σ-algebra of subsets of W. In the usual case, when W is finite, we can take the $\wp(W)$ to be the relevant σ-algebra. To update partial belief, we adopt the standard probabilistic modeling. For $E \subseteq W$ such that $p(E) > 0$, $p(\cdot \mid E)$ is the partial belief function resulting from learning E. We will sometimes use p_E as a shorthand for $p(\cdot \mid E)$. Almost always, partial belief is updated via conditioning:

$$p(A \mid E) = \frac{p(A \cap E)}{p(E)}, \text{ whenever } p(E) > 0.$$

Let $\mathcal{F}_p^+$ be the set of propositions with positive probability according to p, i.e $\{A \subseteq W : p(A) > 0\}$.

5.1 *Belief as Extremal Probability*

The first bridge principle that suggests itself is that full belief is just the maximum degree of partial belief. Expressed probabilistically, it says that at all times a rational agent's beliefs and partial beliefs can be represented by a pair $\langle \mathcal{B}, p \rangle$ satisfying:

$\mathcal{B}(A)$ iff $p(A) = 1$. (Extremal Probability)

Roorda (1995) calls this the *received view* of how full and partial belief ought to interact. Gärdenfors (1986) is a representative of this view, as are van Fraassen (1995) and Arló-Costa (1999), although the latter two accept a slightly non-standard probabilistic modeling for partial belief. For fans of Deductive Cogency, the following observations ought to count in favor of the received view.

Theorem 21 *If $\langle \mathcal{B}, p \rangle$ satisfy extremal probability, then $\mathcal{B}$ is deductively cogent.*

Gärdenfors (1986) proves the following.

Theorem 22 *Suppose that $\langle \mathcal{B}_E, p_E \rangle$ satisfy extremal probability for all $E \in \mathcal{F}_p^+$. Then $\mathcal{B}_{\mathcal{F}_p^+}$ satisfies the AGM postulates.*

In other words: if an agent's partial beliefs validate the probability axioms, she updates by Bayesian conditioning and fully believes all and only those propositions with extremal probability, her qualitative update behavior will satisfy all the AGM postulates (at least whenever Bayesian conditioning is defined). Readers who take the AGM revision postulates to be a *sine qua non* of rational belief update will take this to be good news for the received view.

Roorda (1995) makes three criticisms of the received view. Consider the following three propositions.

1. Millard Fillmore was the 13th President of the United States;
2. Millard Fillmore was a U.S. President;
3. Millard Fillmore either was or was not a U.S. President.

Of course, I am not as confident that Fillmore was the 13^{th} president as I am in the truth of the tautology expressed in (3). Yet there does not seem to be anything wrong with saying that I fully believe each of (1), (2), and (3). However, if extremal probability is right, it is irrational to fully believe each of (1), (2), and (3) and not assign them all the same degree of belief.

Roorda's second objection appeals to the standard connection between degrees of belief and practical decision making. Suppose I fully believe (1). According to the standard interpretation of degrees of belief in terms of betting quotients, I ought to be accept a bet that pays out a dollar if (1) is true, and costs me a million dollars if (1) is false. In fact, if I truly assign unit probability to (1), I ought to accept nearly any stakes whatsoever that guarantee some positive payout if (1) is true. Yet it seems perfectly rational to fully believe (1) and refrain from accepting such a bet. If we accept Bayesian decision theory, extremal probability seems to commit me to all sorts of weird and seemingly irrational betting behavior.

Roorda's final challenge to extremal probability appeals to *corrigibility*, according to which it is reasonable to believe that at least some of my

beliefs may need to be abandoned in light of new information. However, if partial beliefs are updated via Bayesian conditioning, I can never cease to believe any of my full beliefs since if $p(A) = 1$ it follows that $p(A \mid E) = 1$ for all E such that $p(E) > 0$. If we believe in Bayesian conditioning, extremal probability seems to entail that I cannot revise any of my full beliefs in light of new information.objections to be decisive against extremal probability.

5.2 *The Lockean Threshold*

The natural response to the difficulties with the received view is to retreat from full certainty. Perhaps full belief corresponds to partial belief above some *threshold* falling short of certainty. Foley (1993) dubbed this view the *Lockean thesis*, after some apparently similar remarks in Book IV of Locke's *Essay Concerning Human Understanding*. So far, the Lockean thesis is actually ambiguous. There may be a single threshold that is rationally mandated for all agents and in all circumstances. Alternatively, each agent may have her own threshold that she applies in all circumstances—that threshold may characterize how "bold" or "risk-seeking" the agent is in forming qualitative beliefs. A yet weaker thesis holds that the threshold may be contextually determined. We distinguish the strong, context-independent Lockean thesis (SLT) from the weaker, context-dependent thesis (WLT). The domain of the quantifier may be taken as the set of all belief states $\langle \mathcal{B}, p \rangle$ a *particular* agent may find herself in, or as the set of all belief states whatsoever.

STRONG LOCKEAN THESIS (SLT). There is a threshold $\frac{1}{2} < s < 1$ such that all rational $\langle \mathcal{B}, p \rangle$ satisfy

$$\mathcal{B}(A) \text{ iff } p(A) \geq s.$$

WEAK LOCKEAN THESIS (WLT). For every rational $\langle \mathcal{B}, p \rangle$ there is a threshold $\frac{1}{2} < s < 1$ such that

$$\mathcal{B}(A) \text{ iff } p(A) \geq s.$$

Most discussions of the Lockean thesis have in mind the strong thesis. More recent work, especially Leitgeb (2017), adopts the weaker thesis. The strong thesis leaves the correct threshold unspecified. Of course for every $\frac{1}{2} < s < 1$, we can formulate a specific thesis SLT^s in virtue of which the strong thesis is true. For example, $\text{SLT}^{.51}$ is a very permissive version of the thesis, whereas $\text{SLT}^{.95}$ and $\text{SLT}^{.99}$ are more stringent. It is also possible to further specify the weak thesis. For example, Leitgeb (2017) believes that the contextually-determined threshold should be equal to the degree of

belief assigned to the strongest proposition that is fully believed. In light of Deductive Cogency, that corresponds to the orthographically ungainly $\mathrm{WLT}^{p(\cap\mathcal{B})}$.

The strong Lockean thesis gives rise to the well-known *Lottery paradox*, due originally to Kyburg (1961, 1997). The lesson of the Lottery is that the strong thesis is in tension with Deductive Cogency. Suppose that s is the universally correct Lockean threshold. Now think of a fair lottery with N tickets, where N is chosen large enough that $1 - (1/N) \geq s$. Since the lottery is fair, it seems permissible to fully believe that *some* ticket is the winner. It also seems reasonable to assign degree of belief $1/N$ to each proposition of the form "The i^{th} ticket is the winner." According to the Lockean thesis, such an agent ought to fully believe that the first ticket is a loser, the second ticket is a loser, the third is a loser, etc. Since cogency requires belief to be closed under conjunction, she ought to believe that all the tickets are losers. But now she violates cogency by believing both that every ticket is a loser and that some ticket is a winner. Since s was arbitrary, we have shown that no matter how high we set the threshold, there is some Lottery for which an agent must either violate the Lockean thesis or violate Deductive Cogency. According to Kyburg, what the paradox teaches is that we should give up on Deductive Cogency: full belief should not necessarily be closed under conjunction. Many others take the lesson of the Lottery to be that the strong Lockean thesis is untenable.

Several authors attempt to revise the strong Lockean thesis by placing restrictions on when a high degree of belief warrants full belief. Broadly speaking, they propose that a high degree of belief is sufficient to warrant full belief unless some defeating condition holds. For example, Pollock (1995) proposes that, although a degree of belief in P above some threshold is a *prima facie* reason for belief, that reason is defeated whenever P is a member of an inconsistent set of propositions each of which is also believed to a degree exceeding the threshold. Ryan (1996) proposes that a high degree of belief is sufficient for full belief unless the proposition is a member of a set of propositions such that each member has a degree of belief exceeding the threshold, but the probability of their conjunction is below the threshold. Douven (2002) says that it is sufficient except when the proposition is a member of a *probabilistically self-undermining* set. A set $\mathcal{S}$ is probabilistically self undermining iff for all $A \in \mathcal{S}, p(A) > s$ and $p(A \mid B) \leq s$, where $B = \bigcap(\mathcal{S} \setminus \{A\})$. It is clear that any of these proposals would prohibit full belief that a particular lottery ticket will lose.

These proposals are all vitiated by the following sort of example due to Korb (1992). Let A be any proposition with a degree of belief above threshold but short of certainty. Let L_i be the proposition that the i^{th} lottery ticket (of a large lottery with N tickets) will lose. Consider the set $\mathcal{S} = \{\neg A \cup L_i \mid 1 \leq i \leq N\}$. Each member of $\mathcal{S}$ is above threshold,

since L_i is above threshold. Furthermore, the set $\mathcal{S} \cup \{A\}$ meets all three defeating conditions. Therefore, these proposals prohibit full belief in any proposition with degree of belief short of certainty. Douven and Williamson (2006) generalize this sort of example to trivialize an entire class of similar formal proposals.

Buchak (2014) argues that what partial beliefs count as full beliefs cannot merely be a matter of the degree of partial belief, but must also depend on the type of evidence it is based on. According to Buchak, this means there can be no merely formal answer to the question: what conditions on partial belief are necessary and sufficient for full belief? The following example, of a type going back to Thomson (1986), illustrates the point. Your parked car was hit by a bus in the middle of the night. The bus could belong either to the blue bus company or the red bus company. Consider the following two scenarios.

SCENARIO 1. You know that the blue company operates 90% of the buses in the area, and the red bus company operates only 10%. You have degree of belief 0.9 that a blue bus is to blame.

SCENARIO 2. The red and blue companies operate an equal number of buses. A 90% reliable eyewitness testifies that a blue bus hit your car. You have degree of belief 0.9 that a blue bus is to blame.

Buchak (2014) argues that it is rational to have full belief that a blue bus is to blame in the second scenario, but not in the first. You have only statistical evidence in the first scenario, whereas in the second, a causal chain of events connects your belief to the accident (see also Thomson, 1986, Nelkin, 2000, and Schauer, 2003). These intuitions, Buchak observes, are reflected in our legal practice: purely statistical evidence is not sufficient to convict. If you find Buchak's point convincing, you will be unsatisfied with most of the proposed accounts for how full and partial belief ought to correspond (Staffel, 2016).

Despite difficulties with buses and lotteries, the dynamics of qualitative belief under the strong thesis are independently interesting to investigate. For example, van Eijck and Renne (2014) axiomatize the logic of belief for a Lockean with threshold $\frac{1}{2}$. Makinson and Hawthorne (2015) investigate which principles of non-monotonic logic are validated by Lockean agents. Before turning to proposed solutions to the Lottery paradox, we make some observations about qualitative Lockean revision, inspired largely by Shear and Fitelson (2018).

It is a theorem of the probability calculus that $p(H \mid E) \leq P(E \to H)$. So if H is assigned a high degree of belief given E, the material conditional $E \to H$ must have been assigned a degree of belief at least as high *ex ante*. It is easy to see that as a probabilistic analogue of the principle of

Conditionalization from non-monotonic logic or, equivalently, the AGM Inclusion principle. That observation has the following consequence: any belief that the Lockean comes to have after conditioning, she could have arrived at by adding the evidence to her prior beliefs and closing under logical consequence. Therefore Lockean updating satisfies the AGM principle of Inclusion. Furthermore, it follows immediately from definitions that Lockean update satisfies Success and Extensionality.

Theorem 23 *Suppose that* $s \in (\frac{1}{2}, 1)$. *Let* $\mathcal{B}_E = \{A : p(A \mid E) \geq s\}$ *for all* $E \in \mathcal{F}_p^+$. *Then* $\mathcal{B}_{\mathcal{F}_p^+}$ *satisfies Inclusion, Success, and Extensionality.*

In Section 2.2.1, we argued that Inclusion and Preservation capture the spirit of AGM revision. If Lockean revision also satisfied Preservation, we would have a clean sweep of the AGM principles, with the exception of Deductive Cogency.

However, that cannot hold in general. It is possible to construct examples where $p(\neg E) < s$, $p(H) \geq s$, and yet $p(H \mid E) < s$. For Lockean agents this means that it is possible to lose a belief, even when revising on a proposition that is not disbelieved.

Recall the example of Alice, Bob, and the Ford from Section 2.1.1. Let $W = \{a, b, c\}$ corresponding to the worlds in which Alice owns the Ford, Bob owns the Ford, and no one in the office owns the Ford. Suppose the probability function

$$p(a) = \frac{6}{10},$$
$$p(b) = \frac{3}{10},$$
$$p(c) = \frac{1}{10},$$

captures my partial beliefs. For Lockean thresholds in the interval $(.75, .9]$, my full beliefs are exhausted by $\mathcal{B} = \{\{a, b\}, W\}$. Now suppose I were to learn that Alice does not own the Ford. That is consistent with all beliefs in $\mathcal{B}$, but since $p(\{a, b\} \mid \{b, c\}) = \frac{3}{4}$, it follows by the Lockean thesis that $\{a, b\} \notin \mathcal{B}_{\{b,c\}}$. So Lockeanism does not in general validate Preservation. The good news, at least for those sympathetic to Pollock's critique of non-monotonic logic, is that the Lockean thesis allows for undercutting defeat of previous beliefs.

However, Shear and Fitelson (2018) also have some good news for fans of AGM and the Lockean thesis. Two quantities are in the *golden ratio* ϕ if their ratio is the same as the ratio of their sum to the larger of the two quantities, i.e. for $a > b > 0$, if $\frac{a+b}{a} = \frac{a}{b}$ then $\frac{a}{b} = \phi$. The golden ratio is an irrational number approximately equal to 1.618. Its inverse ϕ^{-1} is approximately .618. Shear and Fitelson prove the following intriguing result.

Theorem 24 *Suppose that* $s \in (\frac{1}{2}, \phi^{-1}]$. *Let* $\mathcal{B}_E = \{A : p(A \mid E) \geq s\}$ *for all* $E \in \mathcal{F}_p^+$. *Let*

$$\mathcal{D} = \{E \subseteq W : E \in \mathcal{F}_p^+ \text{ and } \mathcal{B}_E \text{ is deductively cogent}\}.$$

Then $\mathcal{B}_\mathcal{D}$ *satisfies the six basic AGM postulates.*

That shows that for relatively low thresholds, Lockean updating satisfies all the AGM postulates—at least when we restrict to deductively cogent belief sets.

Why has the golden ratio turned up here? That is relatively simple to explain. The AGM Preservation postulate can be factored into the following two principles.

If $\neg E \notin Cn(\mathcal{B})$ and $E \in Cn(\mathcal{B})$ then $\mathcal{B} \subseteq \mathcal{B}_E$. (Cautious Monotony)

If $\neg E \notin Cn(\mathcal{B})$ and $E \notin Cn(\mathcal{B})$ then $\mathcal{B} \subseteq \mathcal{B}_E$. (Preservation B)

We have discussed Cautious Monotony in Section 2.1.1. It is widely accepted as a *sine qua non* of rational non-monotonic reasoning. Surprisingly, there is no Lockean threshold that satisfies Cautious Monotony in general.[8] However, if $p(H \mid E) < s$ it must be that $p(H \cap E) < s \cdot P(E) \leq s$, from which it follows that any violation of Cautious Monotony must be a violation of deductive closure. Moreover, Lockean updating with a threshold in $(\frac{1}{2}, \phi^{-1}]$ satisfies Preservation B. That follows immediately from the fact that for $s \in (\frac{1}{2}, \phi^{-1}]$, if $p(E) < s$ and $p(H \mid E) < s$, then $P(H \to \neg E) \geq s$. The proof of that fact hinges on a neat fact about the golden ratio: if $s > 0$, then $s \leq \phi^{-1}$ iff $s^2 \leq 1 - s$.[9]

5.3 *The Stability Theory of Belief*

For many, sacrificing Deductive Cogency is simply too high a price to pay for a bridge principle, even one so simple and intuitive as the strong Lockean thesis. That occasions a search for bridge principles that can be reconciled with Deductive Cogency. One proposal, due to Leitgeb (2013, 2014, 2015, 2017) and Arló-Costa and Pedersen (2012), holds that rational full belief corresponds to a stably high degree of belief, i.e. a degree of belief that remains high even after conditioning on new information. Leitgeb calls this view the *Humean thesis*, due to Hume's conception of belief as an idea of superior vivacity, but also of superior steadiness.[10] Leitgeb (2017) formalizes Hume's definition, articulating the following version of the thesis:

8 See Lemma 1 in Shear and Fitelson (2018).

9 Suppose that $s \in (\frac{1}{2}, \phi^{-1}]$ and $P(E) < s$ and $P(H \mid E) < s$. Then, $P(E)P(H \mid E) = P(H \cap E) < s^2 \leq 1 - s$, and therefore $1 - P(H \cap E) = P(H \to \neg E) \geq s$.

10 See Loeb (2002, 2010) for a detailed development of the stability theme in Hume's conception of belief..

HUMEAN THESIS (HT). For all rational pairs $\langle \mathcal{B}, p \rangle$ there is $s \geq 1/2$ such that

$$\mathcal{B}(A) \text{ iff } \neg B \notin \mathcal{B} \text{ implies } p(A \mid B) > s.$$

In other words: every full belief must have stably high conditional degree of belief, at least when conditioning on propositions which are not currently disbelieved. Since full belief occurs on both sides of the biconditional, it is evident that this is not a proposed *reduction* of full belief to partial belief, but rather a constraint that every rational agent must satisfy. The Humean thesis leaves the precise threshold s unspecified. Of course for every $\frac{1}{2} < s < 1$, we can formulate a specific thesis HT^s in virtue of which the thesis is true. For example, $\text{HT}^{.5}$ requires that every fully believed proposition remains more likely than its negation when conditioning on propositions not currently disbelieved.

Some form of stability is widely considered to be a necessary condition for *knowledge*. Socrates propounds such a view in the *Meno*. Paxson and Lehrer (1969) champion such a view in the epistemology literature post-Gettier. However, stability is not usually mooted as a condition of *belief*. Raidl and Skovgaard-Olsen (2017) claim that Leitgeb's stability condition is more appropriate in an analysis of knowledge and too stringent a condition on belief. A defender of the Humean thesis might say that every *rational* belief is possibly an instance of knowledge. Since knowledge is necessarily stable, unstable beliefs are *ipso facto* not known.however, be out of step with several decades of work in epistemology. The Gettier cases are celebrated cases of unstable beliefs, but widely believed to be justified. If they are justified, then surely they are rational as well.

Leitgeb demonstrates the following relationships between the Humean thesis, Deductive Cogency, and the weak Lockean thesis.

Theorem 25 *Suppose that* $\langle \mathcal{B}, p \rangle$ *satisfy* HT *and* $\varnothing \notin \mathcal{B}$. *Then,* $\mathcal{B}$ *is deductively cogent and* $\langle \mathcal{B}, p \rangle$ *satisfy* $\text{WLT}^{p(\cap \mathcal{B})}$.

So if an agent satisfies the Humean thesis and does not "fully" believe the contradictory proposition, her qualitative beliefs are deductively cogent and furthermore, she satisfies the weak Lockean thesis, where the threshold is set by the degree of belief assigned to $\cap\mathcal{B}$, the logically strongest proposition she believes. Leitgeb also proves the following partial converse.

Theorem 26 *Suppose that* $\mathcal{B}$ *is deductive cogent and* $\langle \mathcal{B}, p \rangle$ *satisfy* $\text{WLT}^{p(\cap \mathcal{B})}$. *Then,* $\langle \mathcal{B}, p \rangle$ *satisfy* $\text{HT}^{\frac{1}{2}}$ *and* $\varnothing \notin \mathcal{B}$.

Together, these two theorems say that the Humean thesis (with threshold $\frac{1}{2}$) is equivalent to Deductive Cogency and the weak Lockean thesis (with threshold $p(\cap\mathcal{B})$). Since it is always possible to satisfy $\text{HT}^{\frac{1}{2}}$, Leitgeb gives

us an ingenious way to reconcile Deductive Cogency with a version of the Lockean thesis.

Recall the example of the lottery. Let $W = \{w_1, w_2, \ldots, w_N\}$, where w_i is the world in which the i^{th} ticket is the winner. No matter how many tickets are in the lottery, a Humean agent cannot believe any ticket will lose. Suppose for a contradiction that she believes $W \setminus \{w_1\}$, the proposition that the first ticket will lose. Now suppose she learns $\{w_1, w_2\}$, that all but the first and second ticket will lose. This is compatible with her initial belief, but her updated degree of belief that the first ticket will lose must be $\frac{1}{2}$. That contradicts the Humean thesis. So she cannot believe that any ticket will lose. In this Lottery situation the agent cannot fully believe any non-trivial proposition. This example also shows how sensitive the Humean proposal is to the fine-graining of possibilities. If we coarsen W into the set of possibilities $W = \{w_1, w_2\}$, where w_1 is the world in which the first ticket is the winner and w_2 is "the" world in which some other ticket is the winner, the agent can believe that the first ticket will lose without running afoul of the Humean thesis.

Perhaps Buchak (2014) is right and no agent should have beliefs in lottery propositions—these beliefs would necessarily be formed on the basis of purely statistical evidence. Kelly and Lin (forthcoming) give another scenario in which Humean agents seem radically skeptical, but in situations which are evidentially unproblematic. Suppose the luckless Job goes in for a physical. On the basis of a thorough examination, the doctor forms the following dire opinion of his health: her degree of belief that Job will survive exactly n months is $\frac{1}{2^n}$. Therefore, her degree of belief that Job will not survive the year is $\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^{12}} > .999$. Shockingly, the Humean thesis prevents the doctor from forming *any* nontrivial beliefs. Let $\leq n$ be the proposition that Job survives at most n months and let $\geq n$ be the proposition that he survives at least n months. Let B be the strongest proposition that the doctor believes. Suppose for a contradiction that B entails some least upper bound for the number of Job's remaining months, i.e for some n, B entails $\leq n$ and does not entail $\leq n'$ for any $n' < n$. By construction, $p(B| \geq n) = p(n)/p(\geq n) = \frac{1}{2}$ for all n. But since $\geq n$ is compatible with B, the Humean thesis requires that $p(B| \geq n) > \frac{1}{2}$. Contradiction.

The example of the doctor suggests that the price of Humeanism is a rather extreme form of skepticism: in many situations a Humean agent will have no non-trivial full beliefs at all. That criticism is developed extensively in Rott (2017) and Douven and Rott (2018). The doctor also illustrates how the Humean proposal allows arbitrarily small perturbations of partial beliefs to be reflected as huge differences in full beliefs. Suppose the doctor is slightly more confident that Job will not survive a month, i.e. her survival probabilities decrease as $\frac{1}{2} + \epsilon, \frac{1}{4}, \frac{1}{8} - \epsilon, \frac{1}{16}, \frac{1}{32}, \ldots$. Now the

doctor can believe that Job will be dead in two months without running afoul of the Humean thesis.

So far we have inquired only into the synchronic content of the Humean proposal. What sort principles of qualitative belief update does it underwrite? Leitgeb demonstrates an intimate relationship between the AGM revision principles and the Humean thesis: every agent that satisfies the AGM principles, as well as a weak version of the Lockean thesis, must also satisfy the Humean thesis. So if you think that AGM theory is the correct theory of rational qualitative belief update (and you believe that a high degree of partial belief is a *necessary* condition of full belief) you must also accept the Humean thesis.

To present Leitgeb's result we have to introduce a few technical concepts. Say that a proposition A is p-stable^r iff for all $B \in \mathcal{F}_p^+$ such that $A \cap B \neq \varnothing$, $p(A \mid B) > r$. An immediate consequence of this definition is that if A is p-stable^r and A is consistent with $E \in \mathcal{F}_p^+$, then $A \cap E$ is p_E-stable^r. Let

$$\mathcal{S}_p^r = \{A : A \text{ is } p\text{-stable}^r\}.$$

Leitgeb proves that for $r \geq 1/2$, the set $\mathcal{S}_p^r$ is a system of spheres in the sense of Section 2.2.3. That is: there is some least element B of $\mathcal{S}_p^r$ such that all other elements constitute a nested, well-ordered sphere system centered on B. Recall that $\mathcal{S}_p^r(E)$ is defined to be $D \cap E$, where D is the closest sphere to B compatible with E. By the previous observation, $\mathcal{S}_p^r(E)$ is p_E-stable^r.

Leitgeb proves the following.

Theorem 27 *The following are equivalent.*

1. $\mathcal{B}_{\mathcal{F}_p^+}$ *satisfies all AGM postulates and for all* $E \in \mathcal{F}_p^+$, $A \in \mathcal{B}_E$ *only if* $p(A \mid E) > r$.
2. $\cap \mathcal{B}_E = \mathcal{S}_p^r(E) \in \mathcal{S}_{p_E}^r$.

We know from the result of Section 2.2.3 that for any AGM belief revision operation, there is a corresponding system of Grove spheres. Leitgeb has proven that any agent that validates the AGM postulates and the high-probability requirement can be modeled by the system of spheres generated by the p-stable^r propositions. For such an agent, all pairs $\langle \mathcal{B}_E, p_E \rangle$ satisfy the Humean thesis with threshold r. So any agent that violates the Humean thesis must either fail to satisfy the AGM postulates, or the high-probability requirement. Note that the converse is not true: it is not the case that that if all pairs $\langle \mathcal{B}_E, p_E \rangle$ satisfy the Humean thesis, then $\mathcal{B}_{\mathcal{F}_p^+}$ must satisfy the AGM postulates. To prove this, suppose that $\langle \mathcal{B}, p \rangle$ satisfy the Humean thesis and $\cap \mathcal{B} \subset E$ for some $E \in \mathcal{F}_p^+$. If we let $\mathcal{B}_E = \{E\}$, then $\langle \mathcal{B}_E, p_E \rangle$ satisfy the Humean thesis. However, such an agent patently violates Rational and even Cautious Monotony.

5.4 *The Tracking Theory*

Lin and Kelly (2012) propose that qualitative belief update ought to *track* partial belief update. On their picture, partial and full beliefs are maintained and updated by parallel cognitive systems. The first system, governed by the probabilistic norms of Bayesian coherence and conditioning, is precise, slow, and cognitively expensive. That system is engaged for important deliberations requiring a lot of precision and occurring without much time pressure e.g. retirement planning. The second, which in some way maintains and updates full beliefs, is quicker and less cognitively burdensome.[11] That system is engaged in ordinary planning: grocery shopping, or selecting a restaurant for a department event. What keeps these two parallel systems in sync with each other?

Lin and Kelly study *acceptance rules* that specify a mechanism for transitioning gracefully into the qualitative and out of the probabilistic system. An acceptance rule α maps every partial belief state p to a unique qualitative belief state $\alpha(p)$ with which it coheres. For example, the strong Lockean thesis determines an acceptance rule once we specify a threshold. The Humean thesis, on the other hand, underdetermines an acceptance rule, merely imposing constraints on acceptable pairs $\langle \mathcal{B}, p \rangle$. An agent's qualitative updates *track* her probabilistic updates iff

$$\alpha(p)_E = \alpha(p_E),$$

whenever $p(E) > 0$. In other words: acceptance followed by qualitative revision yields the same belief state as probabilistic revision followed by acceptance.

Here is a way to understand the tracking requirement. Suppose that, although an agent maintains a latent probabilistic belief state, most of her cognitive life is spent reasoning with and updating qualitative beliefs. A typical day will go by without having to engage the probabilistic system at all. Suppose Monday is a typical day. Let $\langle \alpha(p), p \rangle$ be the belief state she wakes up with on Monday: her full and partial beliefs are in harmony. Let E be the total information she acquired since waking up. Since qualitative beliefs are updated on the fly, she goes to sleep with the qualitative belief state $\alpha(p)_E$. Overnight, her probabilistic system does the difficult work of Bayesian conditioning and computes the partial belief state p_E, just in case she runs into any sophisticated decision problems on Tuesday. Before waking, she transitions out of her probabilistic system p_E and into the qualitative belief state $\alpha(p_E)$. If she fails the tracking requirement, she may wake up on Tuesday morning with a qualitative belief state that is drastically different from the one she went to sleep with on Monday night.

11 For an objection to the two systems view, see Staffel (2018).

If she tracks, then she will notice no difference at all. For such an agent, no mechanism (other than memory) is required to bring her full and partial beliefs back into harmony on Tuesday morning. Supposing that we *enter* the probabilistic system by conditioning our previous partial belief state p on all new information E, and *exit* by accepting $\alpha(p_E)$, tracking ensures that transitioning in and out of the probabilistic system does not induce any drastic changes in qualitative beliefs. An agent that tracks will notice no difference at all. An agent that does not track may find her full and partial beliefs perpetually falling out of sync, requiring many expensive acceptance operations to bring them back into harmony.

Tracking may be a desirable property, but are there any architectures that exhibit it? Lin and Kelly (2012) answer this question affirmatively. Since Bayesian conditioning is taken for granted, Lin and Kelly must specify two things: a qualitative revision operation and an acceptance rule that jointly track conditioning. We turn now to the details of their proposal. As usual, let W be a set of worlds. A *question* $\mathcal{Q}$ is a partition of W into a countable collection of mutually exhaustive propositions $H_1, H_2, \ldots$, which are the complete *answers* to $\mathcal{Q}$. The partial belief function p is defined over the algebra of propositions $\mathcal{A}$ generated by $\mathcal{Q}$.

First we specify an acceptance rule. Lin and Kelly propose the *odds threshold rule*. The degree of belief function p is used to determine a plausibility order by setting

$$H_i \prec_p H_j \qquad \text{if and only if} \qquad \frac{p(H_i)}{p(H_j)} > t,$$

where t is a constant greater than 1 and $p(H_i), p(H_j) > 0$. This determines an acceptance rule by setting $\alpha(p) = \mathcal{B}_{\prec_p}$. Since the odds threshold rule determines a plausibility order $\prec_p$ and any plausibility order $\prec$ gives rise to a deductively cogent belief state $\mathcal{B}_{\prec}$, the Lottery paradox is avoided. In other words: the bridge principle that any rational $\langle \mathcal{B}, p \rangle$ are related by $\mathcal{B} = \alpha(p)$ ensures that $\mathcal{B}$ is deductively cogent. Furthermore, the odds threshold rule allows non-trivial qualitative beliefs in situations where the stability theory precludes them. Recall the case of the doctor. Consider the odds threshold $2^{10} - 1$. Given this threshold, the hypothesis that Job will survive exactly 1 month is strictly more plausible than the proposition that he will survive at least n months for any $n \geq 10$. This threshold yields the full belief that Job will survive at most 10 months. However, in the case of the Lottery the odds threshold rule precludes any non-trivial beliefs.[12] See Rott (2017) and Douven and Rott (2018) for an extensive comparison of the relative likelihood of forming non-trivial qualitative beliefs on the odds-threshold and stability proposals.

12 The content-dependent threshold rule proposed by Kelly and Lin (forthcoming) may allow non-trivial beliefs in the Lottery situation.

It remains to specify the qualitative revision operation. Lin and Kelly adopt an operation proposed by Shoham (1987). Let $\prec$ be a well-founded, strict partial order over the answers to $\mathcal{Q}$.[13] This is interpreted as a *plausibility ordering*, where $H_i \prec H_j$ means that H_i is strictly *more* plausible than H_j. Every plausibility order $\prec$ gives rise to a belief state $\mathcal{B}_\prec$ by letting $\neg H_i \in \mathcal{B}_\prec$ iff there is some H_j strictly more plausible than H_i and closing under logical consequence. In other words, $\cap\mathcal{B}_\prec$ is the disjunction of the minimal elements in the plausibility order. The plausibility order $\prec$ is updated on evidence E by setting every answer incompatible with E to be strictly less plausible than every answer compatible with E, and otherwise leaving the order unchanged. Let $\prec_E$ denote the result of this update operation. We use the updated plausibility order to define a belief revision rule by setting $\mathcal{B}_E = \mathcal{B}_{\prec_E}$. Then, for all $E, F \subseteq W$, B_E is deductively cogent and satisfies:

$$\cap\mathcal{B}_E \subseteq E; \tag{Success}$$

$$\cap\mathcal{B} \cap E \subseteq \cap\mathcal{B}_E; \tag{Inclusion}$$

$$\text{if } \cap\mathcal{B} \subseteq E \text{ then } \cap\mathcal{B}_E \subseteq \cap\mathcal{B}. \tag{Cautious monotony}$$

However, it does not necessarily satisfy Preservation. To see this suppose that $\mathcal{Q} = \{H_1, H_2, H_3\}$ and $H_1 \prec H_2$ but H_3 is not ordered with H_1 or H_2. Then $\cap\mathcal{B} = H_1 \cup H_3$. However $\cap\mathcal{B}_{\neg H_1} = H_2 \cup H_3 \nsubseteq \cap\mathcal{B}$ even though $\cap\mathcal{B} \cap \neg H_1 \neq \varnothing$.

Lin and Kelly prove that Shoham revision and odds-threshold based acceptance jointly track conditioning.

Theorem 28 *Let $\prec$ equal $\prec_p$ and let $\mathcal{B}_E = \mathcal{B}_{\prec_E}$. Then $\mathcal{B}_{\wp(W)}$ satisfies Deductive Cogency, Success, Cautious Monotony, and Inclusion. Furthermore, $\mathcal{B}_E = \alpha(p)_E = \alpha(p_E)$ for all $E \in \mathcal{F}_p^+$.*

In other words: odds-threshold acceptance followed by Shoham revision yields the same belief state as Bayesian conditioning followed by odds-threshold acceptance.[14] Although the original plausibility ordering $\prec_p$ is built from the probability function p, subsequent qualitative update proceeds without consulting the (conditioned) probabilities. That shows that there are at least some architectures that effortlessly keep the probabilistic and qualitative reasoning systems in harmony.

Fans of AGM will regret that Shoham revision does not satisfy AGM Preservation (Rational Monotony). Lin and Kelly (2012) prove that no "sensible" acceptance rule that tracks conditioning can satisfy Inclusion and

13 A strict partial order is *well-founded* iff every subset of the order has a least element. This is closely related to the stopperedness property discussed in Section 2.1.2.

14 Kelly and Lin (forthcoming) recommend a modification of the odds-threshold rule proposed in Lin and Kelly (2012).

Preservation. According to Lin and Kelly, sensible acceptable rules are *non-skeptical*, *non-opinionated*, *consistent*, and *corner-monotonic*. An acceptance rule is *non-skeptical* iff for every answer H_i to $\mathcal{Q}$ there is a non-negligible set of probability functions p such that $H_i \in \alpha(p)$.[15] An acceptance rule is *non-opinionated* iff there is a non-negligible set of probability functions p where judgement is suspended, i.e. where $\cap\alpha(p) = W$. An acceptance rule is *consistent* iff for all p, $\alpha(p)$ is deductively cogent. The intuition behind corner-monotony is that if H_i is accepted at p, then H_i should still be accepted if H_i is made more probable. More precisely an acceptance rule is *corner-monotone* iff $H_i \in \alpha(p)$ implies that $H_i \in \alpha(p')$ for all p' such that

$$p' = p(\cdot \mid H_i) \cdot q + p(\cdot \mid \neg H_i) \cdot (1 - q),$$

and $q > p(H_i)$. Lin and Kelly (2012) prove the following "no-go" theorem for AGM revision.

Theorem 29 *Suppose that $\mathcal{B}_E = \alpha(p_E)$ for $E \in \mathcal{F}_p^+$. Then $\mathcal{B}_{\mathcal{F}_p^+}$ satisfies Inclusion and Preservation only if α is not sensible.*

5.5 *Decision-Theoretic Accounts*

All of the bridge principles we have seen so far have the following in common: whether an agent's full and partial beliefs cohere is a matter of the full and partial beliefs *alone*. It is not necessary to mention preferences or utilities in order to evaluate a belief state. There is another tradition, originating in Hempel (1962) and receiving classical expression in Levi (1967), that assimilates the problem of "deciding" what to believe to a Bayesian decision-theoretic model. Crucially, these authors are not committed to a picture on which agents literally decide what to believe—rather they claim that an agent's beliefs are subject to the same kind of normative evaluation as their practical decision-making. Contemporary contributions to this tradition include Easwaran (2015a), Pettigrew (2016c), and Dorst (2017). Presented here is a somewhat simplified version of Levi's (1967) account taking propositions, rather than sentences, as the objects of belief.

As usual, let W be a set of possible worlds. The agent is taken to be interested in answering a *question* $\mathcal{Q}$, which is a partition of W into a finite collection of mutually exhaustive answers $\{H_1, H_2, \ldots H_n\}$. Levi calls situations of this sort "efforts to replace agnosticism by true belief," echoing themes in Peirce (1877).

15 A set of probability functions is *non-negligible* iff it contains an open set in the topology generated by the metric

$$||p - q|| = \sqrt{\sum_{H_i \in \mathcal{Q}} (p(H_i) - q(H_i))^2}.$$

> Doubt is an uneasy and dissatisfied state from which we struggle to free ourselves and pass into the state of belief; while the latter is a calm and satisfactory state which we do not wish to avoid, or to change to a belief in anything else. On the contrary, we cling tenaciously, not merely to believing, but to believing just what we do believe.

The agent's partial beliefs are represented by a probability function p that is defined, at a minimum, over the algebra $\mathcal{A}$ generated by the question. Levi recommends the following procedure to determine which propositions are fully believed: disjoin all those elements of $\mathcal{Q}$ that have *maximal* expected epistemic utility and then close under deductive consequence. The *expected epistemic utility* of a hypothesis $H \in \mathcal{A}$ is defined as:

$$E(H) := p(H) \cdot U(H) + p(\neg H) \cdot u(H),$$

where $U(H)$ is the epistemic utility of accepting H when it is true, and $u(H)$ is the utility of accepting H when it is false. How are $u(H), U(H)$ to be determined? Levi is guided by the following principles.

1. True answers have greater epistemic utility than false answers.
2. True answers that afford a high degree of relief from agnosticism have greater epistemic utility than true answers that afford a low degree of relief from agnosticism.
3. False answers that afford a high degree of relief from agnosticism have greater epistemic utility than false answers that afford a low degree of relief from agnosticism.

It is easy to object to these principles. The first principle establishes a lexicographic preference for true beliefs. It is conceivable that, contra this principle, an informative false belief that is approximately true should have greater epistemic utility than an uninformative true belief. The first principle precludes trading content against truthlikeness. It is also conceivable that, contra the third principle, one would prefer to be wrong, but not too opinionated, than wrong and opinionated. The only unexceptionable principle seems to be the second.

To measure the degree of relief from agnosticism, a probability function $m(\cdot)$ is defined over the elements of $\mathcal{A}$. Crucially, $m(\cdot)$ does not measure a degree of belief, but degree of *uninformativeness*. The degree of relief from agnosticism afforded by $H \in \mathcal{A}$, also referred to as the *amount of content* in H, is defined to be the complement of uninformativeness: $cont(H_i) = m(\neg H_i)$. Levi argues that all the elements of $\mathcal{Q}$ ought to be assigned the same amount of content, i.e. $m(H_i) = \frac{1}{n}$ and therefore

$cont(H_i) = \frac{n-1}{n}$ for each $H_i \in \mathcal{Q}$. The set of epistemic utility functions that Levi recommends satisfy the following conditions:

$$\begin{aligned} U(H) &= \quad 1 - q \cdot cont(\neg H), \\ u(H) &= \quad\quad -q \cdot cont(\neg H), \end{aligned}$$

where $0 < q < 1$. All such utility functions are guaranteed to satisfy Levi's three principles. The parameter q is interpreted as a "degree of caution," representing the premium placed on truth as opposed to relief from agnosticism. When $q = 1$ the epistemic utility of suspending judgement, $U(W)$, is equal to zero. This is the situation in which the premium placed on relief from doubt is the maximum. Levi proves that expected epistemic utility $E(H)$ is maximal iff $p(H) > q \cdot cont(\neg H)$. Therefore, Levi's ultimate recommendation is that the agent believe all deductive consequences of

$$\bigcap \{\neg H_i \in \mathcal{Q} : p(\neg H_i) > 1 - q \cdot cont(\neg H_i)\}.$$

From this formulation it is possible to see Levi's proposal as a question-dependent version of the Lockean thesis where the appropriate threshold is a function of content. However, Levi takes pains to make sure that the result of this operation is deductively cogent and therefore avoids Lottery-type paradoxes.

Contemporary contributions to the decision-theoretic tradition proceed differently from Levi. Most recent work does not take epistemic utility to be primarily a function of content. Most of these proposals do not refer to a question in context. Many proposals, such as Easwaran (2015a) and Dorst (2017), are equivalent to a version of the Lockean thesis, where the threshold is determined by the utility the agent assigns to true and false beliefs. Since these are essentially Lockean proposals, they are subject to Lottery-style paradoxes.

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.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine. *Econometrica*, *21*(4), 503–546.

Arló-Costa, H. (1999). Qualitative and probabilistic models of full belief. In S. R. Buss, P. Hájek, & P. Pudlák (Eds.), *Proceedings of logic colloquim* (Vol. 98, pp. 25–43).

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.

Armendt, B. (1980). Is there a Dutch book argument for probability kinematics? *Philosophy of Science*, *47*, 583–588.

Briggs, R. (2017). Normative theories of rational choice: Expected utility. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Buchak, L. (2014). Belief, credence, and norms. *Philosophical Studies*, *169*(2), 285–311.

Carnap, R. (1947). On the application of inductive logic. *Philosophy and Phenomenological Research*, *8*, 133–148.

Cohen, L. J. (1977). *The probable and the provable*. Oxford: Clarendon Press.

Cohen, L. J. (1980). Some historical remarks on the Baconian conception of probability. *Journal of the History of Ideas*, 219–231.

Cohen, L. J. (1992). *An essay on belief and acceptance*. Clarendon Press: Oxford.

de Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. In *Annales de l'institut henri poincaré* (Vol. 7, *1*, pp. 1–68).

de Finetti, B. (1970). *Theory of probability*. New York: Wiley.

de Finetti, B. (1972). *Probability, induction and statistics*. New York: Wiley.

Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society (Series B, Methodological)*, *30*(2), 205–247.

Dorst, K. (2017). Lockeans maximize expected accuracy. *Mind*, *128*(509), 175–211.

Douven, I. (2002). A new solution to the paradoxes of rational acceptability. *British Journal for the Philosophy of Science*, *53*, 391–410.

Douven, I. & Rott, H. (2018). From probabilities to categorical beliefs: Going beyond toy models. *Journal of Logic and Computation*, *28*(6), 1099–1124.

Douven, I. & Williamson, T. (2006). Generalizing the lottery paradox. *British Journal for the Philosophy of Science*, *57*, 755–779.

Doyle, J. (1979). A truth maintenance system. *Artificial intelligence*, *12*(3), 231–272.

Doyle, J. (1992). Reason maintenance and belief revision: Foundations vs. coherence theories. In P. Gärdenfors (Ed.), *Belief revision* (pp. 29–52). Cambridge Tracts in Theoretical Computer Science. Cambridge University Press.

Earman, J. (1992). *Bayes or bust?: A critical examination of bayesian confirmation theory*. MIT Press.

Easwaran, K. (2011a). Bayesianism I: Introduction and arguments in favor. *Philosophy Compass*, *6*(5), 312–320.

Easwaran, K. (2011b). Bayesianism II: Applications and criticisms. *Philosophy Compass*, *6*(5), 321–332.

Easwaran, K. (2015a). Dr. Truthlove or: How I learned to stop worrying and love Bayesian probabilities. *Nous*, *50*(4), 816–853.

Easwaran, K. (2015b). Primitive conditional probabilities. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics*, *75*(4), 643–669.

Eriksson, L. & Hájek, A. (2007). What are degrees of belief? *Studia Logica*, *86*(2), 183–213.

Fine, T. (1973). *Theories of probability: An examination of foundations*. Elsevier.

Foley, R. (1993). *Working without a net: A study of egocentric epistemology*. Oxford University Press.

Frankish, K. (2009). Partial belief and flat-out belief. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (pp. 73–79). Synthese Library. Springer.

Gabbay, D. M. (1985). Theoretical foundations for non-monotonic reasoning in expert systems. In *Logics and models of concurrent systems* (pp. 439–457). Springer.

Gärdenfors, P. (1986). The dynamics of belief: Contractions and revisions of probability functions. *Topoi*, *5*(1), 29–37.

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

Gärdenfors, P. (1992). Belief revision: An introduction. In P. Gärdenfors (Ed.), *Belief revision* (pp. 1–29). Cambridge Tracts in Theoretical Computer Science. Cambridge University 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. (2017). How inductive is Bayesian conditioning? *Manuscript*. Retrieved from https://kgenin.github.io/papers/conditioning%5C_long.pdf

Gettier, E. L. (1963). Is justified true belief knowledge? *analysis*, *23*(6), 121–123.

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

Gyenis, Z., Hofer-Szabó, G., & Rédei, M. (2017). Conditioning using conditional expectations: The Borel-Kolmogorov paradox. *Synthese*, *194*(7), 2595–2630.

Hacking, I. (1975). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press.

Hájek, A. (2003). What conditional probability could not be. *Synthese*, *137*(3), 272–323.

Hájek, A. (2012). Interpretations of probability. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Hájek, A. & Lin, H. (2017). A tale of two epistemologies? *Res Philosophica*, *94*(2), 207–232.

Hansson, S. O. (1999). *A textbook of belief dynamics: Theory change and database updating*. Kluwer Academic Publishers.

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.

Harman, G. (1986). *Change in view: Principles of reasoning*. MIT Press.

Hempel, C. G. (1962). Deductive-nomological vs. statistical explanation. In H. Fiegl & G. Maxwell (Eds.), *Vol 3*. (pp. 98–169). Minnesota Studies in the Philosophy of Science. University of Minnesota Press.

Horgan, T. (2017). Troubles for bayesian formal epistemology. *Res Philosophica*, *94*(2), 233–255.

Horty, J. F. (2012). *Reasons as defaults*. Oxford University Press.

Howson, C. & Urbach, P. (2006). *Scientific reasoning: The bayesian approach*. Open Court Publishing.

Huber, F. (manuscript). *Belief and counterfactuals. a study in means-end philosophy.* Oxford University Press.

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. (2016). Formal representations of belief. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2016). Metaphysics Research Lab, Stanford University.

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

Jeffrey, R. C. (1970). Dracula meets Wolfman: Acceptance vs. Partial Belief. In M. Swain (Ed.), *Induction, acceptance, and rational belief*. Synthese Library. D. Reidl.

Jeffrey, R. C. (1983). *The logic of decision* (2nd). University of Chicago Press.

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, *65*, 575–603.

Kaplan, M. (1996). *Decision theory as philosophy*. Cambridge University Press.

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

Kelly, K. T. & Lin, H. (forthcoming). Beliefs, probabilities, and their coherent correspondence. In I. Douven (Ed.), *Lotteries, knowledge and rational belief: Essays on the lottery paradox*. Cambridge University Press.

Kemeny, J. G. (1955). Fair bets and inductive probabilities. *Journal of Symbolic Logic*, *20*, 263–273.

Keynes, J. M. (1921). A treatise on probability. vol. 8 of collected writings (1973 ed.) London: Macmillan.

Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. New York: Chelsea.

Konek, J. (2019). Comparative probabilities. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*.

Koopman, B. O. (1940). The axioms and algebra of intuitive probability. *The Annals of Mathematics*, *41*(2), 269–292.

Korb, K. B. (1992). The collapse of collective defeat: Lessons from the lottery paradox. In D. Hull, M. Forbes, & K. Okruhlik (Eds.), *Psa: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1, pp. 230–236).

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

Kvanvig, J. L. (2016). Intellectual humility: Lessons from the preface paradox. *Res Philosophica*, *93*(3), 509–532.

Kyburg, H. E. (1961). *Probability and the logic of rational belief*. Wesleyan University Press.

Kyburg, H. E. (1997). The rule of adjunction and reasonable inference. *The Journal of Philosophy*, *94*(3), 109–125.

Lehrer, K. (1965). Knowledge, truth and evidence. *Analysis*, *25*(5), 168–175.

Leitgeb, H. (2013). Reducing belief simpliciter to degrees of belief. *Annals of Pure and Applied Logic*, *164*(12), 1338–1389.

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

Leitgeb, H. (2015). I—the Humean thesis on belief. In *Aristotelian society supplementary volume* (Vol. 89, *1*, pp. 143–185). Wiley Online Library.

Leitgeb, H. (2017). *The stability of belief: How rational belief coheres with probability*. Oxford University Press.

Leitgeb, H. & Pettigrew, R. (2010). An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science*, *77*, 236–272.

Levi, I. (1967). *Gambling with truth: An essay on induction and the aims of science*. MIT Press.

Levi, I. (1977). Subjunctives, dispositions and chances. *Synthese*, *34*(4), 423–455.

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

Lewis, D. (1979). Attitudes de dicto and de se. *The philosophical review*, *88*(4), 513–543.

Lewis, D. (1999). Why conditionalize? In *Papers in metaphysics and epistemology* (pp. 403–407). Cambridge University Press.

Liao, S.-y. (2012). What are centered worlds? *The Philosophical Quarterly*, *62*(247), 294–316.

Lin, H. (2013). Foundations of everyday practical reasoning. *Journal of Philosophical Logic*, *42*(6), 831–862.

Lin, H. (2019). Belief revision theory. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*.

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

Loeb, L. E. (2002). *Stability and justification in hume's treatise*. Oxford University Press on Demand.

Loeb, L. E. (2010). *Reflection and the stability of belief: Essays on descartes, hume, and reid*. Oxford University Press on Demand.

Mahtani, A. (2019). Imprecise probabilities. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*.

Makinson, D. (1965). The paradox of the preface. *Analysis*, *25*(6), 205–207.

Makinson, D. (1994). General patterns in nonmonotonic reasoning. In D. M. Gabbay, C. Hogger, & J. Robinson (Eds.), *Handbook of logic in artificial intelligence and logic programming (vol. 3): Nonmonotonic reasoning and uncertain reasoning* (pp. 35–111). Oxford University Press.

Makinson, D. & Hawthorne, J. (2015). Lossy inference rules and their bounds: A brief review. In *The road to universal logic* (pp. 385–407). Springer.

Moon, A. (2017). Beliefs do not come in degrees. *Canadian Journal of Philosophy*, *47*(6), 760–778.

Moss, S. (2018). *Probabilistic knowledge*. Oxford University Press.

Nelkin, D. K. (2000). The lottery paradox, knowledge, and rationality. *The Philosophical Review*, *109*(3), 373–409.

Paxson, T., Jr. & Lehrer, K. (1969). Knowledge: Undefeated justified true belief. *The Journal of Philosophy*, *66*(8), 225–237.

Peirce, C. S. (1877). The fixation of belief. In N. Houser & C. Kloesel (Eds.), *The essential Pierce: Selected philosophical writings. vol. i.* Indiana University Press.

Perry, J. (1979). The problem of the essential indexical. *Noûs*, 3–21.

Pettigrew, R. (2016a). *Accuracy and the laws of credence*. Oxford University Press.

Pettigrew, R. (2016b). Epistemic utility arguments for probabilism. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Pettigrew, R. (2016c). Jamesian epistemology formalised: An explication of 'the will to believe'. *Episteme*, *13*(3), 253–268.

Pollock, J. L. (1987). Defeasible reasoning. *Cognitive Science*, *11*(4), 481–518.

Pollock, J. L. (1995). *Cognitive carpentry: A blueprint for how to build a person*. MIT Press.

Pollock, J. L. (2006). *Thinking about acting: Logical foundations for rational decision making*. Oxford University Press.

Popper, K. (1955). Two autonomous systems for the calculus of probabilities. *British Journal for the Philosophy of Science*, *6*(3-4), 51–57.

Popper, K. & Miller, D. (1983). A proof of the impossibility of inductive probability. *Nature*, *302*(5910), 687.

Quine, W. V. O. (1990). *Pursuit of truth*. Harvard University Press.

Quine, W. V. O. & Ullian, J. S. (1970). *The web of belief*. Random House.

Raidl, E. & Skovgaard-Olsen, N. (2017). Bridging ranking theory and the stability theory of belief. *Journal of Philosophical Logic*, *46*(6), 577–609.

Ramsey, F. P. (1931). Truth and probability. In R. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (pp. 156–199). Routledge.

Renyi, A. (1955). On a new axiomatic system for probability. *Acta Mathematica Academiae Scientiarum Hungaricae*, *6*, 285–335.

Resnik, M. D. (1987). *Choices: An introduction to decision theory*. University of Minnesota Press.

Roorda, J. (1995). Revenge of Wolfman: A probabilistic explication of full belief. *unpublished*. Retrieved from https://www.princeton.edu/~bayesway/pu/Wolfman.pdf

Ross, D. (1930). *The right and the good*. Oxford University Press.

Rott, H. (2003). Basic entrenchment. *Studia Logica*, *73*(2), 257–280.

Rott, H. (2017). Stability and scepticism in the modelling of doxastic states: Probabilities and plain beliefs. *Minds and Machines*, *27*(1), 167–197.

Ryan, S. (1996). The epistemic virtues of consistency. *Synthese*, *109*, 121–141.

Savage, L. J. (1954). *The foundations of statistics*. Wiley publications in statistics. Wiley.

Schauer, F. (2003). *Profiles, probabilities and stereotypes*. Belknap Press.

Schurz, G. (2011). Abductive belief revision in science. In *Belief revision meets philosophy of science* (pp. 77–104). Springer.

Schwitzgebel, E. (2015). Belief. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.

Shear, T. & Fitelson, B. (2018). Two approaches to belief revision. *Erkenntnis*.

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

Skyrms, B. (2009). Diachronic coherence and probability kinematics. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of belief* (pp. 73–79). Synthese Library. Springer.

Spohn, W. (1988). Ordinal conditional functions: A dynamic theory of epistemic states. In *Causation in decision, belief change, and statistics* (pp. 105–134). Springer.

Spohn, W. (2012). *The laws of belief: Ranking theory and its philosophical applications*. Oxford University Press.

Spohn, W. (2017). Knightian uncertainty meets ranking theory. *Homo Oeconomicus*, *34*(4), 293–311.

Spohn, W. (2019). Defeasible normative reasoning. *Synthese*.

Staffel, J. (2013). Can there be reasoning with degrees of belief? *Synthese*, *190*(16), 3535–3551.

Staffel, J. (2016). Beliefs, buses and lotteries: Why rational belief can't be stably high credence. *Philosophical Studies*, *173*(7), 1721–1734.

Staffel, J. (2018). How do beliefs simplify reasoning? *Noûs*.

Stalnaker, R. (1981). Indexical belief. *Synthese*, *49*(1), 129–151.

Stalnaker, R. (1984). *Inquiry*. MIT Press.

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

Strasser, C. & Antonelli, G. A. (2018). Non-monotonic logic. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Teller, P. (1973). Conditionalization and observation. *Synthese*, *26*, 218–258.

Thoma, J. (2019). Decision theory. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*.

Thomson, J. J. (1986). Liability and individualized evidence. *Law and Contemporary Prolems*, *49*(3), 199–219.

Ullman-Margalit, E. (1983). On presumption. *The Journal of Philosophy*, *80*(3), 143–163.

van Fraassen, B. C. (1995). Fine-grained opinion, probability, and the logic of full belief. *Journal of Philosophical logic*, *24*(4), 349–377.

van Eijck, J. & Renne, B. (2014). Belief as willingness to bet. *CoRR*, *abs/1412.5090*. arXiv: 1412.5090. Retrieved from http://arxiv.org/abs/1412.5090

von Neumann, J. & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Weirich, P. (2004). Belief and acceptance. In I. Niiniluoto, M. Sintonen, & J. Woleński (Eds.), *Handbook of epistemology* (pp. 499–520). Dordrecht: Kluwer Academic Publishers.

Weisberg, J. (2011). Varieties of bayesianism. In D. M. Gabbay, S. Hartmann, & J. Woods (Eds.), *Volume 10: Inductive logic* (pp. 477–553). Handbook of the History of Logic. Elsevier.

10

DOXASTIC LOGIC

Michael Caie

There are at least three natural ways of interpreting the object of study of doxastic logic. On one construal, doxastic logic studies certain general features of the doxastic states of actual agents. On another construal, it studies certain general features of *idealized* doxastic states. While on yet another construal, doxastic logic provides a normative account of what features an agent's doxastic state ought to have on pain of irrationality.

The field of doxastic logic was initiated by Hintikka (1962), where techniques from modal logic were employed to model doxastic and epistemic states and to characterize certain valid principles governing such states. The theory presented by Hintikka provides an account of certain synchronic principles governing doxastic states. That theory, however, is silent on the question of how the doxastic states of an agent over time are or should be related. Later work, initiated by Alchourrón, Gärdenfors, and Makinson (1985), sought to provide an account of how an agent will or should revise her doxastic state in the light of new evidence. According to the accounts developed out of Alchourrón et al., a characterization of an agent's doxastic state should include, in addition to the set of beliefs the agent has, a characterization of the agent's belief revision policy.

One of the characteristic features of the models developed by Hintikka is that they provide a natural way of modeling the higher-order beliefs of an agent, i.e., the agent's beliefs about her own beliefs. The models developed out of Alchourrón et al., however, don't provide a natural way of characterizing an agent's higher-order beliefs about her belief revision policy. More recent work in dynamic doxastic logic has attempted to remedy this defect by providing a semantics for an object language that includes not only a unary belief operator but also a binary belief revision operator.

In Section 1, I'll outline the theory developed by Hintikka and briefly discuss how this theory looks given each of the above construals. In Section 2, I'll consider the theory of belief revision that developed out of Alchourrón et al. In Section 3, I'll discuss more recent work in dynamic doxastic logic. And, finally, in Section 4, I'll consider some paradoxes of doxastic logic and the bearing that these have on some of the accounts considered in Section 1–3.

1 STATIC DOXASTIC LOGIC

In Section 1.1, I'll first provide a quick overview of the basic theory developed by Hintikka (1962). In Section 1.2, I'll then consider how these doxastic models may be extended to characterize the doxastic states of multiple agents and various collective doxastic properties. The presentation of this material will work under the assumption that the theory serves to characterize certain features of an *idealized* doxastic state. Having outlined the basic theory, however, in Section 1.3, I'll consider how the theory looks under alternate interpretations.

1.1 *Basic Doxastic Logic*

Let $\mathcal{L}$ be a propositional language. We assume that $\mathcal{L}$ includes the Boolean connectives $\neg$ and $\vee$, and in addition a unary operator B_α. The intuitive gloss of B_α will be "Alpha believes that…". Other connectives may be defined in the standard manner. Being a sentence of $\mathcal{L}$ is characterized as follows.

- If ϕ is an atomic propositional sentence letter, then ϕ is a sentence.
- If ϕ and ψ are sentences, then so is $\phi \vee \psi$.
- If ϕ is a sentence, then so are $\neg\phi$ and $B_\alpha\phi$.
- Nothing else is a sentence.

A *Kripke model* for our language $\mathcal{L}$ is a tuple $M = \langle W, R_\alpha, [\![\cdot]\!]\rangle$. W is a set of points that we'll call *possible worlds*. R_α is a binary relation on W, i.e., $R_\alpha \subseteq W \times W$, that we'll call the *accessibility relation*. And $[\![\cdot]\!]$ is the *interpretation function* mapping propositional letters to sets of possible worlds.

We can think of the accessibility relation R_α as serving to represent the set of worlds that are doxastic possibilities for an agent α relative to some world w. In particular, if w' is such that $\langle w, w'\rangle \in R_\alpha$, then we can think of w' as being a possible world that is left open given all that the agent believes at w.

The truth of a sentence ϕ at a world w in a Kripke model M (for short: $[\![\phi]\!]^w_m = 1$) may be defined as follows.

- If ϕ is a propositional letter, then $[\![\phi]\!]^w_m = 1$ just in case $w \in [\![\phi]\!]$.
- $[\![\neg\phi]\!]^w_m = 1$ just in case $[\![\phi]\!]^w_m \neq 1$.
- $[\![\phi \vee \psi]\!]^w_m = 1$ just in case $[\![\phi]\!]^w_m = 1$ or $[\![\psi]\!]^w_m = 1$.

- $[\![B_\alpha \phi]\!]^w_m = 1$ just in case $[\![\phi]\!]^{w'}_m = 1$, for every w' such that $wR_\alpha w'$.

We let $\vdash_x \phi$ mean that that there is a sequence of formulas $\phi_1, \ldots, \phi$ such that each item in the sequence is either an axiom of the logical system X or follows from items earlier in the sequence by one of the inference rules of X. Then $\vdash_k$ may be characterized as follows.

AXIOMS OF K

(P) Axioms of propositional logic

(K) $B_\alpha(\phi \to \psi) \to (B_\alpha \phi \to B_\alpha \psi)$.

INFERENCE RULES OF K

(MP) $(\vdash_k \phi \ \wedge \ \vdash_k \phi \to \psi) \Rightarrow \vdash_k \psi$.

(N) $\vdash_k \phi \Rightarrow \vdash_k B_\alpha \phi$.

Let $\mathcal{K}$ be the class of Kripke models, and let $\models_{\mathcal{K}} \phi$ mean that, for every Kripke model M, and every $w \in W$, $[\![\phi]\!]^w_m = 1$. Then two basic results in modal logic are:[1]

THEOREM 1. $\vdash_k \phi \Rightarrow \models_{\mathcal{K}} \phi$.

THEOREM 2. $\models_{\mathcal{K}} \phi \Rightarrow \vdash_k \phi$.

THEOREM 1 tells us that $\vdash_k$ is sound with respect to the class of models $\mathcal{K}$, and THEOREM 2 tells us that $\vdash_k$ is complete with respect to this class of models.

The first assumption that we'll make is that the beliefs of an *idealized* doxastic state may be represented by a Kripke model. Given the soundness of $\vdash_k$, it follows from this assumption that:

(B_N) if ϕ is a logical validity, then ϕ is believed by an idealized doxastic agent;

(B_K) an idealized doxastic agent believes all of the logical consequences of her beliefs.

If, in addition, we assume that, for any model $M \in \mathcal{K}$, there is some idealized doxastic state that is represented by M, then the soundness and completeness of $\vdash_k$ entail that (B_N) and (B_K) provide a complete characterization of those properties that are shared by every idealized doxastic state.

We can characterize certain subsets of $\mathcal{K}$ by properties of R_α.

1 For proofs of these results see, e.g., Hughes and Cresswell (1996), Chellas (1980), or Blackburn, de Rijke, and Venema (2001).

DEF. We say that R_α is *serial* just in case, for every $w \in W$, there is some $w' \in W$ such that $wR_\alpha w'$

DEF. We say that R_α is *transitive* just in case, for every $w, w', w'' \in W$, if $wR_\alpha w'$ and $w'R_\alpha w''$, then $wR_\alpha w''$.

DEF. We say that R_α is *Euclidean* just in case, for every $w, w', w'' \in W$, if $wR_\alpha w'$ and $wR_\alpha w''$, then $w'R_\alpha w''$.

We'll let $\mathcal{K}_D$ be the subset of Kripke models whose accessibility relation is serial, $\mathcal{K}_4$ the subset of Kripke models whose accessibility relation is transitive, and $\mathcal{K}_5$ the subset of Kripke models whose accessibility relation is Euclidean.

Assume that $\mathcal{L}$ has only propositional letters p and q. Then we can represent a Kripke model for $\mathcal{L}$ by a diagram like Figure 1. In this model

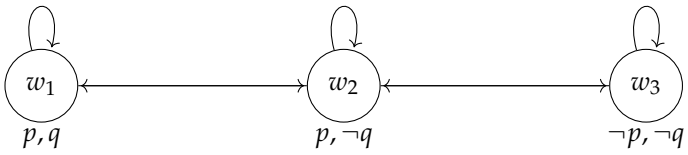


Figure 1: Diagram of a Kripke model

R_α is reflexive and serial, but neither transitive nor Euclidean.

We can further consider the logical systems that arise when one adds certain axioms to K. For example, as additional possible axioms, we have:

(D) $B_\alpha\phi \rightarrow \neg B_\alpha \neg \phi$.

(4) $B_\alpha\phi \rightarrow B_\alpha B_\alpha \phi$.

(5) $\neg B_\alpha \phi \rightarrow B_\alpha \neg B_\alpha \phi$

We'll denote the result of adding some axiom X to K, KX. We have, then, the following soundness and completeness results.

THEOREM 3. $\vdash_{kd} \phi \Leftrightarrow \models_{\mathcal{K}_D} \phi$.

THEOREM 4. $\vdash_{kd4} \phi \Leftrightarrow \models_{\mathcal{K}_D \cap \mathcal{K}_4} \phi$.

THEOREM 5. $\vdash_{kd45} \phi \Leftrightarrow \models_{\mathcal{K}_D \cap \mathcal{K}_4 \cap \mathcal{K}_5} \phi$.

These results tell us the following. First, if we assume that each idealized doxastic state can be modeled by some $M \in \mathcal{K}_D$, then we have:

(B_D) an idealized doxastic agent's beliefs will be consistent.

And, if we further assume that, for each $M \in \mathcal{K}_D$, there is some idealized doxastic state that is represented by M, then it follows that (B_N), (B_K)

and (B_D) provide a complete characterization of those properties that are shared by every idealized doxastic state.

Second, if we assume that, in addition, each idealized doxastic state can be modeled by some $M \in \mathcal{K}_D \cap \mathcal{K}_4$, then we also have:

> (B_4) if an idealized doxastic agent believes ϕ, then she will believe that she believes ϕ.

We'll call this property *positive transparency*.

If we further assume that, for each $M \in \mathcal{K}_D \cap \mathcal{K}_4$, there is some idealized doxastic state that is represented by M, then it follows that (B_N), (B_K), (B_D) and (B_4) provide a complete characterization of those properties that are shared by every idealized doxastic state.

Finally, if we assume that, in addition, each idealized doxastic state can be modeled by some $M \in \mathcal{K}_D \cap \mathcal{K}_4 \cap \mathcal{K}_5$, then we also have:

> (B_5) if an idealized doxastic agent fails to believe ϕ, then she will believe that she fails to believe ϕ.

We'll call this property *negative transparency*.

If we further assume that, for each $M \in \mathcal{K}_D \cap \mathcal{K}_4 \cap \mathcal{K}_5$, there is some idealized doxastic state that is represented by M, then it follows that (B_N), (B_K), (B_D), (B_4), and (B_5) provide a complete characterization of those properties that are shared by every idealized doxastic state.

In the doxastic logic literature, it is typically assumed that every idealized doxastic state can be represented by some element of $\mathcal{K}_D \cap \mathcal{K}_4 \cap \mathcal{K}_5$, and that each element of this set accurately represents some idealized doxastic state. Given this assumption, then, the logic governing the operator B_α is the modal logic KD45.

Note that the following principle is not assumed to hold:

> (T) $\quad B_\alpha\phi \rightarrow \phi$.

That is, we do not assume that our idealized doxastic states are error-free. An ideal belief state, on this picture, need not be one that only includes true beliefs.

> DEF. We say that R_α is *reflexive* just in case, for every $w \in W$, $wR_\alpha w$.

The axiom (T) is guaranteed to hold in any Kripke model whose accessibility relation is reflexive. Importantly, then, we do not assume that a model M representing an idealized doxastic state has a reflexive accessibility relation.

1.2 *Group Beliefs*

So far, our doxastic models have treated the doxastic state of only a single agent. This restriction, however, can be easily relaxed. Instead of a single operator B_α, let $\mathcal{L}$ now contain a series of operators $B_{\alpha_1}, B_{\alpha_2}, \ldots, B_{\alpha_r}$. A Kripke model for $\mathcal{L}$ will be a tuple $\langle W, R_{\alpha_1}, R_{\alpha_2} \ldots, R_{\alpha_r}, [\![\cdot]\!] \rangle$, and truth-at-a-point in such a model will be defined in the obvious way. As with the case of our individual models, we can impose various restrictions for each R_{α_i} such as seriality, transitivity etc. We'll let $\mathcal{K}^\alpha$ be the set of Kripke models for $\mathcal{L}$. We'll let D^α, $\mathcal{K}^\alpha_4$, $\mathcal{K}^\alpha_5$ be, respectively, the set of Kripke models for $\mathcal{L}$ such that each accessibility relation is serial, transitive, and Euclidean.

This type of model allows us to simultaneously represent the doxastic states of multiple agents. Furthermore, it allows us to represent certain collective properties of the doxastic states of groups that cannot be represented by a set of individual Kripke models for each agent in the group.[2] Let's consider how such features may be represented in this sort of model. In particular, we will consider how in such models we can represent the group doxastic properties of *common belief* and *distributed belief*.

1.2.1 *Common Belief*

What is it for ϕ be a matter of common belief amongst a group of agents? The intuitive idea is that common belief is a matter of each agent in the group believing ϕ, and each agent believing that each agent believes ϕ, and each agent believing that each agent believes that each agent believes $\phi \ldots$, and so on, ad infinitum.[3] This sort of group doxastic property can be represented in our models as follows.

> DEF. We will say that w and w' are *n-connected* just in case there is some series of worlds $w_1, \ldots, w_{n+1}$ such that $w = w_1$, $w_{n+1} = w'$ and for each pair $\langle w_i, w_{i+1} \rangle$ there is some R_{α_j} such that $w_i R_{\alpha_j} w_{i+1}$. We'll write $w R_n w'$ to indicate that w and w' are n-connected.

So, for example, the set of 1-connected worlds will just be those pairs of worlds such that there is some j such that $w R_{\alpha_j} w'$, while the set of 2-connected worlds will just be those pairs of worlds that are connected via the belief accessibility relation of at most two agents, etc.

2 See, e.g., Fagin, Halpern, and Vardi (1991), Halpern and Moses (1992), Halpern and Moses (1984), and Halpern, Moses, and Vardi (1995) for important work on the doxastic and epistemic properties of groups.

3 The concepts of common belief and common knowledge and the role that these play in reasoning were introduced in Lewis (1969). See also Aumann (1976) for another influential early treatment of these ideas. See Barwise (1988) for alternative analyses of the notions of common belief and common knowledge.

Let the schematic abbreviation $B_{\alpha^n}\phi$ be inductively characterized as follows:

DEF.

1. $B_{\alpha^1}\phi =_{\text{df}} B_{\alpha_1}\phi \wedge B_{\alpha_2}\phi \wedge \ldots \wedge B_{\alpha_r}\phi$,
2. $B_{\alpha^n}\phi =_{\text{df}} B_{\alpha_1}B_{\alpha^{n-1}}\phi \wedge B_{\alpha_2}B_{\alpha^{n-1}}\phi \wedge \ldots \wedge B_{\alpha_r}B_{\alpha^{n-1}}\phi$.

So B_{α^1} abbreviates the claim that each agent in the group believes ϕ. While B_{α^n} abbreviates the claim that each agent in the group believes that everyone in the group believes ϕ to level $n-1$. As a further definitional abbreviation, we let:

DEF. $M^n_\alpha\phi =_{\text{df}} B_{\alpha^1}\phi \wedge \ldots \wedge B_{\alpha^n}\phi$.

We will read $M^n_\alpha\phi$ as saying that there is *mutual belief of degree n* that ϕ amongst $\alpha_1, \ldots, \alpha_r$.

Given these definitions, it follows that the truth of $B_{\alpha^n}\phi$ and $M_{\alpha^n}\phi$, relative to a point w, in a model M, may be characterized as follows.

$[\![B_{\alpha^n}\phi]\!]^w = 1$ just in case $[\![\phi]\!]^{w'} = 1$, for every w' such that wR_nw'.

$[\![M_{\alpha^n}\phi]\!]^w = 1$ just in case $[\![\phi]\!]^{w'} = 1$, for every w' such that there is some $1 \leq i \leq n$ such that wR_iw'.

So far we haven't added any expressive power to our language. Each operator B_{α^n} and M_{α^n} is merely an abbreviation for some formula already in $\mathcal{L}$. Suppose, however, that we wanted to say that ϕ is a matter of common belief amongst $\alpha_1, \ldots, \alpha_r$. The natural way to express this is to say that, for each n, $M_{\alpha^n}\phi$ holds. Expressing common belief in this way, though, would require quantificational devices or devices of infinite conjunction that our language lacks. This doesn't, however, mean that we can't express the property of common belief in a propositional modal language. To do so, however, we need to add a new operator to our language.

Let $\mathcal{L}$, then, be the language that includes, in addition to each of the operators B_{α_i}, an operator C_α. A Kripke model for our new language $\mathcal{L}$ will still be a tuple $M = \langle W, R_{\alpha_1}, R_{\alpha_2} \ldots, R_{\alpha_r}, [\![\cdot]\!]\rangle$. Given our relations R_{α_i}, we can define the following relation on points in W:

DEF. Let $R_1^+ = \bigcup_{n\geq 1} R_n$.

R_1^+ is the so-called *transitive closure* of R_1.[4] Given this definition, we have that wR_1^+w' just in case there is some n such that wR_nw'. Thus wR_1^+w'

4 This is the smallest transitive relation containing R_1. To see why this is a transitive relation, assume that we have $w_1R_1^+w_2$ and $w_2R_1^+w_3$. Then we have that there is some n such that $w_1R_nw_2$, i.e., w_1 and w_2 are n-connected. And we also have that there is some m such that $w_2R_mw_3$, i.e., w_2 and w_3 are m-connected. But, given this, it follows that we have $w_1R_{n+m}w_3$, i.e., w_1 and w_3 are $(n+m)$-connected.

holds just in case there is some finite length path connecting w and w' via the accessibility relations R_{α_i}.

Note that since R_1^+ is definable in terms of the $R_{\alpha_i} \in M$, we do not need to include this relation in M in order to appeal to it in characterizing the truth of certain sentences in such a model.

We can now characterize the truth of a sentence $C_\alpha\phi$ relative to a world of evaluation in M as follows:

$$[\![C_\alpha\phi]\!]^w = 1 \text{ just in case } [\![\phi]\!]^{w'} = 1, \text{ for every } w' \text{ such that } wR_1^+w'.$$

What are the logical properties governing the common belief operator C_α? We can characterize the logic of this operator as follows. Let K_α be the multi-modal logic characterized by each instance (N), and the relevant instance of (K), for each operator B_{α_i}. (Similarly for KD_α, $KD4_\alpha$ and $KD45_\alpha$.) And let K_α^c (or KD_α^c, $KD4_\alpha^c$ and $KD45_\alpha^c$) be the axiomatic system we get by adding to K_α (or KD_α, $KD4_\alpha$ and $KD45_\alpha$) the following axiom and rule of inference:

(C_1) $\quad C_\alpha\phi \rightarrow M_\alpha^1(\phi \wedge C_\alpha\phi)$.

(R_1) $\quad \vdash_{K_\alpha^c} \phi \rightarrow M_\alpha^1(\psi \wedge \phi) \Rightarrow \vdash_{K_\alpha^c} \phi \rightarrow C_\alpha\psi$.

We, then, have the following soundness and completeness result:[5]

THEOREM 6. $\quad \vdash_{K_\alpha^c} \phi \Leftrightarrow \models_{\mathcal{K}^\alpha} \phi$.

Similar results show that KD_α^c is sound and complete with respect to D^α, that $KD4_\alpha^c$ is sound and complete with respect to $D^\alpha \cap \mathcal{K}_4^\alpha$, and that $KD45_\alpha^c$ is sound and complete with respect to $D^\alpha \cap \mathcal{K}_4^\alpha \cap \mathcal{K}_5^\alpha$. The logic governing C_α, then, is characterized by adding (C_1) and (R_1) to the axioms governing the operators B_{α_i}.

Now it's clear that the structural properties of the accessibility relation R_1^+ will supervene on the structural properties of the accessibility relations R_{α_i}. Importantly, though, the structural properties of R_1^+ may be distinct from those of R_{α_i}. In certain cases, R_1^+ may have additional structural properties to those of R_{α_i}, while in other cases, R_1^+ may lack certain structural properties had by each of the R_{α_i}. Given such discrepancies, then, the logic governing common belief may be distinct from the logic governing individual belief. Let's consider, briefly, some ways in which common belief may inherit some of the logical properties governing individual belief and some ways in which the logic governing common belief may come apart from the logical properties governing individual belief.

First, note that the transitive closure of a serial relation is also serial. If, then, each R_{α_i} is serial, R_1^+ will also be serial. And so, if the logic

5 See Halpern and Moses (1992) for a proof of this. Halpern and Moses (1992), in fact, includes an additional axiom stating $M_\alpha^1\phi \leftrightarrow (B_{\alpha_1}\phi \wedge \ldots \wedge B_{\alpha_1}\phi)$. This, however, is a *definitional* truth and so is not strictly speaking required as an axiom.

governing individual beliefs includes the principle (D), then so will the logic governing common belief. Thus, if the individual agents that we are representing are such that their beliefs are guaranteed to be consistent, then so too will the common beliefs of this group.

We have, then:

(D) $\models_{\mathcal{D}^\alpha} C_\alpha\phi \rightarrow \neg C_\alpha\neg\phi$.

Next, note that, given its definition, R_1^+ is guaranteed to be transitive. In particular, it will satisfy this property whether or not all R_{α_i} do. In this manner, then, R_1^+ may have a structural feature that some R_{α_i} lack. Given that R_1^+ is transitive, we have then:

(4) $\models_{\mathcal{K}^\alpha} C_\alpha\phi \rightarrow C_\alpha C_\alpha\phi$.

Common belief, then, is guaranteed to be positively transparent, even if individual beliefs are not.

Finally, note that while each R_{α_i} being serial entails that R_1^+ is serial, it does *not* follow that if each R_{α_i} is Euclidean then R_1^+ will also be Euclidean.[6] Thus, even if the logic of individual belief entails that individual belief is negatively transparent, it does not follow that common belief must also be negatively transparent.

1.2.2 *Distributed Belief*

What is it for ϕ to be a matter of distributed belief? The intuitive idea is that ϕ is a distributed belief amongst some group of agents just in case ϕ is a consequence of what all of the agents believe.

To express this notion, we'll introduce the operator D_α to our language $\mathcal{L}$. A model for $\mathcal{L}$ will still be a tuple $M = \langle W, R_{\alpha_1}, R_{\alpha_2} \ldots, R_{\alpha_r}, [\![\cdot]\!]\rangle$. We define the following relation amongst the members of W, given our relations R_{α_i}.

DEF. Let wR_dw' just in case for every R_{α_i}, $wR_{\alpha_i}w'$.

Truth-at-a-world for a formula $D_\alpha\phi$, in a model M, can, then, be characterized as follows:

$[\![D_\alpha\phi]\!]^w = 1$ just in case $[\![\phi]\!]^{w'} = 1$, for every w' such that wR_dw'.

We can characterize the logic of this operator as follows. Again let K_α be the multi-modal logic characterized by each instance (N), and the relevant instance of (K), for each operator B_{α_i}. And let K_α^d be the axiomatic system we get by adding to K_α the relevant instance of (K) for the operator D_α, as well an axiom of the following form, for each α_i:

(D_1) $D_\alpha\phi \rightarrow B_{\alpha_i}\phi$.

We, then, have the following soundness and completeness result:[7]

6 See Lismont and Mongin (1994), Colombetti (1993), and Bonanno and Nehring (2000) for proofs and discussion of this result.

7 See Halpern and Moses (1992).

THEOREM 7. $\vdash_{K^d_\alpha} \phi \Leftrightarrow \models_{\mathcal{K}^\alpha} \phi$.

Similarly, if we let $K4_\alpha$ ($K45_\alpha$) be the the multi-modal logic characterized by each instance (N), and the relevant instances of (K) and (4) (and (5)), for each operator B_{α_i}, and let $K4^d_\alpha$ ($K45^d_\alpha$) be the axiomatic system we get by adding to $K4_\alpha$ ($K45_\alpha$) the relevant instances of (K) and (4) (and (5)) for the operator D_α, then we can also show that $K4^d_\alpha$ ($K45_\alpha$) is sound and complete with respect to $\mathcal{K}^\alpha_4$ ($\mathcal{K}^\alpha_4 \cap \mathcal{K}^\alpha_5$).

It's clear that the structural properties of R_d will supervene on the structural properties of the accessibility relations R_{α_i}. While, though, R_d may inherit certain structural properties from R_{α_i}, other structural properties of R_d may be distinct from those of R_{α_i}.

First, note that if each R_{α_i} is transitive, then R_d is transitive. Similarly, if each R_{α_i} is Euclidean, then R_d is Euclidean. It's for this reason that if the logic governing the B_{α_i} includes the principles (4) or (5), so too will the logic governing D_α.

Importantly, however, it doesn't follow from the fact that each R_{α_i} is serial, that R_d is also serial. For it doesn't follow from the fact that, for each w, and each R_{α_i}, there is w' such that $wR_{\alpha_i}w'$, that for each w there is some w' such that wR_dw'. For while, for some w, it may be the case that, for each R_{α_i}, there is some w' such that $wR_{\alpha_i}w'$, this w' need not be the same in each case. Even, then, if the logic governing each B_{α_i} includes the principle (D), it doesn't follow this will be a principle governing D_α. Even, then, if each individual's beliefs are consistent, the distributed beliefs of the group need not be.

1.3 *Some Remarks on the Interpretation of the Formalism*

So far, we've been assuming that our doxastic models represent the doxastic states of certain *idealized* agents. And, as we've noted, there is a standard assumption in the literature that the logic governing such states is KD45. It would, however, be a mistake, I think, to take there to be a substantive question of whether or not the logic governing idealized doxastic states *really* is KD45 or some other logic. Instead, I think it is much more natural to think of these principles as simply *codifying* a certain idealization. On this view, then, there are various types of idealized doxastic states that we might investigate. For example, we might consider those doxastic states that can be represented by some model in $\mathcal{K}$. Doxastic states of this type would be logically omniscient and closed under logical consequence, but perhaps not consistent or perhaps not positively or negatively transparent. Or we might consider those doxastic states that can be represented by some model in $\mathcal{K} \cap \mathcal{D} \cap \mathcal{K}_4$. Doxastic states of this type would be logically omniscient, closed under logical consequence, consistent and

positively transparent, but not negatively transparent. And so on. Now different types of idealized doxastic states may be useful or illuminating for different purposes, but it seems implausible to me that any one of these idealizations stands out as being of significantly greater theoretical importance than all of the others.

One might, however, endorse the bolder hypothesis that our doxastic models are, in fact, intended to represent necessary features of any possible doxastic state. Given this view, the question of whether such models are appropriate and, if so, which constraints should be endorsed, becomes a substantive question. Some have, indeed, argued that the nature of doxastic states makes it the case that they may be represented by models in $\mathcal{K}$.[8] Others have argued that the nature of doxastic states makes it the case that principles such as (B_4) and (B_5) will be satisfied and so models in $\mathcal{K}_4$ or $\mathcal{K}_5$ may serve to represent such states.[9] These claims, however, are quite controversial, and it would take us too far afield now to assess their plausibility.[10] Suffice it to say, there are certain accounts of the nature of doxastic states according to which such states may in fact be accurately represented by the sorts of models we've been considering, while, according to other accounts, certain doxastic states—indeed the types of doxastic states that actual agents tend to have—cannot be accurately represented by the sorts of models that we've been considering.

A third way of interpreting our doxastic models, which should be distinguished from the first interpretation, has it that the role of this class of models is to codify certain general principles that a rational agent's doxastic state *ought* to satisfy. Here we might profitably consider, as an analogous view, the Bayesian account of credal rationality. According to a subjective Bayesian, the class of probability functions defined over some algebra $\mathcal{A}$ represents the class of rationally permissible credal states defined over $\mathcal{A}$. Those features, then, that are common to all such functions represent rational requirements on any credal state defined over such an algebra. Similarly, one might hold that the class of models $\mathcal{K}$ (or the class $\mathcal{K} \cap \mathcal{D} \cap \mathcal{K}_4$ etc.) represent the class of rationally permissible doxastic states. Those features, then, that are common to this class, i.e., the valid formulas

8 See e.g., Stalnaker (1984), Lewis (1974), and Lewis (1999). Both Stalnaker and Lewis, however, recognize that there must be a sense in which agents may have contradictory beliefs or fail to have beliefs that are closed under logical consequence. Lewis (2000) argues that we may think of an agent's doxastic state as consisting of various fragments, where each fragment may be represented by a possible worlds model. An agent, then, may have inconsistent beliefs by having fragments that disagree, and the agent may have beliefs that fail to be logically closed by believing, say, ϕ relative to one fragment and $\phi \rightarrow \psi$ relative to another, but not believing ψ relative to any fragment.

9 See, for example, Shoemaker (1996a, 1996b) for arguments that, at least in certain cases, positive introspection should hold as a constitutive matter.

10 The arguments in Williamson (2000), for example, put serious pressure on the idea that the transparency principles (B_4) and (B_5) will hold for actual agents.

given the class of models, represent, on this view, rational requirements on any doxastic state.

Now this view may seem to be a mere notational variant on the first interpretation. However, concluding this would, I think, be a mistake. What an actual agent ought to believe and what an idealized agent would believe are not the same thing. Here's a somewhat facile, but I think sufficiently instructive example that illustrates this point. An idealized agent would, plausibly, believe that they are idealized. However, a rational agent, who is not an idealized agent, should not be rationally required to believe that they are idealized. Thus, a representation of what an idealized doxastic state would look like is not, thereby, a representation of what doxastic features a rational agent ought to have.

The view that our doxastic models serve to codify rational requirements on doxastic states, again, makes it a substantive question which class of Kripke models, if any, we should take as the appropriate class for formulating our doxastic logic. One may, for example, maintain that doxastic states ought to be such that they're consistent and closed under logical entailment, but deny that doxastic states ought to be transparent on pain of irrationality. Once again, the issues here are subtle and we will simply content ourselves with flagging the issues, without making any attempt to resolve them.

Having noted these three possible roles that our doxastic models may play, it is worth highlighting that different classes of models might, in fact, play different roles. So, for example, one might maintain that the class $\mathcal{K}$ is the smallest class that serves to characterize how a rational agent's doxastic state ought to be. But one might still find it profitable to investigate what features are exhibited by the sorts of idealized doxastic states characterized by, say, $\mathcal{K} \cap \mathcal{D} \cap \mathcal{K}_4$.

In what follows, I will often continue to speak as if the Kripke models, as well as other models we'll introduce, are meant to represent idealized doxastic states. However, in certain cases a normative or descriptive interpretation may seem more natural and so I will sometimes talk as if the models in question are meant to describe such facts. In each case, though, it is worth bearing in mind the alternative interpretations that are available.

2 BELIEF REVISION

So far we've seen how to represent certain features of idealized doxastic states. In particular, we've seen how to represent the beliefs of an idealized doxastic agent, including beliefs about that agent's beliefs, as well as beliefs about other agents' beliefs. The models that we've looked at, however, say nothing about how idealized doxastic states should change given new

information. In this section we'll look at an influential account of belief revision called AGM.[11]

The basic theory of AGM consists of a set of formal postulates that serve to codify rational constraints on belief revision, expansion and contraction. In addition to such formal postulates, however, various authors have provided models of how functions meeting these constraints may be determined. We'll begin, in Section 2.1, by considering the basic postulates of AGM. In Section 2.2, we'll then consider some possible connections between rational belief revision, expansion and contraction. In Section 2.3, we'll consider some models for belief revision and contraction. And, finally, in Section 2.4 we'll look at some additional postulates that have been proposed to handle the phenomenon of iterated belief revision.

2.1 *AGM: The Basic Postulates*

Let $\mathcal{L}$ be a set of sentences closed under the standard Boolean operators, and let $\Gamma \subseteq \mathcal{L}$ be a set of such sentences. We'll denote by $Cl(\Gamma)$ the logical closure of Γ. In standard presentations of this theory, it is assumed that rational agents have belief states that can be (at least partially) modeled by logically closed sets of sentences. In this section, we will follow this practice as well. Note that this marks a departure from our treatments of belief states in the previous section, where such states were modeled as sets of possible worlds. We'll let B be a possible belief set, i.e, a set of sentences such that $B = Cl(B)$. We denote the set of belief sets $\mathcal{B}$.

We'll first consider the AGM postulates governing rational belief expansion. We let $+ : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be a function mapping pairs of belief sets and sentences to belief sets. We'll let B^+_ϕ be the result of applying the function $+$ to the belief set B and sentence ϕ. According to AGM, rational belief expansions must satisfy the following constraints.

(B^+_1) $\phi \in B^+_\phi$.

(B^+_2) $B \subseteq B^+_\phi$.

(B^+_3) If $\phi \in B$, then $B = B^+_\phi$.

(B^+_4) If $A \subseteq B$, then $A^+_\phi \subseteq B^+_\phi$.

(B^+_5) For any operation # satisfying B^+_1–B^+_4, $B^+_\phi \subseteq B^\#_\phi$.

We can think of expansion as an operation that increases the agents belief set B to accommodate belief in ϕ. Then (B^+_1) tells us that, given this

11 This account developed out of Alchourrón et al. (1985). For a comprehensive survey see Gärdenfors (1988). See also Huber (2016) for a helpful treatment of this and other related material.

operation, ϕ will be a part of the resultant belief set. While (B_2^+) tells us that everything that was believed prior to the operation is believed after the operation. Also (B_3^+) tells us that, if ϕ is already believed, given B, then the expansion operation is trivial. Additionally (B_4^+) tells us that the expansion operation is monotone, i.e., that it preserves the subset relation. And, finally, (B_5^+) tells us that, in a specific sense, expansion is the most conservative operation with the preceding characteristics.

It can be shown that (B_1^+)–(B_5^+) uniquely pin down the expansion operator. Thus:

> THEOREM 8. A function $+$ satisfies (B_1^+)–(B_5^+) just in case $B_\phi^+ = Cl(B \cup \{\phi\})$.[12]

Next, we consider belief revision. We let $* : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be a function mapping pairs of belief sets and sentences to belief sets. According to AGM, rational belief revisions must satisfy the following constraints.

(B_1^*) $\phi \in B_\phi^*$.

(B_2^*) $B_\phi^* \subseteq B_\phi^+$.

(B_3^*) If $\neg\phi \notin B$, then $B_\phi^+ \subseteq B_\phi^*$.

(B_4^*) $B_\phi^* = \mathcal{L}$ just in case $\models \neg\phi$.

(B_5^*) If $\models \phi \leftrightarrow \psi$, then $B_\phi^* = B_\psi^*$.

(B_6^*) $B_{\phi\wedge\psi}^* \subseteq (B_\phi^*)_\psi^+$.

(B_7^*) If $\neg\psi \notin B_\phi^*$, then $(B_\phi^*)_\psi^+ \subseteq B_{\phi\wedge\psi}^*$.

Like expansion, we can think of revision as an operation that changes an agent's belief set B to accommodate belief in ϕ. Unlike expansion, though, in belief revision certain beliefs may be discarded to accommodate ϕ.

Constraint (B_1^*) tells us that, given this operation, ϕ will be a part of the resultant belief set. While (B_2^*) tells us that everything that is believed given belief revision will be believed given belief expansion. Also (B_3^*) tells us that the reverse is also true, and so expansion and revision deliver the same output, when ϕ is logically compatible with B. And (B_4^*) tell us that that the result of belief revision will be a consistent belief set just in case ϕ is itself logically consistent. Constraint (B_5^*) tells us that logically equivalent sentences induce the same revision operation on a belief set. And (B_6^*) tells us that everything that is believed after revising a belief set given a conjunction $\phi \wedge \psi$, will be believed after first revising the same belief set given ϕ and then expanding the resultant belief set given ψ.

12 See Gärdenfors (1988).

Finally, (B_7^*) tell us that the reverse is true, and so revising given ϕ and then expanding given ψ delivers the same output as revising given $\phi \wedge \psi$, when ψ is consistent with the result of revision given ϕ.

Unlike with the constraints on $+$, these constraints do *not* suffice to uniquely determine the function $*$. Instead, there is a non-empty, non-singleton, set of functions that satisfy (B_1^*)–(B_7^*).

Finally, we consider belief contraction. We let $- : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be a function mapping pairs of belief sets and sentences to belief sets. According to AGM, rational belief contractions must satisfy the following constraints.

(B_1^-) $B_\phi^- \subseteq B$.

(B_2^-) If $\phi \notin B$, then $B_\phi^- = B$.

(B_3^-) If $\not\models \phi$, then $\phi \notin B_\phi^-$.

(B_4^-) If $\phi \in B$, then $B \subseteq (B_\phi^-)_\phi^+$.

(B_5^-) If $\models \phi \leftrightarrow \psi$, then $B_\phi^- = B_\psi^-$.

(B_6^-) $B_\phi^- \cap B_\psi^- \subseteq B_{\phi \wedge \psi}^-$.

(B_7^-) If $\phi \notin B_{\phi \wedge \psi}^-$, then $B_{\phi \wedge \psi}^- \subseteq B_\phi^-$.

We can think of contraction as an operation that changes an agent's belief set B to accommodate the removal of a belief ϕ.

Constraint (B_1^-) tells us that belief contraction does not introduce any new beliefs. While (B_2^-) tells us that if the agent does not already believe ϕ, then the result of contracting by ϕ leaves the agent's belief set unchanged. Then (B_3^-) tells us that if ϕ is not a logical truth, then ϕ will not be in any belief set that is contracted by ϕ. And (B_4^-) tells us that if a belief set B contains ϕ, then the result of contracting this belief set by ϕ and then expanding the resulting set by ϕ will contain everything that is in B. Next (B_5^-) tells us that logically equivalent sentences induce the same contraction operation on a belief set. And (B_6^-) tells us that every belief that remains when a belief set is contracted by ϕ and by ψ will remain when the belief set is contracted by $\phi \wedge \psi$. Finally, (B_7^-) tells us that if ϕ is not in the belief set that results from contracting a belief set B by $\phi \wedge \psi$, then everything that results from contracting B by $\phi \wedge \psi$ will be in the set that result from contracting B by ϕ.

As with the constraints on $*$, these constraints do not uniquely determine the function $-$. Again, there is a non-empty, non-singleton set of functions that satisfy each of (B_1^-)–(B_7^-).

2.2 *Relations Between Operations*

Given these rational constraints on contraction, revision and expansion, it is natural to ask what connections there might be between these three operations. In this section, we'll consider some possible options.

So far we've been talking as if agents adopt *distinct* policies of rational belief revision, contraction and expansion. Following Levi (1977), however, one might maintain that agents only really adopt policies of contraction and expansion, and that a policy of revision is determined by the latter two policies in the following manner.

> CONSTITUTIVE LEVI IDENTITY. $B^*_\phi =_{\text{df}} (B^-_{\neg\phi})^+_\phi$.

If the claim that the adoption of a rational revision policy simply consists in the adoption of rational contraction and expansion policies is to be at all plausible, it must be the case that, given the putative analysis, it is ensured that the resulting revision policy will indeed be rational given that the contraction and expansion policies are. The following result shows that, given the CONSTITUTIVE LEVI IDENTITY, this is so.

> THEOREM 9. If $-$ satisfies (B^-_1)–(B^-_3) and (B^-_5)–(B^-_7), and $+$ satisfies (B^+_1)–(B^+_5), then, given the CONSTITUTIVE LEVI IDENTITY, $*$ satisfies (B^*_1)–(B^*_7).[13]

Now even if one wants to reject the claim that the adoption of a belief revision policy simply consists in the adoption of expansion and contraction policies, one should, I think, nonetheless hold that there are important rational constraints governing which policies of belief revision, expansion and contraction an agent may simultaneously adopt. In particular, whether one thinks that rational belief revision should be *analyzed* in terms of rational belief contraction and expansion, one should, I think, endorse the following normative constraint.

> NORMATIVE LEVI IDENTITY. If $*$ is an agent's revision policy, $-$ her contraction policy, and $+$ her expansion policy, then the agent ought to be such that $B^*_\phi = (B^-_{\neg\phi})^+_\phi$.

Two points are worth mentioning here.

First, the NORMATIVE LEVI IDENTITY does, indeed, impose a substantive constraint in addition to those imposed by (B^+_1)–(B^+_5), (B^*_1)–(B^*_7), and (B^-_1)–(B^-_7). For there are functions $+$, $*$ and $-$ that satisfy (B^+_1)–(B^+_5), (B^*_1)–(B^*_7), and (B^-_1)–(B^-_7), respectively, but that fail to jointly satisfy the condition that $B^*_\phi = (B^-_{\neg\phi})^+_\phi$.[14]

13 See Gärdenfors (1988), ch. 3.6.

14 This follows from the fact that there is a unique function $+$ satisfying conditions (B^+_1)–(B^+_5), together with the fact that there are multiple functions $*$ and $-$ satisfying (B^*_1)–(B^*_7), and (B^-_1)–(B^-_7) respectively.

Second, THEOREM 9 guarantees that the constraints imposed by the NORMATIVE LEVI IDENTITY are, indeed, consistent with the constraints imposed by (B_1^+)–(B_5^+), (B_1^-)–(B_7^-), and (B_1^*)–(B_7^*). For there are functions $-$ and $+$ that satisfy the constraints imposed by (B_1^-)–(B_7^-), and (B_1^+)–(B_5^+), and it follows from this fact, together with THEOREM 9 that, given such functions, there is a function $*$ that satisfies the constraints imposed by (B_1^*)–(B_7^*), and, in addition, is such that $B_\phi^* = (B_{\neg\phi}^-)_\phi^+$.

Another option, following Harper (1976), is to maintain that agents only really adopt policies of revision and expansion. In particular, one may maintain that an agent's policy of contraction is determined by her policy of revision as follows.

> CONSTITUTIVE HARPER IDENTITY. $B_\phi^- =_{\text{df}} B \cap B_{\neg\phi}^*$.

If the claim that the adoption of a rational contraction policy simply consists in the adoption of a rational revision policy is to be at all plausible, it must the case that, given the putative analysis, it is ensured that the resulting contraction policy will indeed be rational given that the revision policy is. The following result shows that, given the CONSTITUTIVE HARPER IDENTITY, this is so.

> THEOREM 10. If $*$ satisfies (B_1^*)–(B_7^*), then, given the CONSTITUTIVE HARPER IDENTITY, $-$ satisfies (B_1^-)–(B_7^-).[15]

Now, again, even if one wants to reject the claim that the adoption of a belief contraction policy simply consists in the adoption of a revision policy, one should, I think, still hold that there are important rational constraints governing which contraction and revisions policies an agent may simultaneously adopt. In particular, whether one thinks that rational belief contraction should be analyzed in terms of rational belief revision, one should, I think, endorse the following normative constraint:

> NORMATIVE HARPER IDENTITY. If $*$ is an agent's belief revision policy, and $-$ her belief contraction policy, then the agent ought to be such that $B_\phi^- = B \cap B_{\neg\phi}^*$.

Two points are, again, worth mentioning here.

First, the NORMATIVE HARPER IDENTITY provides a substantive constraint in addition to (B_1^*)–(B_7^*) and (B_1^-)–(B_7^-). For there are functions $*$ and $-$ that satisfy the latter constraints but for which the identity $B_\phi^- = B \cap B_{\neg\phi}^*$ fails to hold.[16]

Second, THEOREM 10 guarantees that the constraint imposed by the NORMATIVE HARPER IDENTITY is consistent with the constraints imposed

15 See Gärdenfors (1988) ch. 3.6.

16 This follows from the fact that there are multiple functions $*$ and $-$ satisfying (B_1^*)–(B_7^*), and (B_1^-)–(B_7^-) respectively.

by (B_1^*)–(B_7^*) and (B_1^-)–(B_7^-). For there is some function $*$ satisfying (B_1^*)–(B_7^*), and so, given THEOREM 10, it follows that there is some other function $-$ satisfying (B_1^-)–(B_7^-), such that $B_\phi^- = B \cap B_{\neg\phi}^*$.

2.3 *AGM: Models*

Constraints (B_1^-)–(B_7^-) determine a class of rational contraction functions, while (B_1^*)–(B_7^*) determine a class of rational revision functions. There are, however, other ways of characterizing the classes determined by (B_1^-)–(B_7^-) and (B_1^*)–(B_7^*). In particular, we can provide models of possible features of an agent's doxastic state that might serve to determine which rational revision or contraction policy she adopts, and we can show that, given certain rational constraints on such features, the classes of rationally permissible revision or contraction functions are the same classes as those determined by (B_1^-)–(B_7^-) and (B_1^*)–(B_7^*).

2.3.1 *Sphere Systems*

We first consider a model of how an agent's doxastic state might serve to determine a rational revision policy.[17] In rough outline, we may think of an agent's doxastic state, in addition to determining a belief set, as also determining, for each possible belief set B, an ordering of plausibility amongst various maximal consistent descriptions of how the world might be. Given a doxastic state with such structure, we may think of the agent as adopting a revision policy such that, given a belief set B and a sentence ϕ, the resulting revised belief set is just the intersection of the most plausible maximal consistent descriptions of how the world might be that contain ϕ.

A little more pedantically: Let $\mathcal{W}$ be the set of maximal consistent subsets of $\mathcal{L}$. Following the literature, we'll refer to these as *possible worlds*. It is, however, worth keeping in mind that, as sets of of sentences, these differ from the possible worlds considered in the previous section. For any belief set B, we let $[\![B]\!] = \{w \in \mathcal{W} : B \subseteq w\}$. And for any $P \subseteq \mathcal{W}$, we let $B_P = \cap\{w : w \in P\}$.

Let $\mathbf{S}$ be a set of subsets of $\mathcal{W}$. We call $\mathbf{S}$ a *system of spheres centered on* B just in case $\mathbf{S}$ satisfies the following conditions.

> ORDERING. For every $S, S' \in \mathbf{S}$, either $S \subseteq S'$ or $S' \subseteq S$.
>
> CENTERING. $[\![B]\!] \in \mathbf{S}$, and for every $S \in \mathbf{S}$, $[\![B]\!] \subseteq S$.
>
> UNIVERSALITY. $\mathcal{W} \in \mathbf{S}$.
>
> LIMIT ASSUMPTION. Let ϕ be a sentence. If there is some $S \in \mathbf{S}$ such that $S \cap [\![\phi]\!] \neq \emptyset$, then there is some $S \in \mathbf{S}$ such that (i)

17 See Grove (1988) for the initial development of this model. The model is, in certain respects, notably similar to the semantic theory for counterfactuals developed in Lewis (1973).

$S \cap [\![\phi]\!] \neq \emptyset$, and (ii) for every $S' \in \mathbf{S}$ such that $S' \cap [\![\phi]\!] \neq \emptyset$, $S \subseteq S'$.

ORDERING tells us that the members of $\mathbf{S}$ can be totally ordered by the subset relation. CENTERING tells us that the set of worlds that are compatible with the belief set B is a member of $\mathbf{S}$ that is minimal with respect to the subset ordering on $\mathbf{S}$. UNIVERSALITY tells us that the set of all worlds is itself a member of $\mathbf{S}$. And, finally, the LIMIT ASSUMPTION tells us that for any sentence ϕ if the set of $S \in \mathbf{S}$ such that ϕ is true at some world in S is non-empty, then this set has a least element relative to the subset ordering on $\mathbf{S}$.

If $\not\models \neg\phi$, then we let S_ϕ be the smallest sphere intersecting ϕ. Given UNIVERSALITY and the LIMIT ASSUMPTION, such a sphere is guaranteed to exist. And if $\models \neg\phi$, then we let $S_\phi = \mathcal{W}$. We let $C_S(\phi) = [\![\phi]\!] \cap S_\phi$.

Let $\mathcal{S}$ be a function that maps each $B \in \mathcal{B}$ to a system of spheres centered on B. Call this a *sphere function*. We denote the sphere system determined by a sphere function $\mathcal{S}$, for some belief set B, by $\mathcal{S}_B$. Finally, let $f : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be a function such that there is some sphere function $\mathcal{S}$ such that for every $B \in \mathcal{B}$ and $\phi \in \mathcal{L}$, $f(B, \phi) = \cap C_{\mathcal{S}_B}(\phi)$. We call this a *sphere revision function*.

We can think of a sphere revision function, determined by some sphere function $\mathcal{S}$, as mapping a belief state B and a sentence ϕ to the belief state that is determined by the worlds at which ϕ holds that, according to $\mathcal{S}$, are closest to B.

More precisely, we can think of a sphere revision function, determined by some sphere function $\mathcal{S}$, as operating in the following manner. Given a belief set B and a sentence ϕ, we first look at the sphere system centered on B determined by $\mathcal{S}$. Next we find the smallest sphere that is compatible with ϕ, and consider the set of worlds within this sphere at which ϕ holds. The sphere revision function, then, returns as a belief set the set of sentences that are true at every world within this set.

The following theorem shows that the adoption of a rational revision function can always be modeled in terms of the adoption of a sphere revision function determined by some sphere function $\mathcal{S}$, and that, conversely, the adoption of a sphere revision function determined by some sphere function $\mathcal{S}$ will always correspond to the adoption of a rational revision function.

THEOREM 11. Function f is a sphere revision function just in case f satisfies (B_1^*)–(B_7^*).[18]

We can support the claim that a system of spheres may be thought of as encoding a relation of doxastic plausibility as follows. Call a relation $\leq$ over $\mathcal{W}$ with the following properties, a *B-plausibility ordering*.

18 See Grove (1988).

CONNECTIVITY. For every $w, w' \in \mathcal{W}$, either $w \leq w'$ or $w' \leq w$.

TRANSITIVITY. If $w \leq w'$ and $w' \leq w''$, then $w \leq w''$.

ϕ-MINIMALITY. If $[\![\phi]\!] \neq \emptyset$, then $\{w \in [\![\phi]\!] : w \leq w' \text{ for all } w' \in [\![\phi]\!]\} \neq \emptyset$.

B-MINIMALITY. $w \leq w'$ for all $w' \in \mathcal{W}$ just in case $w \in [\![B]\!]$.

Given a B-plausibility ordering $\leq$, let $S_w = \{w' \in \mathcal{W} : w' \leq w\}$. We let $\mathcal{S}_{\leq} = \{S_w : w \in \mathcal{W}\}$. It can be shown that:

THEOREM 12. For any B-plausibility ordering $\leq$, $\mathcal{S}_{\leq}$ is a system of spheres centered on B, and for any system of spheres $\mathcal{S}$ centered on B, there is a unique B-plausibility ordering $\leq$, such that $\mathcal{S} = \mathcal{S}_{\leq}$.[19]

A system of spheres, thus, encodes an ordering over possible worlds, and it is this plausibility ordering, that, according to this model, serves to determine how a rational agent will revise her belief set B given some sentence ϕ.

Given THEOREM 10, we can also use this model to provide a model of how an agent might adopt a rational contraction function. Given the adoption of a sphere revision function determined by some $\mathcal{S}$, the agent would adopt a policy of contracting a belief set B, given some sentence ϕ, so that her new belief set is given by the set of sentences that are true in all and only the worlds in $C_S(\neg\phi) \cup [\![B]\!]$. That is, such an agent will contract her belief set by ϕ by adding to the set of worlds representing this set, the most plausible $\neg\phi$ worlds. Such a policy will be guaranteed to satisfy (B_1^-)–(B_7^-).

2.3.2 *Epistemic Entrenchment*

Next, we consider a model of how an agent's doxastic state might serve to determine a rational belief contraction policy.[20] In rough outline, we may think of an agent's doxastic state, in addition to determining a belief set, as also determining, for each belief set B, a binary relation on the set of sentences of $\mathcal{L}$ that encodes information about how epistemically entrenched such sentences are, given the belief set B. While the notion of epistemic entrenchment is best thought of as being functionally defined via its role in the following account of belief contraction, one can think of an epistemic entrenchment ordering, roughly, as corresponding to an ordering representing how committed an agent is to retaining certain beliefs, given that they have a belief set B.

19 See Gärdenfors (1988).
20 See Gärdenfors and Makinson (1988) for this type of model.

Given a doxastic state that determines an entrenchment ordering for each belief set B, we may think of the agent as adopting a contraction policy such that, given a belief set B and a sentence ϕ, the agent restricts B to the subset of elements ψ such that either ψ is a theorem or $\psi \vee \phi$ is more epistemically entrenched than ϕ.

A little more pedantically: let $\leq$ be a binary relation over $\mathcal{L}$. We let $\phi < \psi =_{\text{df}} \phi \leq \psi \wedge \psi \not\leq \psi$. We call $\leq$ a *B-entrenchment relation* just in case it satisfies the following postulates.

E1. If $\phi \leq \psi$ and $\psi \leq \xi$, then $\phi \leq \xi$.

E2. If $\phi \models \psi$, then $\phi \leq \psi$.

E3. For all ϕ, ψ, either $\phi \leq \phi \wedge \psi$ or $\psi \leq \phi \wedge \psi$.

E4. If $B \neq \mathcal{L}$, then $\phi \in B$ just in case $\phi \leq \psi$, for all ψ.

E5. If $\psi \leq \phi$, for all ψ, then $\models \phi$.

Let $\preceq$: $\mathcal{B} \to \mathcal{P}(\mathcal{L} \times \mathcal{L})$ be a function that maps each $B \in \mathcal{B}$ to a binary relation over $\mathcal{L}$. We denote each such relation by $\preceq_B$. If each $\preceq_B$ is a B-entrenchment relation, we'll call $\preceq$ an *entrenchment function*. Let $C_{\preceq}$: $\mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be a function such that $C_{\leq}(B, \phi) = \{\psi : \psi \in B$ and either $\phi < \phi \vee \psi$ or $\models \psi\}$. Let $f : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be a function such that there is some entrenchment function $\preceq$ such that for every $B \in \mathcal{B}$ and $\phi \in \mathcal{L}$, $f(B, \phi) = C_{\leq}(B, \phi)$. We call this an *entrenchment contraction function*.

The following theorem shows that the adoption of a rational contraction function can always be modeled in terms of the adoption of an entrenchment contraction function determined by some entrenchment function $\preceq$, and that, conversely, the adoption of an entrenchment contraction function determined by some some entrenchment function $\preceq$ will always correspond to the adoption of a rational contraction function.

THEOREM 13. Function f is an entrenchment contraction function just in case f satisfies (B_1^-)–(B_7^-).[21]

Given THEOREM 9, we can also use this model to provide a model of how an agent might adopt a rational revision function. Given the adoption of an entrenchment contraction function determined by some $\preceq$, the agent would adopt the policy of revising a belief set B given ϕ, by first contracting B to the subset of elements $\psi \in B$ such that either ψ is a theorem or $\psi \vee \neg\phi$ is more epistemically entrenched that $\neg\phi$, and then expanding the resulting set by ϕ. Such a policy will be guaranteed to satisfy (B_1^*)–(B_7^*).

21 See Gärdenfors and Makinson (1988).

2.4 *Iterated Belief Revision*

If an agent has adopted a revision policy $*$ satisfying (B_1^*)–(B_7^*), then not only is it determined how the agent should revise her current belief set B, given some information ϕ, but it is also determined how the agent should revise this new belief set, given additional information ψ. For a revision policy satisfying (B_1^*)–(B_7^*) determines how any belief set should be revised given any piece of information. It has, however, been suggested that the AGM postulates provide implausible results when we consider which patterns of iterated belief revision they count as rationally permissible and rationally mandated.[22] In response to these putative problems, various emendations of, or additions to, the AGM postulates have been suggested. In this section, we'll consider some putative problems with iterated belief revision that arise for AGM and look at a few solutions that have been suggested.

2.4.1 *Problems with Iterated Belief Revision in AGM*

There are two types of problems that the AGM revision postulates have been thought to have with iterated belief revision. On the one hand, the AGM revision postulates have been thought to be too permissive, vindicating as rational certain patterns of iterated belief revision that would seem to be irrational. On the other hand, the AGM revision postulates have been thought to be too restrictive, ruling out as irrational certain patterns of iterated belief revision that would seem to be rational. Let me say a bit more about each of these worries in turn.

The first problem stems from the fact that (B_1^*)–(B_7^*) put very few constraints on iterated belief revision. To see this, let b be some particular belief set. We'll, then, let (b_1^*)–(b_7^*) be the postulates that result from (B_1^*)–(B_7^*) by saturating the variable B ranging over elements of $\mathcal{B}$ with the particular element $b \in \mathcal{B}$. Then (b_1^*)–(b_7^*) provide constraints on a function $b* : \mathcal{L} \to \mathcal{B}$ mapping sentences to belief sets. In particular, they provide constraints on functions that tell us how the belief set b should be revised given new information. Call such a function a *b-revision function*.

> DEF. For any function $f : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$, let f^b be the function such that $< l_x, b_y > \in f^b \leftrightarrow < b, l_x, b_y > \in f$.
>
> DEF. For any function $f : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$, and any function $g : \mathcal{L} \to \mathcal{B}$, let $f/g_b : \mathcal{B} \times \mathcal{L} \to \mathcal{B}$ be the function such that if $b_x \neq b$, then $< b_x, l_y, b_z > \in f \leftrightarrow < b_x, l_y, b_z > \in f/g_b$, while $< b, l_y, b_z > \in f/g_b \leftrightarrow < l_y, b_z > \in g$.

22 See, for example, Boutilier (1996), Darwiche and Pearl (1997), and Stalnaker (2009).

We can think of f^b as the b-revision function determined by the revision function f. And we can think of f/g_b as the revision function that results from swapping g_b for the b-revision function determined by f. We, then, have the following results.

- For any f satisfying (B_1^*)–(B_7^*) and any $b \in \mathcal{B}$, f^b will satisfy (b_1^*)–(b_7^*).
- For any g satisfying (b_1^*)–(b_7^*), and any f satisfying (B_1^*)–(B_7^*), f/g_b will also satisfy (B_1^*)–(B_7^*).

What this shows is that functions satisfying (B_1^*)–(B_7^*) can be thought of as the result of freely choosing, for each $b \in \mathcal{B}$, some function satisfying (b_1^*)–(b_7^*). Thus (B_1^*)–(B_7^*) allow us to mix-and-match b-revision functions as we like.

This degree of freedom, however, has problematic consequences. For a b-revision function $f : \mathcal{L} \to \mathcal{B}$ can be seen as encoding *conditional beliefs*. We'll say that the conditional belief $\phi|\psi$ is accepted by f just in case $\phi \in f(\psi)$. The fact that (B_1^*)–(B_7^*) allow for arbitrary mixing and matching of b-revision functions, shows that (B_1^*)–(B_7^*) impose almost no constraints on which conditional beliefs an agent should maintain or give up when changing her belief set in light of some new information.

Here's an example that illustrates the problem.[23]

> Flying Bird. You initially believe of some animal in the distance that it's neither a bird, $\neg B$, nor can it fly, $\neg F$. You, however, have the conditional belief that it can fly, given that it's a bird, $F|B$. That is, you are disposed to come to believe that the animal can fly, were you to learn that it's a bird. Now you learn that the animal can indeed fly, and as a result you give up the conditional belief $F|B$, and form the conditional belief $\neg F|B$.

This sort of transition will seem to many to be irrational. Learning that the consequent of a conditional belief is true would seem to provide no evidence against that conditional belief. However, this transition will be sanctioned as rationally permissible given (B_1^*)–(B_7^*). Examples such as this have convinced a number of authors that further constraints, in addition to (B_1^*)–(B_7^*), are required to adequately constrain rational revision functions.

To see why one might think that (B_1^*)–(B_7^*) are not only too permissive but also too restrictive, note that if an agent adopts a revision policy satisfying (B_1^*)–(B_7^*), then which belief set she should have, given some information ϕ, is a function of her current belief set. This, however, has the following consequence. Given the adoption of a revision policy satisfying (B_1^*)–(B_7^*), an agent who starts out with a belief set B and who then receives

23 This type of example and others may be found in Darwiche and Pearl (1997).

a series of information $\psi_1, \psi_2, \ldots, \psi_3$ that she uses to successively revise her beliefs and who, as a result, winds up, again, with belief set B, is rationally required to revise this belief set given some information ϕ in exactly the same manner as she would have revised this belief set, given ϕ, prior to receiving the series of information $\psi_1, \psi_2, \ldots, \psi_3$. One might, however, think that a series of information that would ultimately leave an agent's belief set unchanged could rationally lead to a change in the agent's conditional beliefs.[24] This sort of change, however, is ruled out as irrational by the AGM postulates.

2.4.2 *Iterated Belief Revision Functions*

We'll first consider an amendment to the AGM account of rational revision that is meant to address the first problem. We'll, then, consider an alternative amendment that addresses both the first and the second problem.

In response to worries about the excessive permissiveness of AGM, Boutilier (1996) proposes a much more restricted account of rational belief revision. Perhaps the simplest way to present the account is by appeal to Grove's model in which a revision policy is represented by a set of total pre-orders over the space of possible worlds.

Let $\preceq$ be a function mapping each B to a B-plausibility ordering $\preceq_B$. As we noted earlier, each function satisfying (B_1^*)–(B_7^*) may be represented by some such function $\preceq$. For each B, let $C(B, \phi) = \{w : \phi \in w \text{ and } w \preceq_B w', \text{ for each } w' \text{ such that } \phi \in w'\}$. Boutilier (1996) suggests that in order for $\preceq$ to represent a rational revision function, in addition to each $\preceq_B$ being a B-plausibility ordering, it must also satisfy the following.

> Boutilier's Constraint. $\preceq_{C(B,\phi)}$ must be such that for all $w, w' \notin C(B, \phi)$, $w \preceq_{C(B,\phi)} w'$ just in case $w \preceq_B w'$.

The idea here is that a rational agent, in revising her beliefs in response to some information ϕ, should adjust her plausibility ordering over worlds in such a way that the most plausible ϕ-worlds are ranked highest in her new plausibility ordering but otherwise leaves the plausibility ordering amongst worlds untouched. An agent who adjusts her belief state in this manner will effectively make the minimal adjustments to her conditional beliefs as is necessary in order to accommodate ϕ.

It is easy enough to see that the constraints that Boutilier (1996) suggests rule out as irrational the problematic case of revision in Flying Bird. It has been argued, however, that Boutilier's account of belief revision demands that too many conditional beliefs be preserved, and that this has undesirable consequences about when an agent may be rationally required

24 For worries in this vicinity see Levi (1988) and Darwiche and Pearl (1997).

to give up certain beliefs. Darwiche and Pearl (1997) give the following example.

> SEQUENTIAL RED BIRD. You are initially uncertain about whether a certain animal is a red, R, or is a bird, B. You then get information that the animal is a bird, B. Then you get information, from a different source, that the animal is red, R. However, further consultation with an expert indicates that, in fact, the first piece of evidence was wrong and, in fact, the animal is not a bird, $\neg B$. As a result, you wind up believing that the animal is not a bird, $\neg B$, but that the animal is red, R.

Intuitively, this process of revision would seem to be perfectly rational. The constraints on revision proposed by Boutilier (1996), however, deem this process of revision irrational. To see this, consider the following model. Let $w_1 = R \wedge B$, $w_2 = \neg R \wedge B$, $w_3 = R \wedge \neg B$ and $w_4 = \neg R \wedge \neg B$. At t_1, your plausibility ordering is such that:

$$(t_1) \quad w_1 = w_2 = w_3 = w_4.$$

Upon getting the information B, at t_2 you minimally adjust your ordering, in accord with the Boutillier model, so that:

$$(t_2) \quad w_1 = w_2 < w_3 = w_4.$$

Then, at t_3, upon getting the information R, you again minimally adjust your ordering so that:

$$(t_3) \quad w_1 < w_2 < w_3 = w_4.$$

Finally, upon getting the information $\neg B$, you once again minimally adjust your plausibility ordering, so that at t_4 we have:

$$(t_4) \quad w_3 = w_4 < w_1 < w_2.$$

And so, what we find is that, upon making these minimal adjustments, you will fail to believe R, since this is a proposition that is false at some world that is amongst the most plausible according to your plausibility ordering at t_4.

The problem may be diagnosed as follows. When you start out uncertain about R and B, you lack the conditional belief $R|\neg B$. For, in your state at the time, you would not come to believe that the animal is red if you were to learn that the animal is not a bird. On the Boutillier model, however, when you get information ϕ, you should minimally adjust your conditional beliefs, i.e., you should only change your conditional beliefs insofar as such a change is forced on you by taking what were previously the most plausible ϕ-worlds to now be the most plausible worlds tout court. Since

coming to believe B and then R does not force one to accept the conditional belief $R|\neg B$, the Boutillier model requires that you continue to lack the conditional belief $R|\neg B$. And so, when you come to believe $\neg B$, since you lack the appropriate conditional belief that would sanction your continuing to believe R, you must give up this belief.

The point that would seem to clearly emerge from this type of example is that if we want to allow that an agent may rationally preserve certain beliefs that she forms over time, we need to allow the agent to adjust her conditional beliefs in certain non-minimal ways that are precluded given the Boutillier model.

In response to the problems of iterated revision faced, on the one hand, by Alchourrón et al. (1985), and, on the other hand, by Boutilier (1996), Darwiche and Pearl (1997) offer an alternative account of iterated revision. Their account is intended to offer stricter constraints on iterated belief revision than those imposed by Alchourrón et al. (1985), while allowing for certain permissible variations in how an agent's conditional beliefs may be updated over time that are ruled out by Boutilier (1996). In addition, their theory is designed to accommodate the second worry about iterated belief revision for AGM considered in Section 2.4.1.

According to Darwiche and Pearl (1997), belief revision should not be thought of, fundamentally, in terms mapping one belief *set* to another. Instead, belief revision should be thought of as mapping one belief *state* to another, where a belief state here is something that determines a belief set but, in addition, encodes information about the agent's conditional beliefs. We can model such a state as a plausibility ordering over the set of possible worlds.[25]

Let $\mathcal{G}$ be the set of belief states. For each $G \in \mathcal{G}$, we'll let $Bel(G)$ be the belief set determined by G. Let $\circ : \mathcal{G} \times \mathcal{L} \to \mathcal{G}$, be a function mapping pairs of belief states and sentences to belief states. Paralleling the AGM postulates, Darwiche and Pearl suggest the following constraints for such a function.

(G_1°) $\phi \in Bel(G_\phi^\circ)$.

(G_2°) $Bel(G_\phi^\circ) \subseteq Bel(G)_\phi^+$.

(G_3°) If $\neg\phi \notin Bel(G)$, then $Bel(G)_\phi^+ \subseteq Bel(G_\phi^\circ)$.

(G_4°) $Bel(B_\phi^\circ) = \mathcal{L}$ just in case $\models \neg\phi$.

(G_5°) If $\models \phi \leftrightarrow \psi$, then $G_\phi^\circ = G_\psi^\circ$.

25 We can, of course, think of the acceptance of a belief revision function on the AGM picture as adopting a policy for mapping one belief state to another. However, importantly, on the AGM account all that matters for this mapping is what the belief set looks like.

(G°_6) $Bel(G^\circ_{\phi\wedge\psi}) \subseteq Bel(G^\circ_\phi)^+_\psi$.

(G°_7) If $\neg\psi \notin Bel(G^\circ_\phi)$, then $Bel(G^\circ_\phi)^+_\psi \subseteq Bel(G^\circ_{\phi\wedge\psi})$.

In addition, however, Darwiche and Pearl also propose the following constraints.

(G°_8) If $\phi \models \psi$, then $Bel((G^\circ_\psi)^\circ_\phi) = Bel(G^\circ_\phi)$.

(G°_9) If $\phi \models \neg\psi$, then $Bel((G^\circ_\psi)^\circ_\phi) = Bel(G^\circ_\phi)$.

(G°_{10}) If $Bel(G^\circ_\phi) \models \psi$, then $Bel((G^\circ_\psi)^\circ_\phi) \models \psi$.

(G°_{11}) If $Bel(G^\circ_\phi) \not\models \neg\psi$, then $Bel((G^\circ_\psi)^\circ_\phi) \not\models \neg\psi$.

Darwiche and Pearl (1997) then show how revision functions satisfying these constraints may be modeled. Let $\preceq$ now be a function that maps each *belief state* G to a total pre-order on the set of possible worlds $\preceq_G$. (Again we'll think of possible worlds as maximal consistent sets of $\mathcal{L}$.) We let $w_1 \prec_G w_2 =_{df} w_1 \preceq_G w_2$ and $w_2 \not\preceq_G w_1$.

DEF. We say that $\preceq$ is a *faithful assignment* just in case:

(i) if $Bel(G) \subseteq w_1$ and $Bel(G) \subseteq w_2$, then $w_1 \preceq_G w_2$ and $w_2 \preceq_G w_1$;

(ii) $Bel(G) \subseteq w_1, Bel(G) \not\subseteq w_2$, then $w_1 \prec_G w_2$.

Let $f : \mathcal{G} \times \mathcal{L} \to \mathcal{G}$. We again let $B_P = \cap\{w : w \in P\}$, given a set of worlds P. And for each $G \in \mathcal{G}$, and each $\preceq$ we let $C_\preceq(G, \phi) = \{w : \phi \in w$ and $w \preceq_G w'$, for each w' such that $\phi \in w'\}$. Darwiche and Pearl (1997) show:

THEOREM 14. Function f satisfies (G°_1)–(G°_7) just in case there exists a faithful assignment $\preceq$ such that $Bel(G^f_\phi) = B_{C_\preceq(G,\phi)}$.

In addition, Darwiche and Pearl (1997) show:

THEOREM 15. If f satisfies (G°_1)–(G°_7), then f satisfies (G°_8)–(G°_{11}) just in case f and any corresponding faithful assignment $\preceq$ such that $Bel(G^f_\phi) = B_{C_\preceq(G,\phi)}$ satisfy:

(iii) if $\phi \in w_1$ and $\phi \in w_2$, then $w_1 \preceq_G w_2$ if and only if $w_1 \preceq_{G^f_\phi} w_2$;

(iv) if $\neg\phi \in w_1$ and $\neg\phi \in w_2$, then $w_1 \preceq_G w_2$ if and only if $w_1 \preceq_{G^f_\phi} w_2$;

(v) if $\phi \in w_1$ and $\neg\phi \in w_2$, then if $w_1 \prec_G w_2$, then $w_1 \prec_{G^f_\phi} w_2$;

(vi) if $\phi \in w_1$ and $\neg\phi \in w_2$, then if $w_1 \preceq_G w_2$, then $w_1 \preceq_{G^f_\phi} w_2$.

Given THEOREM 14 and THEOREM 15, we can think of an agent's belief state as being representable by a total pre-order over the space of possible worlds, while the agent's rational revision policy may be represented as a function mapping a pair of such a pre-order and a sentence to another pre-order. A rational revision policy will be representable by a function, $\circ$, that maps each such pre-order, G, and each sentence, ϕ, to a pre-order G°_{ϕ}, such that:

- the minimal worlds in G°_{ϕ} are the minimal ϕ-worlds in G;
- the ordering amongst the ϕ-worlds in G°_{ϕ} is exactly the ordering of the ϕ-worlds in G;
- the ordering amongst the $\neg\phi$-worlds in G°_{ϕ} is exactly the ordering of the $\neg\phi$-worlds in G;
- any strict or weak preference for a ϕ-world w_1 over a $\neg\phi$-world w_2 in G is preserved in G°_{ϕ}.

Note, however, that the Darwiche and Pearl's account does not require that strict or weak preferences for $\neg\phi$-worlds over ϕ-worlds, given G, be preserved in G°_{ϕ}. More specifically, unlike on the Boutillier model, the Darwiche and Pearl account allows that a ϕ-world w_1, which is non-minimal in G and which may not be strictly preferable to some $\neg\phi$-world w_2 may be strictly preferable to w_2 relative to G°_{ϕ}. And this allows Darwiche and Pearl to deal with the problematic case SEQUENTIAL RED BIRD.

Again let $w_1 = R \wedge B$, $w_2 = \neg R \wedge B$, $w_3 = R \wedge \neg B$ and $w_4 = \neg R \wedge \neg B$. At t_1, your plausibility ordering is such that:

(t'_1) $\quad w_1 = w_2 = w_3 = w_4$.

And, again, given the Darwiche and Pearl model, upon getting the information B, at t_2 you will adjust your ordering so that:

(t'_2) $\quad w_1 = w_2 < w_3 = w_4$.

At t_3, however, upon getting the information R, the Darwiche and Pearl model allows you to adjust your ordering such that:

(t'_3) $\quad w_1 < w_2 < w_3 < w_4$.

Compare this to the ordering that is required by the Boutillier model: $w_1 < w_2 < w_3 = w_4$. The key difference here is that, upon getting the information R, Boutilier (1996) requires that you only promote the most plausible R-world. Darwiche and Pearl (1997), however, allows that each of the R-worlds may be promoted. And, given the ordering $w_1 < w_2 < w_3 < w_4$, upon getting the information $\neg B$ at t_4 you will adjust your ordering so that:

(t_4') $\quad w_3 < w_4 < w_1 < w_2$.

And so what we find is that at the end of this process you will believe $\neg B$ and you will believe R.

The postulates proposed in Darwiche and Pearl (1997), however, are not without problems. In particular, (G_9°) would seem to be subject to potential counterexamples. Thus consider the following case:[26]

> CONJUNCTIVE RED BIRD. You are initially uncertain about whether a certain animal is a red, R, or is a bird, B. Moreover you start out assuming that information about whether or not the animal is a bird, gives you no information about the animal's color. In particular, you do not have the conditional belief $R|\neg B$. You then get information that the animal is a red bird, $R \wedge B$. However, further consultation with an expert indicates that, in fact, the animal is not a bird, $\neg B$. As a result, you wind up believing that the animal is not a bird, $\neg B$, but that the animal is red, R.

Intuitively this would seem to be a rational progression of belief revision. This progression, however, is ruled out as irrational given (G_9°). To see this, note that since $\neg B$ is incompatible with $R \wedge B$, (G_9°) requires that the result of your revising, given $\neg B$, the belief state you have after incorporating $R \wedge B$, be the same as the belief state that would have resulted had you first gotten the information $\neg B$. But since you start out lacking the conditional belief $R|\neg B$, (G_9°), then, precludes your continuing to believe R once you accept $\neg B$.

The problem here would seem to be that upon getting some information, say $R \wedge B$, it may be rational for one to take various parts of that information to be independent of others in the sense that one takes it that one part is true, conditional on some other part turning out to be false. But (G_9°) precludes assuming this sort of independence. It's hard to see, however, why such assumptions of independence should be rationally precluded.

3 DYNAMIC DOXASTIC LOGIC

The models developed in Section 1 allowed us to represent an agent's beliefs, including various higher-order beliefs about the agent's own beliefs. Those models, however, failed to represent important features of an agent's doxastic state. In particular, they failed to provide any representation of an agent's conditional beliefs. The AGM models, on the other hand, allowed us to capture this feature of an agent's doxastic state. However, the AGM

26 See, e.g., Stalnaker (2009) for this type of example.

models failed to provide any representation of an agent's higher-order beliefs.

In this section, we'll begin by presenting models, in the style of Hintikka, that allow us to represent both an agent's unconditional beliefs and her conditional beliefs, and also allow us to represent various higher-order conditional and unconditional beliefs. We'll, then, consider how to add to the language dynamic operators that serve to express how an agent's beliefs, both conditional and unconditional, would be revised in light of new information.

3.1 *Doxastic Plausibility Models*

Let $\mathcal{L}$ be a propositional language including the standard Boolean connectives. In addition, we'll assume that $\mathcal{L}$ contains a binary operator $B_\alpha(\cdot,\cdot)$. As a notational simplification, we will write the second argument as a superscript, so that $B_\alpha(\phi,\psi) =_{\text{df}} B_\alpha^\psi\phi$. The intuitive gloss of $B_\alpha^\psi\phi$ will be "Alpha believes ϕ, conditional on ψ."

A *plausibility model* for $\mathcal{L}$ is a tuple $M = \langle W, \leq, [\![\cdot]\!]\rangle$. W, as before, is a set of worlds, and $[\![\cdot]\!]$ is the *interpretation function* mapping propositional letters to sets of possible worlds. $\leq$ is a *ternary* relation on W. We write this as: $w_1 \leq_w w_2$. The intuitive gloss on this is that, relative to Alpha's plausibility ordering in w, w_1 is at least as plausible as w_2. We assume that, for each w, $\leq_w$ is connected, transitive and satisfies ϕ-minimality. For ease of reference, we list these conditions again.

> CONNECTIVITY. For every $w', w'' \in \mathcal{W}$, either $w' \leq_w w''$ or $w'' \leq_w w'$.
>
> TRANSITIVITY. If $w' \leq_w w''$ and $w'' \leq_w w'''$, then $w' \leq_w w'''$.
>
> ϕ-MINIMALITY. For each $Q \subseteq W$ such that $Q \neq \emptyset$, $\{w' \in Q : w' \leq_w w'',$ for all $w'' \in Q\} \neq \emptyset$.

The truth of a sentence ϕ at a world w in a plausibility model M may be defined inductively in the standard manner. Here we simply give the condition for $B_\alpha^\psi(\phi)$.

> DEF. $[\![\phi]\!]_m =_{\text{df}} \{w : [\![\phi]\!]_m^w = 1\}$.
>
> DEF. For each $Q \subseteq W$, we let $Min_{\leq_w}(Q) =_{\text{df}} \{w' \in Q : w' \leq_w w'',$ for all $w'' \in Q\}$

We then say:

> $[\![B_\alpha^\psi\phi]\!]_m^w = 1$ just in case $Min_{\leq_w}([\![\psi]\!]_m) \subseteq [\![\phi]\!]_m$.

We'll take the notion of unconditional belief to be defined in terms of conditional belief as follows:

DEF. $B_\alpha \phi = B_\alpha^\top \phi$.

Our models here, of course, look quite a lot like the Grove models from Section 2.3.1. There are, however, some differences that are worth highlighting. One, not terribly important, difference is that, in these models, we once again take possible worlds to be primitive entities, instead of maximally consistent sets of sentences. Another, more significant, difference is that our doxastic plausibility models, unlike the Grove models, are defined for a language with an iterable operator that expresses conditional belief. Our models, then, are able to represent the conditional and unconditional beliefs of an agent who has conditional and unconditional beliefs about her own conditional and unconditional beliefs.

We can provide an axiomatic theory for our language $\mathcal{L}$ that is sound and complete with respect to the class of plausibility models so characterized. We'll call this theory C.

AXIOMS OF C

(C_1) $B_\alpha^\phi \phi$.

(C_2) $(B_\alpha^\phi \psi \wedge B_\alpha^\psi \phi) \rightarrow (B_\alpha^\phi \xi \leftrightarrow B_\alpha^\psi \xi)$.

(C_3) $(B_\alpha^{\phi \vee \psi} \phi) \vee (B_\alpha^{\phi \vee \psi} \psi) \vee (B_\alpha^{\phi \vee \psi} \xi \leftrightarrow (B_\alpha^\phi \xi \wedge B_\alpha^\psi \xi))$.

INFERENCE RULES OF C

(TI) If $(\phi_1 \wedge \ldots \wedge \phi_n) \rightarrow \psi$ is a tautology, then $\vdash_c \phi_1 \wedge \ldots \wedge \phi_n \Rightarrow \vdash_c \psi$.[27]

(DWC) If $\vdash_c (\phi_1 \wedge \ldots \wedge \phi_n) \rightarrow \psi$ then $\vdash_c (B_\alpha^\xi \phi_1 \wedge \ldots \wedge B_\alpha^\xi \phi_n) \rightarrow B_\alpha^\xi \psi$.

Axiom (C_1) tells us that, for every ϕ, Alpha believes ϕ conditional on ϕ. Axiom (C_2) tells us that if Alpha believes ψ conditional on ϕ, and ϕ conditional on ψ, then, for any ξ, Alpha believes ξ conditional on ϕ just in case they believe ξ conditional on ψ. Axiom (C_3) tells us that Alpha is such that, conditional on $\phi \vee \psi$, either they believe ϕ, or they believe ψ, or, for every ξ, they believe ξ just in case they believe ξ conditional on ϕ and conditional on ψ. Rule (TI) tells us that the system C is closed under logical entailment. And, finally, (DWC) tells us that Alpha's conditional beliefs are closed under entailment given C.

Let $\mathcal{P}$ be the set of plausibility models satisfying the constraints we've laid down. We, then, have the following result.

THEOREM 16. $\vdash_c \phi \Leftrightarrow \models_{\mathcal{P}} \phi$.[28]

27 It is assumed, for both rules, that if $n = 0$, then $(\phi_1 \wedge \ldots \wedge \phi_n) = \top$, and so the conditional $(\phi_1 \wedge \ldots \wedge \phi_n) \rightarrow \psi$ is equivalent to ψ.

28 For a proof of this result see Lewis (1971). Lewis' proof concerns the logic of conditionals, but the same proof applies when we replace the conditional $\phi > \psi$ with $B_\alpha^\phi(\psi)$.

As with the case of our earlier Kripke models, we can characterize subsets of $\mathcal{P}$ by imposing constraints on the plausibility relation $\leq$.

> DEF. We say that $\leq$ is *minimally homogeneous* just in case for every w, every $z \in Min_{\leq_w}(W)$ is such that $\leq_w = \leq_z$. DEF. We say that $\leq$ is *minimally weakly homogeneous* just in case for every w, every $z \in Min_{\leq_w}(W)$ is such that if $w_1 \not<_w w_2$ then $w_1 \not<_z w_2$.

Let $\mathcal{P}_H$ be the members of $\mathcal{P}$ such that $\leq$ is minimally homogeneous. And let $\mathcal{P}_{WH}$ be the members of $\mathcal{P}$ such that $\leq$ is minimally weakly homogeneous. Now consider the following positive and negative introspection principles.

$(C_4) \quad B^{\phi}_{\alpha}(\psi) \rightarrow B_{\alpha}(B^{\phi}_{\alpha}(\psi))$.

$(C_5) \quad \neg B^{\phi}_{\alpha}(\psi) \rightarrow B_{\alpha}(\neg B^{\phi}_{\alpha}(\psi))$.

Principle (C_4) tells us that if Alpha believes ψ conditional on ϕ, then Alpha believes that they believe ψ conditional on ϕ. And (C_5) tells us that if Alpha does not believe ψ conditional on ϕ, then Alpha believes that they do not believe ψ conditional on ϕ.

We can show:

> THEOREM 17. Principle (C_5) is valid relative to the class $\mathcal{P}_{WH}$.

To see why this result holds, note that for there to be conditional beliefs in z that are not in w there need to be strict preferences amongst worlds in z that are not strict preferences in w. That is, if two worlds differ only in that certain strict preferences, $w_1 <_w w_2$, relative to w, are weak preferences, $w_1 \leq_z w_2$ and $w_2 \leq_z w_1$, relative to z, then while there will be certain conditional beliefs had at w that will not be had at z there will be no additional conditional beliefs at z. Thus, if, in accordance with the condition of minimal weak homogeneity, each of the most plausible worlds z, relative to w, imposes no strict preferences that are not imposed in w, then any conditional belief that Alpha fails to have in w, will also be such that Alpha fails to have it in z. And so, if $\leq$ is minimally weakly homogeneous, then, for any w, if Alpha fails to believe some ϕ conditional on ψ, then she will also fail to believe ϕ conditional on ψ relative to the most plausible worlds, given w, and so Alpha, at w, will believe that she fails to believe ϕ conditional on ψ.

We can also show the following.

> THEOREM 18. Principles (C_4) and (C_5) are valid relative to the class $\mathcal{P}_H$.

This result should be obvious, since any two worlds w and z, such that $\leq_w = \leq_z$, will agree about all conditional belief facts. It's worth, however, pointing out that there is no weaker condition on $\leq$ that will ensure the validity of (C_4). The reason for this is that if two worlds w and z are such that $\leq_w \neq \leq_z$, then there will be some possible assignment such that the conditional beliefs relative to w will differ from those at z. To see this, assume that we have $w_1 \leq_w w_2$ and $w_1 \not\leq_z w_2$, and so $w_2 <_z w_1$. Let ϕ and ψ be atomic sentences such that $I(\phi) = \{w_1, w_2\}$ and $I(\psi) = \{w_2\}$. Then, given our assumptions about $\leq$, relative to these assignments, we will have $[\![B^{\phi}_{\alpha}\psi]\!]^z_m = 1$ and $[\![B^{\phi}_{\alpha}\psi]\!]^w_m = 0$. Thus, minimal homogeneity is the weakest condition on $\leq$ that will ensure that, for every w, every conditional belief in w is a conditional belief in each of the minimal worlds (relative to w).

Our plausibility models can clearly be generalized to the multi-agent setting. And in such models, various conditional generalizations of the group doxastic properties of common and distributed belief can be represented. We won't, however, consider such models here. Instead, we'll move on to consider how certain dynamic operators, representing facts about how an agent's conditional beliefs would be revised given new information, may be added to our language.

3.2 *Dynamic Operators*

Let us add to our language $\mathcal{L}$ the following binary operator $[^*_\alpha]$. The rough intuitive reading of $[^*_\alpha\phi]\psi$ will be "ψ holds after Alpha revises its belief state given information ϕ." A model for our augmented language will still be a tuple $M = \langle W, \leq, [\![\cdot]\!]\rangle$, with $\leq$ subject to the same constraints. In order to characterize truth-at-a-world for formulas of the form $[^*_\alpha\phi]\psi$, however, we first need to characterize an operation on the set of models $\mathcal{P}$.

For illustrative purposes, we'll assume that rational belief revision for idealized agents works in the manner described in Boutilier (1996). That is, given an agent with a plausibility ordering over the space of worlds, such an agent will revise their belief state, given information ϕ, by minimally adjusting their plausibility ordering so that the order remains the same except that the most plausible ϕ worlds are now the most plausible worlds tout court.

> DEF. For each $w \in W$ and $Q \subseteq W$, let $C(\leq_w, Q) = \{z \in W : z \in Q$ and for all $x \in Q$, $z \leq_w x\}$.
>
> DEF. For each $w \in W$ let $\leq^{*q}_w$ be the binary relation on W such that (i) for every $z \in C(\leq_w, Q)$ and every $x \in W$, $z \leq^{*q}_w x$, and (ii) for every $x, z \in W - C(\leq_w, Q)$ $z \leq^{*q}_w x$ iff $z \leq_w x$.[29]

29 Note that given that $\leq_w$ is transitive, connected and satisfies ϕ-minimality, so too will $\leq^{*q}_w$.

We can now characterize the truth of a sentence $[^*_\alpha\phi]\psi$ at a world w in a model $M = \langle W, \leq, [\![\cdot]\!]\rangle$. We say:

> $[\![[^*_\alpha\phi]\psi]\!]^w_m = 1$ iff $[\![\psi]\!]^w_{m'} = 1$, where $M' = \langle W, \leq', [\![\cdot]\!]\rangle$ is such that for each $z \in W$ $\leq'_z = \leq^{*[\![\phi]\!]}_z$.

Operators such as B_α in our earlier Kripke models, or B^ϕ_α in our current plausibility models, function by shifting the world of evaluation. Operator $[^*_\alpha\phi]$ is, however, quite different in nature. Instead of shifting the world parameter of evaluation, $[^*_\alpha\phi]$ shifts the *model* of evaluation. We'll call operators that have the semantic function of shifting models of evaluation in this manner *dynamic operators*.

It has been shown, in van Bentham (2007), that if we add to C the following so-called reduction axioms, we get an axiomatic system that is sound and complete relative to the class of models $\mathcal{P}$ given this semantics.

(C_6) $[^*_\alpha\phi]\psi$ for each atomic ψ.

(C_7) $[^*_\alpha\phi]\neg\psi \leftrightarrow \neg[^*_\alpha\phi]\psi$.

(C_8) $[^*_\alpha\phi]\psi \wedge \xi \leftrightarrow [^*_\alpha\phi]\psi \wedge [^*_\alpha\phi]\xi$.

(C_9) $[^*_\alpha\phi]B^\psi_\alpha(\xi) \leftrightarrow [(B^\phi_\alpha\neg[^*_\alpha\phi]\psi) \wedge (B^{[^*_\alpha\phi](\psi)}_\alpha[^*_\alpha\phi]\xi)] \vee [(\neg B^\phi_\alpha\neg[^*_\alpha\phi]\psi) \wedge (B^{\phi\wedge[^*_\alpha\phi]\psi}_\alpha[^*_\alpha\phi]\xi)]$.

Axiom (C_6) tells us that atomic statements will not change their truth-value when Alpha revises its belief state given information ϕ. Axiom (C_7) tells us $\neg\psi$ will hold when Alpha revises its belief state given information ϕ just in case ψ does not hold when Alpha revises its belief state given information ϕ. Axiom (C_8) tells us that a conjunction will hold when Alpha revises its belief state given information ϕ just in case both of the conjuncts hold. These latter conditions are all fairly intuitive. Unfortunately, axiom (C_9) is much more unwieldy and lacks a simple intuitive gloss. This principle tells us that at least one of the following two conditions must obtain.

(i) Alpha believes ξ, conditional on ψ, when Alpha revises its belief state given information ϕ just in case, (a) conditional on ϕ, Alpha believes that it's not the case that if they revise their belief state given ϕ, then ψ will hold, and (b) conditional on ψ holding, if Alpha revises its belief state given ϕ, then Alpha believes that if they revise their belief state given ϕ, then ξ holds.

(ii) (a) It is not the case that, conditional on ϕ, Alpha believes that it's not the case that, if Alpha revises their beliefs given ϕ, then ψ, and (b) Alpha believes, conditional on the conjunction of ϕ and the claim that if Alpha revises their beliefs given ϕ, then ψ will hold, that if Alpha revises their beliefs given ϕ, then ξ will hold.

One thing that these reduction axioms highlight is that the addition of $[^*_\alpha]$ to our language $\mathcal{L}$ in fact adds no real expressive power. For the reduction axioms show that any formula involving such an operator is equivalent to some formula that doesn't contain this operator. The equivalent $[^*_\alpha]$-free formulas may, however, be extremely complex. The introduction of $[^*_\alpha]$, then, provides a way of expressing, in a concise manner, claims that might otherwise lack a simple expression.

I said earlier that the rough gloss on $[^*_\alpha\phi]\psi$ will be "ψ holds after Alpha revises its belief state given information ϕ." However, the semantics for $[^*_\alpha]$ encodes certain idealizing assumptions about what happens when an agent revises her beliefs given new information. In particular, the semantics we've outlined entails that if ϕ is an atomic sentence, then, when an agent gets new information ϕ, not only will the agent come to believe ϕ, but the agent will believe that they believe ϕ, and believe that they believe that they believe ϕ, and so on. Let B^n_α abbreviate n iterations of B_α. Then, if ϕ is atomic, we have that, for any n, $[\![[^*_\alpha\phi]B^n_\alpha\phi]\!]^w_m = 1$, for all worlds w and models M. New information, at least when it concerns some atomic proposition, will, on this model, be transparent to an agent.

The reason that such formulas are valid, given our semantics, is that an evaluation of the truth of $[^*_\alpha\phi]\psi$ in a model M at a world w, requires us to assess ψ at w relative to a model M' which differs from M in that the best ϕ-worlds relative to $\leq_w$ are the best worlds relative to each $\leq_z$. On this semantic theory, $[^*_\alpha\phi]$ effects a global shift on the plausibility ordering. In order to avoid the assumption that new information will not only be believed, but also believed to be believed etc., one would need to take the operator $[^*_\alpha\phi]$ to simply shift the the model M to a model M' whose plausibility ordering only differs relative to the world of evaluation w. We won't, however, look at these alternative semantic treatments here.

We've seen, then, how we can introduce a dynamic operator into our language that, in a certain sense, corresponds to the belief revision policy of Boutilier (1996). As noted earlier, though, any belief revision policy satisfying the AGM postulates can be thought of as a function mapping a belief state encoding conditional beliefs to another such state. Given any such policy f, then, we can introduce a dynamic operator $[^f]$. A formula of the form $[^f\phi]\psi$ will be true in a model M at a world w just in case ψ is true in a model M' at w, where M' is the model which shifts each $\leq_z$ to $f(\leq_z)$.

The application of dynamic operators in doxastic and epistemic logic is a rapidly developing area of study. In addition to expressing the sorts of revision that AGM was concerned with, dynamic operators can also be used to express other sorts of doxastic and epistemic changes, such as the doxastic results of so-called public announcements, which make certain pieces of information common knowledge amongst a group of agents.

The literature here is vast and growing, and a thorough survey is beyond the scope of this work. Our goal here has, instead, been to simply give a sense of how such dynamic operators function. The following, though, provides a small sample of work in this tradition: Baltag, Moss, and Solecki (1998), Segerberg (1998), Segerberg (2001), van Ditmarsche (2005), Baltag and Smets (2006a), Baltag and Smets (2006b), Rott (2006), Leitgeb and Segerberg (2007), van Bentham (2007), van Ditmarsche, van der Hoek, and Kooi (2008), Beltag and Smets (2008), van Bentham (2011) Girard and Rott (2014).

4 DOXASTIC PARADOXES

In this final section, we'll look at two doxastic paradoxes and consider, on the one hand, how some of the tools developed in the previous sections may be brought to bear to analyze these cases, and, on the other hand, how such paradoxes may serve to call into question certain assumptions made earlier about the principles governing the doxastic states of idealized agents.

4.1 *Moore's Paradox*

As Moore (1942) famously noted, there is something decidedly odd about the sentence 'ϕ and I don't believe ϕ'. What is puzzling about the case is that, while claiming that ϕ and I don't believe ϕ would seem, in some way, to be incoherent, the claim itself is perfectly consistent. There is nothing that prevents it from being true that ϕ and I don't believe ϕ.

Hintikka (1962) argued that the oddity of Moore paradoxical sentences such as $\phi \wedge \neg B_\alpha \phi$ can be explained by the fact that such claims are unbelievable for agents whose doxastic states meet certain constraints. Thus, let us assume that Alpha is an agent whose doxastic state is consistent, closed under logical consequence, and satisfies positive introspection. Given these assumptions, we can show that the following can never be true $B_\alpha(\phi \wedge \neg B_\alpha \phi)$. For assume that it is. Then since Alpha's doxastic state is closed under logical consequence we have $B_\alpha \phi$ and $B_\alpha \neg B_\alpha \phi$. And, since Alpha's doxastic state satisfies positive introspection, we have $B_\alpha B_\alpha \phi$. But, then, contrary to our assumption, Alpha's doxastic state is inconsistent.

Besides being unbelievable for certain agents, Moore paradoxical sentences have other odd features. Let us assume that our agent Alpha's idealized doxastic state is consistent, logically closed, and satisfies positive and negative introspection. Then such an agent's doxastic state may be represented by a KD45 model. More specifically, we can represent the agent's

doxastic state, as well as other facts about the world, by a particular point w in a some KD45 model $M = \langle W, R_\alpha, [\![\cdot]\!] \rangle$.

Now given an idealized agent such as Alpha, whose doxastic state may be represented by a particular point w in a some KD45 model M, we can represent the change in such an agent's doxastic state that results from getting some true information ϕ by the point w in a model M_ϕ. In particular, let $W^\phi = W \cap [\![\phi]\!]_m$, $R_\alpha^\phi = R_\alpha \cap W^\phi \times W^\phi$, and $[\![\cdot]\!]^\phi = [\![\cdot]\!] \cap W^\phi$. Then the model representing Alpha's doxastic state, after Alpha has received some true information ϕ will be $M_\phi = \langle W^\phi, R_\alpha^\phi, [\![\cdot]\!]^\phi \rangle$. We can think of M_ϕ as the model that results when one removes from M all of the worlds in which ϕ is false and then minimally adjusts the accessibility relation and valuation function.

Interestingly, there are certain sentences ϕ that, while true relative to w and M, may be false relative to w and M_ϕ. Indeed, there are certain sentences ϕ that are *guaranteed* to be false relative to w and M_ϕ. Call such sentences *self-refuting*. If ϕ is an atomic sentence, then the Moore paradoxical sentence $\phi \wedge \neg B_\alpha \phi$ is a paradigmatic case of a self-refuting sentence.

Let M be a KD45 model and w such that $[\![\phi \wedge \neg B_\alpha \phi]\!]_m^w = 1$. Let $M' = M_{\phi \wedge \neg B_\alpha \phi}$. Then it's guaranteed that $[\![\phi \wedge \neg B_\alpha \phi]\!]_{m'}^w = 0$. For, since any world in M in which ϕ is false makes $\phi \wedge \neg B_\alpha \phi$ false, each $\neg\phi$-world in M will be removed from M'. But, then, since ϕ is atomic, it follows that ϕ must be true for every world in M'. But this guarantees that we have $[\![B_\alpha \phi]\!]_{m'}^w = 1$ and so $[\![\phi \wedge \neg B_\alpha \phi]\!]_{m'}^w = 0$. Indeed, we can see that this reasoning establishes that, for each $w' \in W'$, $[\![\phi \wedge \neg B_\alpha \phi]\!]_{m'}^{w'} = 0$.

Moore paradoxical sentences, then, are not only unbelievable for certain idealized agents, they are also such that if they are true and learned to be so by such an agent then they become false.[30] While Moore paradoxical sentences may be true, their truth is, in a particular manner, unstable.

The fact that a Moore paradoxical sentence $\phi \wedge \neg B_\alpha \phi$ fails to hold for each point in the model $M_{\phi \wedge \neg B_\alpha \phi}$ is relevant for the assessment of certain principles of belief revision. For recall that, in AGM, it is assumed that $\phi \in B_\phi^*$. That is, upon revising their belief set by ϕ, an ideal agent will believe ϕ. This, of course, seems prima facie quite plausible, but Moore paradoxical sentences would seem to provide a counterexample to this claim. For, given that one's doxastic state upon learning $\phi \wedge \neg B_\alpha \phi$ is represented by $M_{\phi \wedge \neg B_\alpha \phi}$, we've seen that, upon revising one's belief set given $\phi \wedge \neg B_\alpha \phi$, this sentence will not be believed.

Now there are a few ways of responding to this worry.

First, one could grant that this is a counter-example to (B_1^*) formulated as an unrestricted principle governing belief revision. However, one could

30 Holliday and Icard (2010) show that, in a certain sense, for introspective agents all self-refuting formulas are Moore paradoxical in character.

claim that this principle, properly understood, should only apply to sentences that don't contain any belief operators. And, indeed, we can show that any sentence ϕ that doesn't contain such operators will be guaranteed to hold at any point w in M_ϕ, if it held at w in M.[31] And, furthermore, as we earlier noted, the languages that the proponents of AGM initially considered simply had no resources for talking about particular agent's beliefs or revision policies.

Another response, though, would be to argue that, properly construed, belief revision should concern *propositions*. The correct principle in the vicinity, then, is that if one learns some proposition ϕ, then one's revised belief state, in light of this, should include that proposition. One may argue, then, that if we think of the objects of belief as propositions and so revision policies as concerning which propositions one should believe, given new information, the problem for (B_1^*), so construed, disappears. For while Moore paradoxical sentences are self-refuting, Moore-paradoxical *propositions* are not.[32] For a Moore-paradoxical proposition will be time-indexed. But, if one learns between t_1 and t_2, that ϕ held at t_1 but one did not believe ϕ, this claim will remain true and may be consistently believed at t_2.

The AGM account of belief revision was formulated on the assumption that the objects of belief are sentences. Moorean phenomena, however, make it apparent that, if one wants to maintain one of the most basic principles of the theory, then the correct formulation of this theory should, instead, take the objects of belief to be propositions.

4.2 *The Burge-Buridan Paradox*

So far we've assumed that an idealized agent will have beliefs that are consistent, logically omniscient, closed under logical consequence, and that satisfy positive introspection. A close cousin of Moore's paradox, however, would seem to show that these constraints cannot be jointly satisfied if the expressive power of the language over which our doxastic models are defined is enriched in a certain manner.

31 Indeed the class of formulas with this property is larger than the class of formulas lacking any belief operators. See Holliday and Icard (2010). So there is room to enlarge the scope of this principle to certain sentences containing belief operators.

32 To be clear, by 'proposition' I mean an *eternal* proposition, i.e., something that determines a function from worlds to truth-values. The present points don't hold if one thinks that the objects of belief are temporal propositions, i.e., things that only serve to determine a function from world, time pairs to truth-values.

We'll call sentences such as the following *Burge-Buridan sentences*: "I don't believe that this sentence is true."[33] If we consider an agent who can entertain the proposition expressed by a sentence such as this, we can show, given plausible auxiliary assumptions, that this agent cannot satisfy all of the constraints we've imposed on idealized doxastic states.

So far, we have been working with propositional languages. To treat the Burge-Buridan sentence in a formal setting, however, we need to add to our language $\mathcal{L}$ a single predicate $T(\cdot)$, as well as a single term β. The intuitive interpretation of $T(\beta)$ will be that the sentence referred to by β is true. Being a sentence of our predicate language $\mathcal{L}$ may be defined in the standard manner.

We will stipulate that in our language $\mathcal{L}$ the term β refers to the sentence $\neg B_\alpha T(\beta)$. As an instance of the T-schema, then, we have:

$$\text{(T)} \quad T(\beta) \leftrightarrow \neg B_\alpha T(\beta)$$

Given a conception of logic on which the valid principles governing truth count as logical truths, it is quite plausible that (T) will count as a logical truth.[34] Assume, then, that our idealized agent Alpha satisfies logical omniscience. Then, we have $B_\alpha(T(\beta) \leftrightarrow \neg B_\alpha T(\beta))$. Now we can show that Alpha's doxastic state cannot also be consistent, logically closed and satisfy positive transparency. Our proof will proceed by cases. First, assume $\neg B_\alpha T(\beta)$. Then, it follows by closure that $B_\alpha T(\beta)$ which, of course, contradicts our assumption. Next, assume $B_\alpha T(\beta)$. Then, by closure we have $B_\alpha \neg B_\alpha T(\beta)$. But, by positive introspection, we also have $B_\alpha B_\alpha T(\beta)$. And so Alpha fails to have a consistent doxastic state.

Although Alpha cannot have a doxastic state that is consistent, logically omniscient, logically closed and positively transparent, there is no problem, in principle, with Alpha having a doxastic state that satisfies only the first three constraints. To do so, we provide a model in which these properties will be satisfied.

A doxastic model for $\mathcal{L}$ is a tuple $M = \langle W, R_\alpha, D, [\![\cdot]\!] \rangle$. W and R_α are the same as in our earlier propositional doxastic models, while D is a set of objects, and $[\![\cdot]\!]$ is a function which assigns, to propositional letters, subsets of W, to singular terms, elements of D, and, to unary predicates, functions mapping elements of w to subsets of D. Truth in such a model is defined in the obvious way.

---

33 This type of sentence was first discussed in the modern literature in Burge (1978), who attributes the paradox it raises to Buridan. For other discussion see, e.g., Burge (1984), Conee (1987), Sorensen (1988), Caie (2011), and Caie (2012).

34 Note that this instance of the T-schema is compatible with classical logic. This is established by the model given below. Even, then, if one thinks that cases such as the Liar paradox should lead us to reject certain instances of the T-schema as invalid, we don't have similar reason to reject *this* instance of the T-schema.

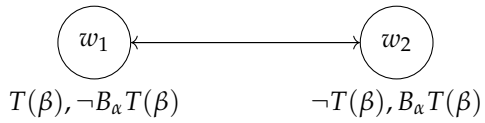


Figure 2: Modeling the Burge-Buridan sentence

Now, let $W = \{w_1, w_2, \}$, let $R_\alpha = \{\langle w_1, w_2\rangle, \langle w_2, w_1\rangle\}$ and let $[\![T]\!] = \{\langle w_1, \{\beta\}\rangle, \langle w_2, \emptyset\rangle\}$. Then we have $[\![T(\beta)]\!]_m^{w_1} = [\![\neg B_\alpha T(\beta)]\!]_m^{w_1} = 1$ and $[\![\neg T(\beta)]\!]_m^{w_2} = [\![B_\alpha T(\beta)]\!]_m^{w_2} = 1$. We may picture this model as in Figure 2.

In the model under consideration, $[\![\beta]\!] = \neg B_\alpha T(\beta)$. This corresponds to our stipulation that the sentence of $\mathcal{L}$, $\neg B_\alpha T(\beta)$, will be the denotation of the term β. Moreover, for each world w, $[\![T(\beta)]\!]_m^w = 1$ just in case $[\![\neg B_\alpha T(\beta)]\!]_m^w = 1$. This corresponds to the assumption that the T-schema for β is believed by Alpha to hold. Alpha moreover will believe all propositional logical truths, as well as any other logical truth that follows from the assumption that the T-schema holds. Since the relation R_α is serial, it follows that Alpha's doxastic state is consistent. And, as with any possible worlds doxastic model, Alpha's beliefs will be closed under logical consequence.

Thus, while idealized agents can be consistent, logically omniscient, and have beliefs that are closed under logical consequence, they cannot always be, in addition, positively transparent.

Now, there are certain ways around this result. For example, if we weaken our background logic governing the Boolean connectives, then we can show that the Burge-Buridan sentences do not preclude an idealized agent from satisfying positive and negative transparency, in addition to consistency, omniscience, and logical closure.[35] However, in order for this to be a non-ad hoc move, the weakening of the background logic would need to be sufficiently independently motivated. And whether this is so is a controversial matter.

We've considered two classes of sentences, the Moore-paradoxical and the Burge-Buridan sentences. It's worth noting, however, that the latter class is really a subclass of the former. In general, a Moore-paradoxical sentence is one that has the following form $\phi \wedge \neg B\phi$. A Burge-Buridan sentence, on the other hand, has the form $\neg BT(\beta)$, where β refers to that very sentence. On the surface, of course, this does not seem to have the form of a Moore-paradoxical sentence. However, given the plausible assumption that $T(\beta)$ and $\neg BT(\beta)$ are logically equivalent, then we get that $\neg BT(\beta)$ is, in fact, equivalent to $T(\beta) \wedge \neg BT(\beta)$. Thus a Burge-Buridan sentence, while not having the overt form of a Moore-paradoxical sentence, is equivalent to a Moore-paradoxical sentence. This sub-class of the

35 See Caie (2012) for a proof of this.

Moore-paradoxical sentences, however, have striking consequences that other members of the class of Moore-paradoxical sentences lack. For it's only with these degenerate cases of Moore-paradoxicality that we find that transparency assumptions come into conflict with other plausible principles governing idealized doxastic states.

REFERENCES

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

Aumann, R. (1976). Agreeing to disagree. *The Annals of Statistics*, 1236–1239.

Baltag, A., Moss, L., & Solecki, S. (1998). The logic of piblic announcements, common knowledge and private suspicions. In *Proceedings of the 7th conference on theoretical aspects of rationality and knowledge* (pp. 43–56). Morgan Kaufmann Publishers.

Baltag, A. & Smets, S. (2006a). Conditional doxastic models: A qualitative approach to dynamic belief revision. *Electronic Notes in Theoretical Computer Science*, *165*, 5–21.

Baltag, A. & Smets, S. (2006b). Dynamic belief revision over multi-agent plausibility models. *Proceedings of LOFT*, *6*, 11–24.

Barwise, J. (1988). Three views of common knowledge. In *Proceedings of the 2nd conference on theoretical aspects of reasoning about knowledge* (pp. 365–379). Morgan Kaufmann Publishers Inc.

Beltag, A. & Smets, S. (2008). A qualitative theory of dynamic interactive belief revision. *Logic and the Foundations of Game and Decision Theory*, *3*, 9–58.

Blackburn, P., de Rijke, M., & Venema, Y. (2001). *Modal logic*. Cambridge University Press.

Bonanno, G. & Nehring, K. (2000). Common belief with the logic of individual belief. *Mathematical Logic Quarterly*, *46*(1), 49–52.

Boutilier, C. (1996). Iterated revision and minimal revision of conditional beliefs. *Journal of Philosophical Logic*, *25*, 262–305.

Burge, T. (1978). Buridan and epistemic paradox. *Philosophical Studies*, *34*, 21–35.

Burge, T. (1984). Epistemic paradox. *Journal of Philosophy*, *81*(1), 5–29.

Caie, M. (2011). *Paradox and belief* (Doctoral dissertation, University of California, Berkeley).

Caie, M. (2012). Belief and indeterminacy. *The Philosophical Review*, *121*(1), 1–54.

Chellas, B. (1980). *Modal logic: An introduction*. Cambridge University Press.

Colombetti, M. (1993). Formal semantics for mutual belief. *Artificial Intelligence*, *63*(341-353).

Conee, E. (1987). Evident, but rationally unacceptable. *Australasian Journal of Philosophy*, *65*, 316–326.

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

Fagin, R., Halpern, J., & Vardi, M. (1991). A model-theoretic analysis fo knolwedge. *Journal of ACM*, *38*(2), 382–428.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. 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). San Francisco: Morgan Kaufmann.

Girard, P. & Rott, H. (2014). *Belief revision and dynamic logic*. (ms.)

Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic*, *17*, 157–170.

Halpern, J. & Moses, Y. (1984). Knowledge and common knowledge in a distributed environment. In *Proceedings of the 3rd acm confernece on principles of distributed computing* (pp. 50–61).

Halpern, J. & Moses, Y. (1992). A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, *54*(319-379).

Halpern, J., Moses, Y., & Vardi, M. (1995). *Reasoning about knowledge*. MIT Press.

Harper, W. (1976). Rational conceptual change. *PSA*, *2*, 462–494.

Hintikka, J. (1962). *Knowledge and belief: An introduction to the logic of the two notions*. Cornell University Press.

Holliday, W. H. & Icard, T. F. (2010). Moorean phenomena in epistemic logic. In L. Beklemishev, V. Goranko, & V. Shehtman (Eds.), *Advances in modal logic* (Vol. 8, pp. 178–199).

Huber, F. (2016). Ranking theory. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*.

Hughes, G. E. & Cresswell, M. J. (1996). *A new introduction to modal logic*. Psychology Press.

Leitgeb, H. & Segerberg, K. (2007). Dynamic doxastic logic: Why, how and where to? *Synthese*, *155*, 167–190.

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

Levi, I. (1988). Iteration of conditionals and the ramsey test. *Synthese*, *76*, 49–81.

Lewis, D. (1969). *Convention*. Cambridge University Press.

Lewis, D. (1971). Completeness and decidability of three logics of counterfactual conditionals. *Theoria*, *37*(1), 74–85.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Lewis, D. (1974). Radical interpretation. *Synthese*, *23*, 331–44.

Lewis, D. (1999). Reduction of mind. In *Papers on metaphysics and epistemology*. Cambridge University Press.

Lewis, D. (2000). Logic for equivocators. In *Papers in philosophical logic*. Cambridge University Press.

Lismont, L. & Mongin, P. (1994). On the logic of common belief and common knowledge. *Theory and Decision*, *37*, 75–106.

Moore, G. E. (1942). A reply to my critics. In P. Schilpp (Ed.), *The philosophy of g.e. moore* (Vol. 4, pp. 535–677). The Libraray of Living Philosophers. Northwester University.

Rott, H. (2006). Shifting priorities: Simple representations of twenty-seven iterated theory change operations. In H. Lagerlund, S. Lindstrom, & R. Sliwinski (Eds.), *Modality matters* (Vol. 53, pp. 359–385). Uppsala Philosophical Studies.

Segerberg, K. (1998). Irrevocable belief revision in dynamic doxastic logic. *Notre Dame Journal of Formal Logic*, *39*, 287–306.

Segerberg, K. (2001). The basic dynamic doxastic logic of agm. In M. Williams & H. Rott (Eds.), *Frontiers in belief revision* (pp. 57–84). Dordrecht: Kluwer.

Shoemaker, S. (1996a). Moore's paradox and self-knowledge. In *The first-person perspecive and other essays* (pp. 74–96). Cambridge University Press.

Shoemaker, S. (1996b). On knowing one's own mind. In *The first-person perspecive and other essays* (pp. 25–49). Cambridge University Press.

Sorensen, R. (1988). *Blindspots*. Oxford University Press.

Stalnaker, R. (1984). *Inquiry*. MIT Press.

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

van Bentham, J. (2007). Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, *17*(2), 129–155.

van Bentham, J. (2011). *Logical dynamics of information and interaction*. Cambridge University Press.

van Ditmarsche, H. (2005). Prolegomena to dynamic logic for belief revision. *Synthese*, *147*, 229–275.

van Ditmarsche, H., van der Hoek, W., & Kooi, B. (2008). *Dynamic epistemic logic*. Springer.

Williamson, T. (2000). *Knowledge and it's limits*. Oxford University Press.

11

CONDITIONALS

R. A. Briggs

Conditionals are sentences that propose a scenario (which may or may not be the actual scenario), then go on to say something about what would happen in that scenario.[1] In English, they are typically expressed by 'if... then...' statements. Examples of conditionals include:

1. If the Axiom of Choice is true, then every set can be well ordered.
2. You will probably get lung cancer if you smoke.
3. If the syrup forms a soft ball when you drop it into cold water, then it is between 112 and 115 degrees Celsius.
4. If kangaroos had no tails, they would topple over.
5. When I'm queen, you will be sorry.

In general, a conditional is formed from two smaller statements: an *antecedent* (the supposition that typically comes directly after 'if') and a *consequent* (the statement that typically comes later in the sentence, and is sometimes preceded by 'then'). In the above examples, the antecedents are:

1. The Axiom of Choice is true.
2. You smoke.
3. The syrup forms a soft ball when you drop it into cold water.
4. Kangaroos had no tails. (Or perhaps: Kangaroos have no tails.)
5. I'm queen.

while the consequents are:

1. Every set can be well ordered.
2. You will probably get lung cancer. (Or perhaps: You will get lung cancer.)
3. The syrup is between 112 and 115 degrees Celsius.
4. Kangaroos would topple over. (Or perhaps: Kangaroos topple over.)
5. You will be sorry.

1 I take this framing, which emphasizes the contents of conditionals rather than their grammatical form, from (Fintel, 2011).

1 WHY CARE ABOUT CONDITIONALS?

Conditionals are useful for a variety of everyday tasks, including decision making, prediction, explanation, and imagination.

When making a decision, you should aim to choose an act such that, if you (were to) perform it, a good outcome is (or would be) likely to result. Decision theory codifies this intuition in formal terms, and often makes explicit use of conditionals (Gibbard & Harper, 1981; Vinci, 1988; Bradley, 2000; Cantwell, 2013).

Conditionals are also useful for deriving predictions and explanations from theoretical models. If I am not sure which model of climate change to accept, I can use conditionals to reason about how much the earth's temperature will increase if each of the models under consideration is true. To check whether a model explains the data I have already observed, I can use conditionals to check whether, if a given model is true, my data should be expected. (For a defense of conditionals in scientific explanation, see Woodward, 2004; for a defense of conditionals in historical explanation, see Reiss, 2009, and Nolan, 2013.)

Children's pretend play is both developmentally important, and closely related to reasoning with conditionals. Amsel and Smalley (2000), Dias and Harris (1990), Gopnik (2009), Harris (2000), Lillard (2001), and K. Walton (1990) argue that children's pretense (for example, pretending a banana is a telephone), involves constructing an alternative scenario to what is known or believed to be true, and then reasoning about what would happen in that scenario. While children can express their thoughts about pretend scenarios without the explicit use of conditionals, conditionals are particularly well suited to expressing these thoughts. Weisberg and Gopnik (2013) argue that the ability to reason about non-actual scenarios is crucial to learning from and planning for the actual world, since it enables children to generate and compare a range of alternative models of reality. Krzyzanowska (2013) argues that the mechanism that lets children evaluate conditionals is the same as the one that lets them attribute false beliefs to others.

In addition to playing a crucial role in everyday reasoning and cognitive development, conditionals do work in philosophical analyses of a variety of concepts. Any philosophical idea that relies on the notion of dependence is ripe for a conditional analysis: to say that one thing e depends on a second thing c is arguably to say that if c is one way, then e is some corresponding way, and if c is a different way, then e is a correspondingly different way. Conditionals famously appear in analyses of causation (see Menzies, 2014, and Collins, Hall, and Paul, 2004, for overviews), dispositions (Prior, Pargetter, & Jackson, 1982; Choi, 2006, 2009), knowledge (Nozick, 1981; Sosa, 1999), and freedom (Moore, 1912; Ayer, 1954).

Finally, conditionals figure in several common patterns of reasoning, to which we now turn.

2 COMMON PATTERNS OF REASONING

The following argument forms look compelling in ordinary, natural-language arguments (though we will see that all of them have putative counterexamples). Different formal theories of conditionals yield different verdicts about which are valid.

2.1 *Modus Ponens*

Modus ponens is the inference form:

1. If A, then C.
2. A.

$\therefore$ C.

Modus ponens is one of the most central—arguably the most central—of the inference forms involving conditionals. Bobzien (2002) traces its roots back to Aristotle's hypothetical syllogisms, and through the logic of the Peripatetics and antiquity. Gillon (2011) notes that modus ponens was a common inference pattern in Pre-Classical Indian philosophy, and quotes a representative argument in which the third-century Buddhist logician Moggaliputta Tissa explicitly notes the inconsistency of simultaneously believing 'if A, then C', 'A', and 'not C'. Ryle (1950) even advances theory of conditionals based entirely on their ability to license modus ponens: an utterance of 'if A then C' is an 'inference ticket' that allows one to move from the premise A to the conclusion C.

Despite its perennial popularity, there are apparent counterexamples to modus ponens. One sort (McGee, 1985) involves nested conditionals. Suppose you see a fisherman with something caught in his net. You are almost sure it is a fish, but the next likeliest option is that it is a frog. McGee argues that you should accept the premises of the following argument, but not the conclusion (since, if the animal has lungs, then it is not a fish but a frog).

1. If that is a fish, then if it has lungs, it's a lungfish.
2. That is a fish.

$\therefore$ If it has lungs, it's a lungfish.

Another type of apparent counterexample (Kolodny & MacFarlane, 2010; Darwall, 1983) involves 'ought's or 'should's. Consider this variant of Darwall's example.

1. If you want to hurt my feelings, you should make fun of the way my ears stick out.
2. You want to hurt my feelings.

∴ Therefore, you should make fun of the way my ears stick out.

Even if you do want to hurt my feelings, you shouldn't make fun of the way my ears stick out, because it's wrong to hurt my feelings. Dowell (2011), and Lauer and Condoravdi (2014) object to the Darwall example (and other, related examples) on the grounds that they equivocate on different meanings of 'should'.

Yet another type of apparent counterexample to modus ponens, discussed by D. Walton (2001), involves defeasible inferences, like the famous Tweety Bird example from cognitive science (Brewka, 1991)

1. If Tweety is a bird, then Tweety flies.
2. Tweety is a bird.

∴ Tweety flies.

The first premise of the Tweety bird argument says that there is a defeasible connection between being a bird and flying—one that can be overridden by extra information, e.g., that Tweety is a penguin. Thus, the premises are true, and the conclusion false, in the case where Tweety is a penguin.

2.2 *Modus Tollens*

Modus tollens is the inference form:

1. If A, then C.
2. Not C.

∴ Not A.

Modus ponens and modus tollens seem to have originated together (see Bobzien, 2002, and Gillon, 2011), and are closely related. Both inferences posit a three-way inconsistency between 'if A, then C', 'A' and 'not C'. Affirm two of these inconsistent claims, and you'll have to deny the third.

Yalcin (2012a) presents a putative counterexample to modus tollens. Consider an urn that contains 100 marbles—some red, some blue, some big, and some small—in the following proportions.

	blue	red
big	10	30
small	50	10

A marble is chosen at random and placed under a cup; no other information about the situation is available.

In Yalcin's scenario, it is reasonable to accept the premises, but not the conclusion, of this instance of modus tollens.

1. If the marble is big then it's likely red.
2. The marble is not likely red.

∴ The marble is not big.

2.3 *Conditional Proof*

Conditional proof (sometimes called the *deduction theorem* in formal logic) lets us establish conditional conclusions without relying on any conditional assumptions. Suppose that an argument from the premises X and A to the conclusion C is valid. Then conditional proof lets us conclude that the argument from X to 'if A, then C' is valid. (Unlike modus ponens and modus tollens, which let us reason from the truth of some propositions to the truth of another proposition, conditional proof lets us reason from the validity of one argument to the validity of another.)

Stalnaker (1975) gives an argument that can easily be worked into a counterexample to conditional proof (though he does not present it that way). The following argument is valid, since in classical logic, anything follows from a contradiction:

1. The butler did it.
2. The butler didn't do it.

∴ The gardener did it.

But the following argument is not valid:

1. The butler did it.

∴ If the butler didn't do it, then the gardener did it.

Although conditional proof in its full generality looks implausible, a restricted version is more appealing: if A all by itself entails C, then 'if A, then C' is a truth of logic. (Koons (2014) makes a similar suggestion about conditional proof in nonmonotonic logic.)

2.4 *Transitivity, Contraposition, and Strengthening the Antecedent*

Transitivity is the inference form:

1. If A, then B.
2. If B, then C.

∴ If A, then C.

Contraposition is:

1. If A, then C.

$\therefore$ If not C, then not A.

And strengthening the antecedent is:

1. If A, then C.

$\therefore$ If A and B, then C.

All three inference forms seem to fail for ordinary conditionals in English. For transitivity, we have the following counterexample (Stalnaker, 1968, p. 106):

1. If J. Edgar Hoover had been born a Russian, then he would have been a communist.
2. If J. Edgar Hoover had been a communist, then he would have been a traitor.

$\therefore$ If J. Edgar Hoover had been born a Russian, then he would have been a traitor.

For contraposition, we have the following counterexample (adapted from Adams, 1988):

1. If it rains, then it does not rain hard.

$\therefore$ If it rains hard, then it does not rain.

And for strengthening the antecedent, we have the following counterexample (Stalnaker, 1968, p. 106):

1. If this match were struck, then it would light.

$\therefore$ Therefore, if this match had been soaked in water overnight and it were struck, then it would light.

Not everyone accepts these putative counterexamples as genuine. Brogaard and Salerno (2008) argue that the meaning of a conditional depends partly on a contextually determined set of relevant possible worlds. They claim that the putative counterexamples involve a context shift between the premises and the conclusion, but in any fixed context, the arguments are valid.

Fintel (2001), Gillies (2007), and Williams (2008) cite linguistic evidence in support of the context shift hypotheses: changing the order of the premises and conclusions in the counterexample arguments changes whether they seem true or false. Counterexamples to antecedent strengthening are closely related to so-called *Sobel sequences* (named for Sobel 1970). A Sobel sequence consists of two sentences of the following form (Gillies, 2007).

(a) If Sophie had gone to the New York Mets Parade, she would have seen Pedro Martínez.

(b) But if Sophie had gone to the New York Mets Parade and gotten stuck behind a tall person, she would not have seen Pedro Martínez.

It seems perfectly reasonable to assert (a) followed by (b). But once someone has asserted (b), an assertion of (a) seems inappropriate—after all, if Sophie had gone to the parade, who's to say she would not have gotten stuck behind a tall person?

Fintel, Gillies, and Williams claim that Sobel sequences involve a context shift: once someone asserts (b), the context changes to make (a) false, but (a) and (b) are never true in the same context. Moss (2012) proposes an alternative explanation: once (b) has been asserted, (a) might be true, but is no longer known, since asserting (b) changes the standards a belief must meet in order to count as knowledge.

2.5 *Simplification of Disjunctive Antecedents*

Simplification of disjunctive antecedents ('simplification' for short; Nute, 1975) is the argument form:

1. If A or B, then C.

$\therefore$ If A, then C.

Simplification seems appealing on its face: surely, to say that the bus will be late if it rains or snows is to say that the bus will be late if it rains, and the bus will be late if it snows.

However, one can easily generate counterexamples by substituting the same sentence for B and C. Suppose I have enough money to visit either Disneyland or Graceland, but not enough to visit both. Then the premise of the following argument is true, while its conclusion is false.

1. If I visit Disneyland or I visit Graceland, then I'll visit Graceland.

$\therefore$ If I visit Disneyland, then I'll visit Graceland.

Counterexamples to strengthening the antecedent can be used to generate counterexamples to simplification (Fine, 1975). Suppose we have both of the following:

1. If A, then C.
2. Not: if A and B, then C.

A is logically equivalent to [(A and B) or (A and not B)], so by 1, we have:

3. If [(A and B) or (A and not B)], then C.

But by simplification, the truth of 3 would have to entail the falsity of 2.

So there is a three-way tension between the validity of simplification, the invalidity of strengthening the antecedent, and the substitution of

logical equivalents. All three ways out of the puzzle are represented in the literature: Loewer (1976) and Mckay and Inwagen (1977) reject simplification; defenders of strict conditional accounts (Section 4.1) accept strengthening the antecedent; and Nute (1975) and Alonso-Ovalle (2009) reject substitution.

3 THE INDICATIVE/COUNTERFACTUAL DISTINCTION

Conditionals in English can be divided into two categories, exemplified by the following pair of sentences (Adams, 1970):

(DD) If Oswald did not shoot Kennedy, then someone else did.

(HW) If Oswald had not shot Kennedy, then someone else would have.

Although (DD) and (HW) are built up from the same antecedent and consequent, they mean different things. (DD) would be acceptable to most people familiar with US history: Kennedy was shot, so someone must have shot him—if not Oswald, then someone else. But (HW) is more controversial; it is accepted by conspiracy theorists, but rejected by those who believe that Oswald acted alone. Sentences like (DD) are called *indicative*; sentences like (HW) are called *counterfactual* (or sometimes *subjunctive*).

It's not clear how to classify conditionals whose antecedents concern the future. Consider the following sentence, as uttered by a conspirator before the Kennedy assassination.

(DW) If Oswald does not shoot Kennedy, then someone else will.

Dudman (1983, 1984) and Bennett (1988) argue that future-tensed conditionals like (DW) belong with counterfactuals like (HW); Bennett (2003, 2001; yes the same Bennett!) argues that they belong with indicatives like (DD); Edgington (1995) argues that there exist distinct categories of future-tensed indicatives and future-tensed counterfactuals.

Philosophers also disagree about the precise relationship between indicatives and counterfactuals. Some favor what Bennett (2003) calls 'Y-shaped analyses', which first explain what is common to indicatives and counterfactuals, and then bifurcate to explain how this common core can produce two different kinds of conditionals. Others (notably Gibbard, 1981, and Bennett, 2003) argue that we need completely separate theories of indicatives and counterfactuals—that there is no interesting core shared by both.

In what follows, I will write '$A \mathrel{\square\!\!\rightarrow} C$' to indicate a counterfactual conditional; '$A \rightarrow C$' to abbreviate an indicative conditional; and 'if A, then C' where I wish to remain neutral. I turn now to a popular class of theories, typically aimed at explaining counterfactual conditionals, but sometimes extended to cover indicatives.

4 SELECTION FUNCTIONS

One way to give a theory of conditionals is to spell out their *truth conditions*, i.e., the circumstances under which they are true. Formally, philosophers represent the truth conditions of a sentence as a function from possible worlds (i.e., ways the world might be) to truth values. Fully specifying the truth conditions for every conditional would be too tall an order: to understand the truth conditions for 'if ontogeny recapitulates phylogeny, then snakes develop vestigial legs', we would have to understand the truth conditions of 'ontogeny recapitulates phylogeny' and 'snakes develop vestigial legs', and that job falls outside the scope of a theory of conditionals. So theories of conditionals adopt a more modest aim: to give a recipe for deriving the truth conditions for 'if A, then C' from the truth conditions of (arbitrary) A and C.

The concept of a selection function (Stalnaker, 1968) provides a way of assigning truth conditions to a conditional based on the truth conditions of its antecedent and consequent. The basic idea is that, to evaluate 'if A, then C', we should first consider a set of *selected* possible worlds where A is true. (Henceforth, I will use 'A-worlds' as shorthand for 'worlds where A is true'.) Intuitively, the selected worlds represent ways the actual world might be if A were true. We then check whether, at all the selected worlds, C is true. If so, then the counterfactual conditional 'if A, then C' is true at the actual world; otherwise, it is false at the actual world.

More formally, we can model this process in terms of a selection function f that maps ordered pairs consisting of a possible world and a proposition onto sets of possible worlds. 'If A, then C' is true at a possible world w if and only if C is true at every world in $f(A, w)$. Different ways of interpreting the selection function yield different theories of conditionals.

4.1 *Strict Conditionals*

One natural way to interpret the selection function is to check *all* possible A-worlds, and say that 'if A, then C' is true at world w just in case C is true at all of them. (Since what is possible may depend on what is actual, the truth value of the conditional may vary from world to world.) This approach yields the *strict conditional* interpretation of the selection function, first developed by C. Lewis (1918). The strict conditional approach classifies transitivity, contraposition, and antecedent-strengthening as valid—which its opponents claim is a mistake (see D. Lewis, 1973a, pp. 4–12).

The strict conditional interpretation also gives questionable results about which counterfactuals are true. If I were to leap out of the second-story window of my office, I would hurt myself—but the strict conditional account says this is not so. There are possible worlds where I leap out the

second-story window and remain unharmed: some where there is a safety net underneath the window, some where I am thoroughly ensconced in protective bubble wrap, some where my body is much less fragile than ordinary human bodies, some where the Earth's gravitational field is weak... but none of them is the sort of world that would result, if I were to leap out the second-story window. Hájek (manuscript) sums up the problem this way: on the strict conditional interpretation, most counterfactuals are false.[2]

4.2 *Closest Worlds*

An alternative to the strict conditional approach, typically used for counterfactuals, defines the selection function in terms of similarity among possible worlds. For every world w, we can rank worlds from most similar to w ('closest') to least similar ('farthest away'). D. Lewis (1973a) holds that every such ranking is a *total preorder*: two worlds can be equally similar to w, but they must be comparable, so that either they are equally similar or one is more similar than the other. (Stalnaker, 1968, discusses the special case of the logic where no two worlds are equally close to each other; Pollock, 1976, discusses a generalization where worlds may be incomparable in terms of closeness.) $A \boxrightarrow C$ is true at w just in case C is true at all the A-worlds that are most similar to w.

Formally, the closest-worlds interpretation can be modeled using a system of 'spheres'—sets of worlds such that every world in the set is closer to w than every world outside it (D. Lewis, 1973a). Then $f(A, w)$ is the intersection of the set of A-worlds with the smallest sphere containing at least one A-world.[3]

Unlike the strict conditional interpretation, the closest-worlds interpretation of the selection function can explain why transitivity, contraposition, and antecedent-strengthening seem invalid. On the closest-worlds interpretation, they *are* invalid, and we can use diagrams (adapted from D. Lewis, 1973a) to illustrate why.

To see why transitivity is invalid, consider a system of spheres model centered on a particular world w, depicted in Figure 1a. (Worlds are points in the diagram, and spheres are concentric circles.) The A-worlds are the points inside the shape labeled A, the B-worlds are the points inside

2 Hájek argues that the problem extends beyond strict conditional accounts; it also affects the closest-worlds account in Section 4.2. K. S. Lewis (2015) argues that we can save the closest-worlds account by ignoring worlds that are deemed irrelevant by a contextually-determined standard of relevance.

3 Some technical difficulties arise when there is no smallest sphere containing at least one A-world, but only a limitless sequence of ever-smaller spheres; see D. Lewis (1973b, pp. 424–425); Stalnaker (1981, pp. 96–99); Warmbrod (1982); and Díez (2015) for discussion.

the shape labeled *B*, and the *C*-worlds are the points inside the shape labeled *C*. All the closest *A*-worlds to *w* are *B*-worlds, and all the closest *B*-worlds are *C*-worlds; yet none of the closest *A*-worlds are *C*-worlds. Figure 1b shows a counterexample to contraposition, and Figure 1c shows a counterexample to antecedent strengthening.

Defenders of the closest-worlds theory have the burden of spelling out what 'closeness' amounts to. D. Lewis (1973a) claims that closeness is based on similarity among worlds: to say that one world is closer to *w* than another is to say that the first world is more similar to *w* than the second. But Fine (1975) presents an example where greater similarity does not make for greater closeness. (I have taken a few liberties with the details of the example.)

On September 26, 1983, at the height of the Cold War, a Soviet early-warning system went off, falsely reporting that missiles had been launched at Russia from the US (Aksenov, 2013). The officer who saw the alarm, Stanislav Petrov, did not report it to his superiors, and so Russia did not launch missiles in retaliation. The following conditional seems true:

PETROV If Petrov had informed his superiors at the time of the false alarm, then there would have been a nuclear war.

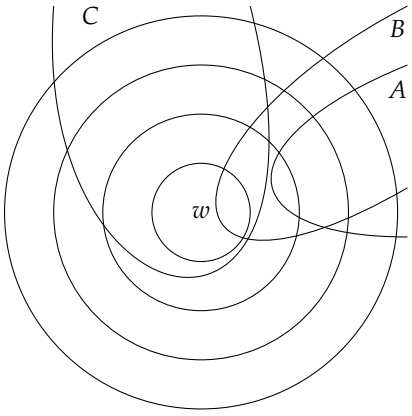
After all, Petrov's superiors were poised to launch the missiles in the event of an attack, and it seems that the phone lines and missile system were in working order. The only missing ingredient was the report from Petrov.

But among the worlds where Petrov informs his superiors at the time of the false alarm, those where the Soviet missile launch is prevented by a happy accident—incompetence by Petrov's superiors, or a broken telephone, or a malfunction of the Soviet missile system—are more similar to the actual world than those where the launch goes through. Worlds where the missile launch is prevented by a happy accident agree with the actual world about the total number of nuclear wars in the 20th Century—surely a more important dimension of similarity than the functioning or malfunctioning of one measly telephone line.[4]

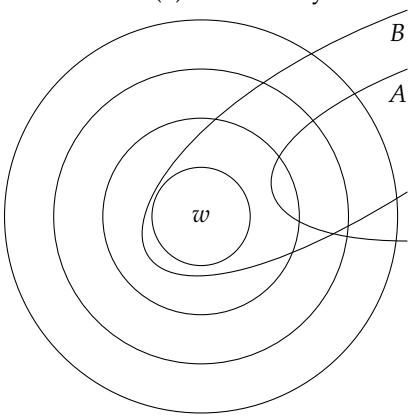
4.3 *Past Predominance*

To handle the PETROV example, a natural thought goes, we need an account of the selection function that treats the past differently from the future. When Petrov made his choice, the missile launch system was already in

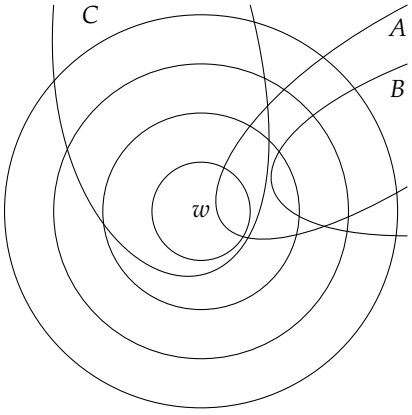
4 Defenders of the closest-worlds interpretation reply that we should understand 'similarity' so that agreeing about the total number of nuclear wars in the 20th Century does not make for greater similarity than agreeing about the functioning or malfunctioning of one measly telephone line; see D. Lewis (1979) and Arregui (2009).



(a) Transitivity



(b) Contraposition



(c) Strengthening the antecedent

Figure 1: Counterexamples to transitivity, contraposition, and strengthening the antecedent in the closest-worlds framework

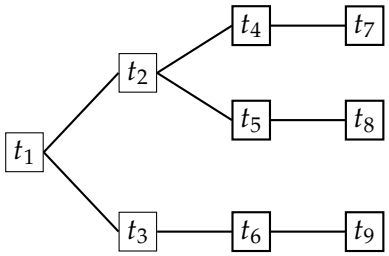


Figure 2: A model of branching time

working order—but it was not yet determined whether there would be a war.

Thomason and Gupta (1980) propose an account of the selection function that takes seriously the past-future asymmetry. They model the universe using branching time, where each moment has only one possible past, but multiple possible futures. (Cross, 1990, shows that the assumption of branching time is dispensable; past predominance can also be modeled using ordinary possible worlds.) Figure 2 depicts such a model. The nodes t_1, t_2, ..., t_9 are moments. Paths through the tree—in this example, $\{t_1, t_2, t_4, t_7\}$, $\{t_1, t_2, t_5, t_8\}$, and $\{t_1, t_3, t_6, t_9\}$—are called *histories*.

We can think of each possible world as containing information about which moment is present, as well as information about which history is actual. (On this way of understanding the model, even when the present moment has more than one possible future, there is a fact of the matter about which future will occur.)

Thomason and Gupta adopt a *past predominance* principle, which says that if a world x is in $f(A, w)$, then it must diverge from w as late as possible—there can be no other A-world whose history overlaps w for a longer span than $f(A, w)$.[5]

The past predominance view can explain the PETROV example. Consider the following interpretation of our diagram: the actual history is $\{t_1, t_2, t_4, t_7\}$. At t_1, it is not yet settled whether the early warning system goes off. The early warning system goes off t_2, and Petrov must decide what to do. (At t_3, which belongs to an alternative history, there is never any alarm.) At t_4, Petrov decides not to notify his superiors, and so at t_7, there is no nuclear war. (At t_5, which belongs to another alternative history, Petrov decides to notify his superiors, and a nuclear war ensues at t_8.)

Now consider the conditional PETROV, as uttered at t_7. Its antecedent is false at the actual world, which has the history $\{t_1, t_2, t_4, t_7\}$. The closest

5 For technical reasons, Thomason and Gupta also assume that $f(A, w)$ is a singleton set, and posit that each world contains a *choice function*, which specifies not just what the future will be like, but what the future would have been like had the past gone differently. I pass over the details.

worlds where its antecedent is true must have the history $\{t_1, t_2, t_5, t_8\}$, which diverges from the actual world's history at the last possible moment yet still makes the antecedent true. Since the actual present moment is t_7, it seems reasonable to select t_8 as the present moment at all the closest worlds. Since there is a nuclear war at t_8, the consequent of PETROV is true at all the the closest worlds; hence PETROV is true at the actual world.

4.4 *Causal Models*

A class of examples called *Morgenbesser cases* (Slote, 1978, 27n) suggest that the selection function should respect causal as well as temporal constraints. Edgington (2004) gives a representative Morgenbesser case.

Our heroine misses a flight to Paris due to a car breakdown. She complains to the repairman: 'If I had caught the plane, I would have been halfway to Paris by now!' But he corrects her: 'I was listening to the radio. It crashed. If you had caught that plane, you would be dead by now.'

The repairman claims that the following counterfactual is true.

LETHAL If the heroine had caught that plane, she would be dead by now.

He is right. It's not clear that past predominance can explain why he's right: the plane crash occurs after our heroine would have made her flight.[6] What matters is that the plane crash is causally independent of whether she makes her flight. This is why, when assessing what would have happened if our heroine had made her flight, we should hold the plane crash fixed.

Pearl (2009) proposes a causal theory of counterfactuals that accounts for Morgenbesser cases. His theory relies on the concept of a *causal model*, consisting of a set of *variables*, which represent what circumscribed parts of the world are like, and a set of *structural equations*, which represent direct causal links between variables. Each variable is assigned an *actual value*; we can think of variables as questions about parts of the world, their possible values as possible answers to those questions, and their actual values as the correct answers in the actual world. Note that although I introduced selection semantics as a recipe for assigning truth values to conditionals at worlds, Pearl's theory is a recipe for assigning truth values to conditionals at model-valuation pairs.[7]

6 But see Phillips (2007) for an argument that past predominance *can* provide an adequate explanation.

7 Pearl's theory can be understood as a version of the situation semantics defended by Barwise and Perry (1981). Instead of assigning truth values to propositions at worlds, it assigns truth values to propositions at situations, which represent ways that circumscribed parts of the world could be.

We can understand Pearl's theory by first building a causal model of Edgington's plane example, then using the model to evaluate the conditional LETHAL. The model will include the following variables.

$$\textsc{car} = \begin{cases} 1 & \text{if the car is working,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\textsc{catch} = \begin{cases} 1 & \text{if our heroine catches her plane,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\textsc{crash} = \begin{cases} 1 & \text{if there is a crash,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\textsc{location} = \begin{cases} 0 & \text{if our heroine ends up stuck at the side of the road,} \\ 1 & \text{if our heroine ends up in Paris,} \\ 2 & \text{if our heroine ends up dead.} \end{cases}$$

CAR and CRASH are what Pearl calls *exogenous* variables; their values are determined by factors outside the model. CATCH and LOCATION are *endogenous* variables; their values are determined by the values of other variables in the model.

For each of the endogenous variables, the model specifies a structural equation. In the plane example, the structural equations are as follows.

$$\textsc{catch} = \textsc{car}$$

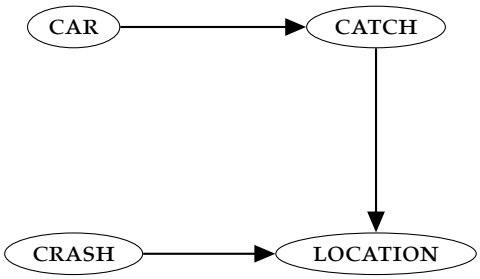
$$\textsc{location} = \begin{cases} 0 & \text{if } \textsc{catch} = 0, \\ 1 & \text{if } \textsc{catch} = 1 \text{ and } \textsc{crash} = 0, \\ 2 & \text{if } \textsc{catch} = 1 \text{ and } \textsc{crash} = 1. \end{cases}$$

(NB: the structural equations are asymmetric. The variable on the left-hand side has its value causally determined by the variables on the right-hand side.)

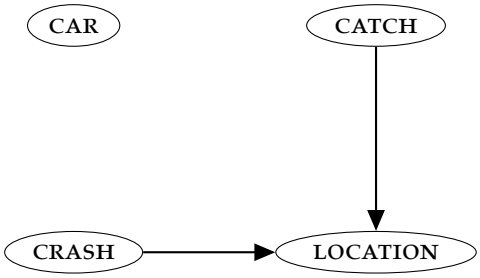
In the plane example, the variables take on the following values.

$$\begin{aligned} \textsc{car} &= 0, \\ \textsc{catch} &= 0, \\ \textsc{crash} &= 1, \\ \textsc{location} &= 0. \end{aligned}$$

We can summarize information about the variables and structural equations using the causal graph in Figure 3a. An arrow from one variable to



(a) The actual model



(b) The submodel generated by intervening to make the antecedent true

Figure 3: Causal graph used to evaluate the counterfactual LETHAL: 'If the heroine had caught that plane, she would be dead by now'.

another indicates that the first variable exerts direct causal influence on the second, but unlike the structural equations, the causal graph doesn't specify the nature of that influence.

Given a pair consisting of a model and an assignment of values to variables in the model, we can use a selection function to assign truth values to conditionals. (This time, the selection function takes in a model, and returns a singleton containing one new model.) Pearl's account is restricted to counterfactuals whose antecedents are either 'literals', which say that a particular variable takes on a particular value, or conjunctions of literals. (So 'the heroine's car breaks down and the plane crashes' is an acceptable antecedent, while 'the heroine's car breaks down or the plane crashes' is not.)

Where $\langle M, V\rangle$ is a model paired with an assignment of values to variables, and A is an antecedent with the appropriate form, we can generate a *submodel* $\langle M_A, V_A\rangle$ by 'intervening' on $\langle M, V\rangle$ to make A true. Intuitively, we can imagine an intervention as an action by someone outside the model who 'reaches in' to make the antecedent true, without tinkering with variables that are causally independent of the antecedent. For instance, a cabbie could intervene to set CATCH $= 1$ by driving our heroine to the airport regardless of whether or not her car has broken down.

Formally, the submodel M_A is a model with the same variables as M, but different structural equations. If X is one of the variables mentioned in

A, and X is endogenous, we delete the structural equation corresponding to X, and make X exogenous instead. (This corresponds to the idea that an intervention makes A true regardless of whether its typical causes obtain; the intervening cabbie enables the heroine to get to the airport whether or not her car is in working order.) We then set the value V_A of each X mentioned in A to the value specified by A. (This corresponds to the idea that the intervention makes the antecedent true.) If a variable is not causally influenced (either directly or indirectly) by any of the variables mentioned in the antecedent, then V_A assigns it the same value as V. (This corresponds to the idea that an intervention is *minimal*, so that only the variables mentioned in the antecedent are directly effected.) Finally, if a variable is causally influenced by one of the variables mentioned in the antecedent, then its value V_A is fixed by the structural equations. (This corresponds to the idea that an intervention is minimal in another sense: it does not interfere with the downstream effects of the variables mentioned in the antecedent.)

We are now ready to evaluate the counterfactual

LETHAL If the heroine had caught that plane, she would be dead by now.

in our original model. To check whether LETHAL is true in the original model, we intervene to make its antecedent true—i.e., to set CATCH $= 1$. We then check whether the consequent is true (i.e., LOCATION $= 2$) in the resulting submodel.

First, we delete the structural equation for CATCH, turning CATCH into an exogenous variable. Our only remaining structural equation is

$$\text{LOCATION} = \begin{cases} 0 & \text{if CATCH} = 0, \\ 1 & \text{if CATCH} = 1 \text{ and CRASH} = 0, \\ 2 & \text{if CATCH} = 1 \text{ and CRASH} = 1. \end{cases}$$

(The graph for the resulting submodel is shown in Figure 3b.)

Second, we set the values of the variables. The antecedent requires that

$$\text{CATCH} = 1.$$

Since neither CAR nor CRASH is downstream from CATCH, we have

$$\text{CAR} = 0,$$
$$\text{CRASH} = 1.$$

Finally, the value of LOCATION is fixed by the structural equation. Since CATCH $= 1$ and CRASH $= 1$, we have

$$\text{LOCATION} = 2.$$

Therefore, in the submodel, the protagonist is dead, so in the original model, had she caught her plane, she would have been dead.

The procedure described is a type of selection semantics: given an antecedent and a model-valuation pair, we call on a 'submodel' selection function that returns the singleton set of another model-valuation pair (a submodel). Galles and Pearl (1998) argue that this selection semantics is formally equivalent to the closest-worlds account. However, there is a key difference between the two accounts: the selection semantics lets us assigns truth conditions to counterfactuals built up from arbitrary sentences, while the causal modeling account only lets us assign truth values to counterfactuals whose antecedents are literals, or conjunctions of literals. Schulz (2011) and Briggs (2012) propose ways of extending the language to counterfactuals with logically complex antecedents; their proposed theories are logically inequivalent to the closest-worlds semantics. Huber (2013) proposes an alternative way of extending the language that makes it logically equivalent to the closest-worlds account.

5 COUNTERPOSSIBLE CONDITIONALS

Selection semantics has trouble with *counterpossible* conditionals—that is, conditionals whose antecedents are impossible. It counts all counterpossible conditionals as trivially true. Where A is impossible, there are no possible A-worlds. Therefore, if we feed the selection function an impossibility A and a world w and ask it to return a set of possible A-worlds, it returns the empty set. Trivially, all the A-worlds in the empty set are C-worlds, so that trivially $A \mathrel{\Box\!\!\rightarrow} C$ is true in the original world.

But counterpossibles seem to have non-trivial truth conditions: some are true, while others are false. Examples of true counterpossibles include:

> If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would not have cared (Nolan, 1997, p. 544).
>
> If I were a horse, then I would have hooves (Krakauer, 2012, p. 10).
>
> If wishes were horses, beggars would ride (Krakauer, 2012, p. 10).
>
> If intuitionistic logic were the correct logic, then the law of excluded middle would no longer be unrestrictedly valid (adapted from Brogaard & Salerno, 2013).

Corresponding examples of false counterpossibles include:

> If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would have taken notice.

If I were a horse, then I would have scales.

If wishes were horses, no one would own any horses.

If intuitionistic logic were the correct logic, then the law of excluded middle would still be unrestrictedly valid.

Assigning non-trivial truth values to counterpossibles doesn't just capture linguistic intuitions; it also enables counterpossibles to do valuable philosophical work. Non-trivial counterpossibles help us assess rival philosophical, mathematical, and logical theories by telling us what would follow if those theories were true (Krakauer, 2012; Brogaard & Salerno, 2013; Nolan, 1997). They explain how necessary events and omissions of impossible events are causally relevant to the actual world—how a mathematician's failure to disprove Fermat's Last Theorem prevented her from getting tenure, how my failure to be in two places at once caused me to miss a colloquium talk, or how the copresence of a mental property and its subvening physical property can result in a subject's raising his arm (Bernstein, 2016). They can be used to give an account of essences: an essential property is one such that, if the bearer had lacked it, then the bearer would not have existed (Brogaard & Salerno, 2013, forthcoming). (Non-trivial counterpossibles save this account from certain implausible commitments—e.g., that living in a world where $2 + 2 = 4$ is trivially part of everyone's essence.)

Not everybody agrees that counterpossibles have non-trivial truth values, however. Williamson (2007, p. 172) argues that apparent examples of non-trivial counterpossibles collapse under closer scrutiny. In a slight variant on Williamson's example,[8] imagine that a student is mulling over a graded arithmetic test. Of the 12 problems on the test, the student has gotten the last one wrong: 'what is $5 + 7$?' The student, who answered '11', laments: 'If only $5 + 7$ were 11, I would have gotten a perfect score!' This seems to be true, and furthermore, it seems false that if $5 + 7$ were 11, the student would have gotten one of the problems wrong. But appearances are deceptive. Suppose that $5 + 7$ were 11. Then in answering all the problems right, the student would have given five right answers followed by seven more right answers, for a total of 11 right answers. Since there are 12 problems on the test, the student would have gotten one problem wrong after all. (For an extended rebuttal of Williamson's argument, see Salerno and Brogaard, forthcoming.)

8 Thanks to Sharon Berry for suggesting this version in conversation.

5.1 *Impossible Worlds*

Nolan (1997) gives an account of counterpossibles by supplementing the closest-worlds account with impossible worlds—ways the world couldn't be. We can then say that $A \boxrightarrow B$ is true at w just in case B is true at all the closest possible or impossible A-worlds to w. Two questions then arise: what are impossible worlds, and what makes them closer or further away from the actual world?

The ontology of impossible worlds has spawned its own literature: they may be collections of individuals like our actual world (Yagisawa, 2010), or they may be sets of sentences in some suitable language (Hintikka, 1975; Melia, 2001; Sider, 2002; see Berto, 2013, for a general overview and discussion.) Another pressing question for theorists of counterpossibles concerns the logical structure of impossible worlds. Is it the case that for every set of sentences, there is some impossible world where all and only the sentences in the set are true? Or is there more logical structure we can impose on impossible worlds?

Proponents of impossible worlds typically don't require that the impossible worlds be closed under classical logical consequence—in other words, they don't require that whenever some propositions are true at an impossible world, all the classical logical consequences of those propositions are true at the world too. If impossible worlds had to be closed under classical logical consequence, then whenever A was impossible by the rules of classical logic, $A \boxrightarrow C$ would be trivially true. Nolan (1997, p. 547) argues that we should not require impossible worlds to be closed under any kind of logical consequence, since for every putative logical truth, there are non-trivial facts about what the world would be like if that logical truth did not obtain. A similar line of reasoning suggests that some impossible worlds have truth-value gluts: we can speculate about what would happen if there were true contradictions, so there must be impossible worlds at which there are true contradictions.

Bjerring (2013) argues that some impossible worlds have truth-value gaps. Otherwise, he argues, our theory of counterpossibles would misclassify certain conditionals as true, such as this one: 'If intuitionistic logic were correct, then the Law of Excluded Middle would hold.' (The Law of Excluded Middle says of every proposition that either it or its negation holds; intuitionists famously deny it.)

What about closeness? Nolan (1997) proposes the

Strangeness of Impossibility Condition Any possible world is more similar [closer] to the actual world than any impossible world (Nolan, 1997, p. 550).

The Strangeness of Impossibility Condition ensures that where A is a possible proposition, supplementing the closest-worlds account with impossible worlds has no effect on how we evaluate $A \boxrightarrow C$. So long as A is possible, the set of closest possible A-worlds coincides with the set of closest possible or impossible A-worlds.

Bjerring (2013, p. 348) proposes another constraint on closeness, which implicitly relativizes closeness to the antecedent of a counterfactual. Given a collection of logical systems $L_1, L_2, \ldots, L_n$, where L_1 is classical logic, and where W_{L_i} is the set of worlds deductively closed under L_i's entailment relation, Bjerring endorses the

Relative Closeness Condition For any counterfactual whose antecedent presupposes that some logic L_i is correct (true, adequate), a world in modal space W_{L_i} is closer to the actual world than any world in modal space W_{L_j} , where $W_{L_i} \neq W_{L_j}$, and where $i \geq 1$ and $j > 1$.[9]

Brogaard and Salerno (2013) develop a theory on which impossible worlds are close to the actual world to the extent that they

1. minimize discrepancies with relevant background facts about the actual world (where the relevance of background facts is fixed by context), and

2. minimize violations of relevant *a priori* entailment (where relevant *a priori* entailment is spelled out in more detail in the paper).

As an illustration of these conditions, Brogaard and Salerno use them to evaluate the counterpossible conditional 'if water had not been H_2O, then water would have been a monkey'. This counterpossible is false. Their theory delivers the correct verdict, they claim, because it is *a priori* that water is not a monkey.

To derive this verdict, they consider two impossible worlds where the antecedent is true. At w_1, water is some chemical compound XYZ (different from H_2O), while at w_2, water is a monkey.

w_1	w_2
water is not H_2O	water is not H_2O
water is XYZ	water is a monkey

9 As stated by Bjerring, the Relative Closeness Condition seems to presuppose that W_{L_i} and W_{L_j} do not intersect. We can get rid of this presupposition by modifying the condition slightly:

Relative Closeness Condition* For any counterfactual whose antecedent presupposes that some logic L_i is correct (true, adequate), a world in modal space W_{L_i} is closer to the actual world than any world outside W_{L_i}.

Since there are more *a priori* truths that hold at w_1 than at w_2, and since both agree with the actual world about the same number of propositions, w_1 is closer to the actual world than w_2. (Brogaard and Salerno tacitly assume that there are no antecedent worlds closer to the actual world than w_1 or w_2.) Thus, at least one of the closest impossible worlds where water is not H_2O is one where water fails to be a monkey, so the conditional is false at the actual world.

5.2 *Relevant Logic*

Relevant logics are motivated by the thought that the conditional 'if A, then C' claims that the truth of A is connected to the truth of C. Relevant logics originated as rivals to the material conditional account, on which the conditional 'if A, then C' is true just in case A is false or C is true (see Section 6). However, some of the same intuitions that favor relevant logics over the material conditional account also favor them over the closest-worlds account. After all, the reason it seems wrong to say 'if Hobbes had squared the circle, sick children in the mountains of South America would have cared' is that there is no connection between Hobbes's squaring the circle and the interests of sick South American children. Likewise, the reason it seems right to say 'if I were a horse, I would have hooves' is because something's being a horse is connected to its having hooves.

Relevant logics are often characterized in proof-theoretic terms. But Routley and Meyer (1973, 1972a, 1972b) develop a versatile semantics for the conditionals of relevant logics, which generalizes the strict conditional semantics of Section 4.1. Recall that on the strict conditional interpretation, $A \boxdot\!\!\rightarrow C$ is true at w just in case C is true at all possible A-worlds (relative to w). We can rewrite the selection function in terms of a two-place accessibility relation among worlds: we say that Rwx just in case world x is possible according to world w, and that $f(A, w)$ is the set of all A-worlds x such that Rwx.

Routley and Meyer interpret the conditional in terms of a three-place accessibility relation among worlds. 'If A, then C' is true at w just in case, for all worlds x and y such that $Rwxy$ and x is an A world, C is true at y. Different restrictions on relation R generate different relevant logics. (For some logics, we need impossible worlds where both a sentence and its negation fail to be true, or impossible worlds where both sentence and its negation are true.)

This three-place R relation is formally useful, but does it mean anything? Beall et al. (2012) propose three interpretations of $Rwxy$, which spring

from different ways of grouping w, x, and y.[10] All three interpretations can be illustrated with the conditional

THERMITE If you light a bucket of thermite with a titanium fuse, then a huge explosion will ensue.

GROUPING THE SECOND AND THIRD WORLDS TOGETHER: $Rw\langle xy\rangle$. 'If A, then C' says at the actual world w, there are no counterexamples where A is true and C is false. We typically think of counterexamples as involving a single world which makes some things true and other things false, but relevant logicians split the labor between two worlds x and y, so that whatever holds at x is true, while whatever fails to hold at y is false. In the example of THERMITE, we might think of potential counterexamples as divided into an earlier part x, when a bucket of thermite may or may not be lit with a titanium fuse, and a later part y, when there may or may not be an explosion. If the actual world w admits some possible two-part scenarios that begin with the lighting of thermite with a titanium fuse, but fails to end in a huge explosion, then these scenarios are counterexamples that falsify THERMITE.

GROUPING THE FIRST AND SECOND WORLDS TOGETHER: $R\langle wx\rangle y$. 'If A, then C' says that using one's current information to draw inferences from A will yield the information that C. To say that $Rwxy$ is to say that when the rules of w are applied to the information in x, it is possible to infer y (or some information that entails y). In the case of THERMITE, we can imagine w as a parcel of information specifying the actual laws of nature, and x as another parcel of information specifying that a bucket of thermite has been lit with a titanium fuse. If sticking these parcels of information together licenses the conclusion that there has been a huge explosion (and does so no matter how we fill in x, the information that the thermite has been lit), then the conditional THERMITE is true.

GROUPING THE FIRST AND THIRD WORLDS TOGETHER: $Rw\rangle x\langle y$. 'If A, then C' says that C is necessary relative to A, or that C is necessary in an A-ish way. The conditional THERMITE does not say it is absolutely necessary that a huge explosion will ensue. The world w may permit a possible scenario y in which no huge explosions occur. However, once we enrich w with some additional information x, specifying that a bucket of thermite has been lit with a titanium fuse, we can consider what is possible under that supposition. If there is some way of filling in the antecedent

10 For a discussion of other ways of interpreting the ternary relation, with references, see Jago (2013).

that makes y a possibility, then y is possible not just absolutely, but under the supposition that the antecedent of THERMITE is true.

Mares and Fuhrmann (1995) propose a theory of counterfactuals that combines the closest-worlds interpretation of the selection function with the relevant interpretation of the conditional: $A \boxdot\!\!\rightarrow B$ is true at a world w just in case the relevant conditional 'if A, then B' is true at all closest A-worlds to w. Mares (1994) argues that this theory has useful applications to conditional analyses of causation, and to theories of conditional obligation.

6 THE MATERIAL CONDITIONAL ACCOUNT OF INDICATIVES

According to the material conditional account defended by Grice (1989) and Jackson (1987), an indicative conditional $A \rightarrow C$ is true just in case either its antecedent A is true, or its consequent C is false. (The material conditional account is almost always offered as a theory of indicative conditionals alone, since counterfactual conditionals with false antecedents can be false. Even though I don't keep a horse, it is false that if I were to keep a horse, it would breathe fire.) The material conditional account has a simple explanation for the apparent validity of all the the inferences discussed in Section 2 (modus ponens, modus tollens, conditional proof, strengthening the antecedent, transitivity, contraposition, and simplification): these inferences really are valid.

Furthermore, there are persuasive arguments for the conclusion that an indicative conditional $A \rightarrow C$ is true if and only if the corresponding material conditional 'not A or C' is true. Suppose the indicative conditional is true. Then it can't have a true antecedent and a false consequent; that would be a violation of modus ponens. So the indicative conditional entails the material conditional. But when I know that either C holds or A doesn't, I can infer that if A, then C. So the material conditional entails the indicative. (Stalnaker, 1975, p. 136, calls this the direct argument.) Since the material and indicative conditionals entail each other, they must be equivalent.

Gibbard (1981) provides a formal argument for the equivalence of the indicative and material conditionals based on three logical principles. Where 'not A' is abbreviated $\neg A$ and 'A or B' is abbreviated $A \vee B$, the principles are:

PSEUDO MODUS PONENS $A \rightarrow C$ entails $\neg A \vee C$.

IMPORT-EXPORT $A \rightarrow (B \rightarrow C)$ is equivalent to $(A \wedge B) \rightarrow C$.

CONDITIONAL PROOF If A entails C, then $A \rightarrow C$ is a logical truth.

To show that $A \rightarrow C$ and $\neg A \vee C$ are equivalent, Gibbard only needs to show that each entails the other. By Pseudo Modus Ponens, $A \rightarrow C$ entails $\neg A \vee C$. The proof that $\neg A \vee C$ entails $A \rightarrow C$ is as follows.

1. $((\neg A \vee C) \wedge A)$ entails C. (By tautological reasoning.)
2. It is a truth of logic that $((\neg A \vee C) \wedge A) \rightarrow C$. (By 1 and Conditional Proof.)
3. It is a truth of logic that $(\neg A \vee C) \rightarrow (A \rightarrow C)$. (By 2 and Import-Export.)
4. It is a truth of logic that $\neg(\neg A \vee C) \vee (A \rightarrow C)$. (By 3 and Pseudo Modus Ponens.)
5. $(\neg A \vee C)$ entails $(A \rightarrow C)$. (By 4 and tautological reasoning.)

Despite these points in its favor, the material conditional account faces substantial difficulties. It seems to yield wrong predictions about logical validity, often called 'paradoxes of material implication'.

For example,the material conditional account entails that all of the following are truths of logic:

> Either the unburied dead will walk the Earth if I bury a chicken head in my backyard, or the unburied dead will walk the Earth if I fail to bury a chicken head in my backyard (McGee, 2005).
>
> Either you are virtuous if you are rich, or you are rich if you are virtuous.
>
> One of these three things holds: if you grant voting rights to children, you will grant them to guinea pigs; if you grant voting rights to guinea pigs, you will grant them to inanimate objects; or if you grant voting rights to inanimate objects, you will take them away from adult human beings.

Furthermore, the material conditional account entails that all of the following inferences are valid. (The proof of God's existence is due to Edgington, 1986.)

1. I will not do my chores today.

$\therefore$ If I do my chores today, then the world will implode.

1. Dinner will be delicious.

$\therefore$ If I burn the veggie burgers and pour sand into the sweet potatoes, then dinner will be delicious.

1. If God does not exist, then it's not the case that if I pray, my prayers will be answered.
2. I do not pray.

$\therefore$ God exists.

In addition to yielding bad predictions about validity, the material conditional account yields bad predictions about the probabilities of conditionals. Suppose I draw a card at random from a 52-card deck. The material conditional 'either I do not draw a red ace, or I draw the ace of hearts' has probability 51/52. (The only way for me to make it false is to draw the ace of diamonds.) Therefore, by the material conditional account, I should assign probability 51/52 to the indicative conditional 'if I draw a red ace, then it will be the ace of hearts'. But the indicative conditional 'if I draw a red ace, then it will be the ace of hearts' should get probability 1/2, since half the time when I draw a red ace, it will be an ace of hearts.

More generally, the material conditional account falls afoul of

THE THESIS Whenever A and C are propositions, the probability of the indicative conditional $A \rightarrow C$ is equal to the conditional probability of C given A, understood as

$$Pr(A|C) = \frac{Pr(A \wedge C)}{Pr(C)}.$$

THE THESIS is a plausible way of unpacking the so-called *Ramsey test*, based on a famous remark by Ramsey (1978, 143n):

> If two people are arguing 'If p will q?'; and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q; so that in a sense 'If p, q' and 'If p, [not q]' are contradictories.

Unfortunately, the material conditional account is straightforwardly incompatible with THE THESIS, and with the Ramsey test more generally. The probability that a material conditional is true is not, in general, the conditional probability of the consequent given the antecedent. (The probability of the material conditional may be anywhere between that conditional probability and 1.) Furthermore, where A is highly unlikely, the material conditional 'not A or C' is both highly believable and highly assertible, whether or not adding A to one's stock of knowledge would justify a high degree of confidence in C.

Grice (1989) and Jackson (1987) explain these wrong predictions by distinguishing between true sentences and sentences that can appropriately be asserted. According to Grice, in a situation where I will not do my chores today, it is technically true that if I do my chores today, then the

world will implode. Likewise, in a situation where dinner will be delicious, it is technically true that if I burn the veggie burgers and pour sand into the sweet potatoes, then dinner will be delicious. Nonetheless, it is misleading to assert a conditional when I know that its antecedent is false, or when I know that its consequent is true, because it is misleading to assert a weak claim when I could have asserted a stronger one. Refusing to assert the stronger claim is liable to mislead my audience into thinking that I do not know it. The supposedly paradoxical arguments are valid. When their premises are true, their conclusions may be bad, but this does not make their conclusions false.

Grice's proposed mechanism for explaining away the problem is useful in other domains: it can explain why some non-conditional assertions are misleading. For instance, if you ask where John is, and I know that he is in the library, it is misleading for me to reply 'He is either at the pub, or in the library.' A similar trick works for negated conjunctions, as an example by D. Lewis (1976) shows. If I point out a harmless mushroom that I plan to keep for myself, and remark 'You won't eat that and live', knowing that my assertion will prevent you from eating it, then I am guilty of misleading you, though what I say is technically true.

Jackson (1987) is not satisfied with Grice's explanation, since sometimes, it is all right to assert an indicative conditional even if you know that the antecedent is false, or the consequent is true. I know that Oswald killed Kennedy, but can nonetheless assert that if Oswald didn't kill Kennedy, someone else did. Jackson has a different explanation for why technically true conditionals might sound wrong. While the material conditional account captures the truth conditions of an indicative conditional, the meaning of 'if... then...' goes beyond its truth conditions. Built into the meaning of an English indicative conditional is the implication that it would still be appropriate to assert the material conditional, even if its antecedent were known. (Jackson calls this feature 'robustness'.) The Oswald-Kennedy conditional is robust, because even if I had reason to doubt that Oswald killed Kennedy, I would still have good reasons to believe that either Oswald or someone else killed him.

7 THE NO TRUTH VALUES (NTV) ACCOUNT OF INDICATIVES

Suppose you are convinced that the material conditional account gives the wrong truth conditions for the indicative conditionals. You might hope that there was some other account of the truth conditions for indicative conditionals—one that could better explain the truth of THE THESIS. Unfortunately, a collection of so-called 'triviality theorems' suggests that no truth conditions whatsoever will do the trick. Triviality theorems motivate Edgington (1986, 1995) and Appiah (1985) to claim that indicative condi-

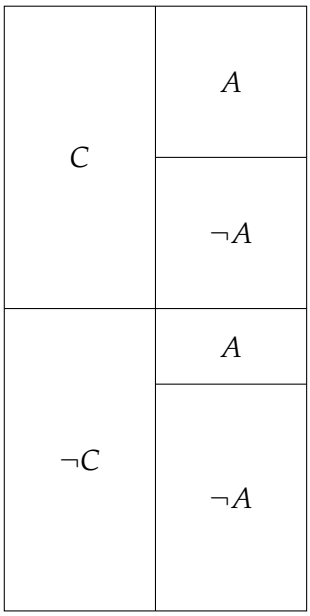


Figure 4: A probability space

tionals lack truth values altogether. (Edgington, 2008, goes on to develop a Y-shaped theory on which counterfactual conditionals also lack truth values altogether.)

In general, triviality theorems show that if THE THESIS is true in general, then every probability function is *trivial*: it assigns positive probability to at most two mutually exclusive alternatives. But it is absurd to claim that every probability function is trivial. (Here is a non-trivial probability function: the one that assigns probability $1/6$ to each possible outcome of the roll of a single die.) Therefore, we must reject THE THESIS.

To see how triviality theorems work, we can consider an early result by D. Lewis (1976), illustrated by system of diagrams adapted from Edgington (1995). Edgington visualizes probabilities using rectangles, divided into horizontal segments representing propositions. The height of a segment represents the probability of the corresponding proposition; the entire rectangle is normalized to have height 1. In Figure 4 the proposition C has probability $1/2$. C is subdivided into the propositions $A \wedge C$ (probability $1/4$) and $A \wedge \neg C$ (probability $1/4$). $\neg C$ (also with probability $1/2$) is subdivided into the propositions $\neg C \wedge A$ (probability $1/8$) and $\neg C \wedge \neg A$ (probability $3/8$).

Figure 5 shows how to calculate the probability of A conditional on C by erasing the bottom half of the diagram, and stretching out the remaining part of the rectangle so its height is 1 (in effect multiplying the height of each of its sub-regions by $\frac{1}{Pr(C)}$). The new height of the A region is $Pr(A|C)$.

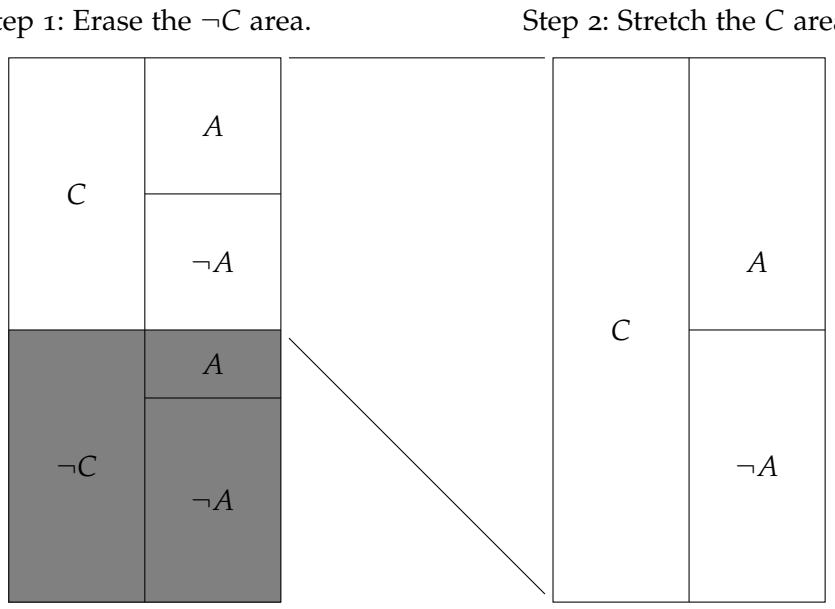


Figure 5: Calculating conditional probability

According to the Law of Total Probability (illustrated in Figure 6), for any two propositions X and Y,

$$Pr(Y) = Pr(Y|X) \times Pr(X) + Pr(Y|\neg X) \times Pr(\neg X). \quad (1)$$

Consider any two propositions A and C such that $P(A \wedge C) > 0$, and $P(A \wedge \neg C) > 0$. Plugging in A for X and $A \rightarrow C$ for Y in Equation 1 yields:

$$Pr(A \rightarrow C) = Pr(A \rightarrow C|C) \times Pr(C) + Pr(A \rightarrow C|\neg C) \times Pr(\neg C). \quad (2)$$

In other words, we can split the probability space into a C part and a $\neg C$ part, and figure out the probability of $A \rightarrow C$ by averaging its probabilities conditional on each part, a procedure illustrated in Figure 7. Consider the probability distribution Pr_C such that for all propositions X, $Pr_C(X) = Pr(X|C)$ (shown in the top center of Figure 7). Using the fact that $Pr_C(A) > 0$, the fact that $Pr_C(C) = 1$, and the definition of conditional probability, we can show that

$$Pr_C(C|A) = 1. \quad (3)$$

Thus, by THE THESIS and Equation 3,

$$Pr_C(A \rightarrow C) = 1. \quad (4)$$

By the definition of Pr_C and equation 4,

$$Pr(A \rightarrow C|C) = 1. \quad (5)$$

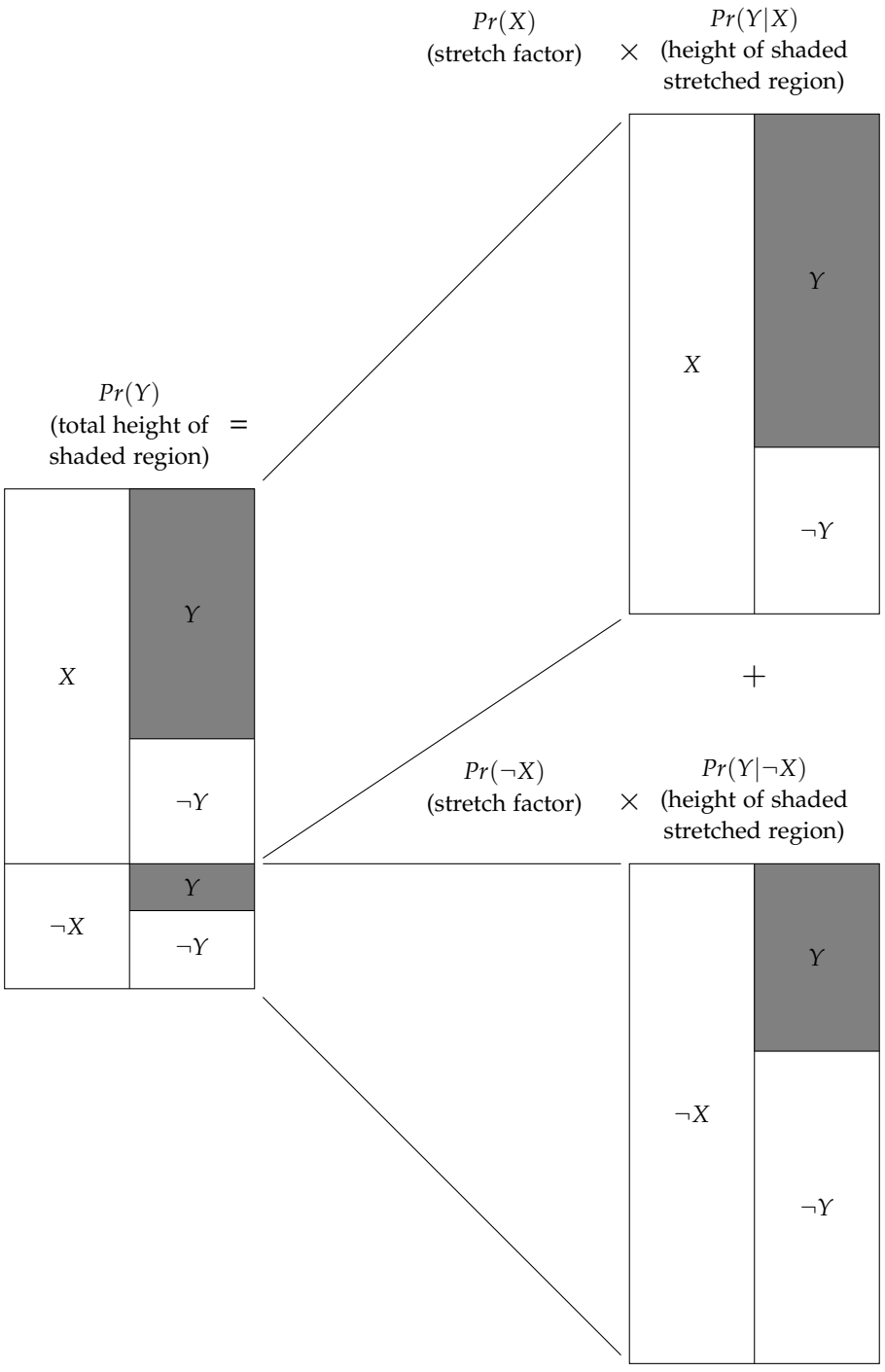


Figure 6: The Law of Total Probability

Likewise, when we consider the probability distribution $Pr_{\neg C}$ such that for all X, $Pr_{\neg C}(X) = Pr(X|\neg C)$ (shown in the bottom center of Figure 7), we see by the fact that $Pr_C(A) > 0$, the fact that $Pr_C(C) = 1$, and the definition of conditional probability that

$$Pr_{\neg C}(C|A) = 0. \tag{6}$$

Thus, by THE THESIS, and Equation 6,

$$Pr_{\neg C}(A \rightarrow C) = 0. \tag{7}$$

And by the definition of $Pr_{\neg C}$ and Equation 7,

$$Pr(A \rightarrow C|\neg C) = 0. \tag{8}$$

Using Equations 5 and 8 to make the appropriate substitutions into equation (2), we get:

$$Pr(A \rightarrow C) = 1 \times Pr(C) + 0 \times Pr(\neg C) = Pr(C). \tag{9}$$

But by THE THESIS,

$$Pr(A \rightarrow C) = Pr(C|A). \tag{10}$$

Substituting $Pr(C|A)$ for $Pr(A \rightarrow C)$ on the left-hand side of Equation 9, we get:

$$Pr(C|A) = Pr(C) \tag{11}$$

—in other words, A and C are probabilistically independent.

The above proof shows that Equation 11 holds for arbitrary propositions A and C, provided both $Pr(A \wedge C)$ and $Pr(A \wedge \neg C)$ are both greater than 0. Therefore Equation 11 should hold for all pairs of propositions A and C such that $Pr(A \wedge C)$ and $Pr(A \wedge \neg C)$ are both greater than 0. But this is only possible in trivial probability spaces. So one of our assumptions must have gone wrong, and the natural place to pin the blame is on THE THESIS.

There are various possible ways out of Lewis's triviality theorem. The proof assumes that the conditional $A \rightarrow C$ has a single set of truth conditions, which remain stable across Pr, Pr_C, and $Pr_{\neg C}$. Defenders of THE THESIS might reject this assumption and claim that the truth-conditions of conditionals are context-dependent. The proof also assumes that THE THESIS holds for all probability functions and all conditionals. Defenders of THE THESIS might retreat and claim that it is true for only some conditionals, or some probability functions.

Unfortunately, both escape routes are treacherous. New triviality theorems can be derived from much weaker assumptions; for a helpful survey, see Hall and Hájek (1994). There are even triviality results that use non-probabilistic variants of THE THESIS (Gärdenfors, 1988), and trivializing

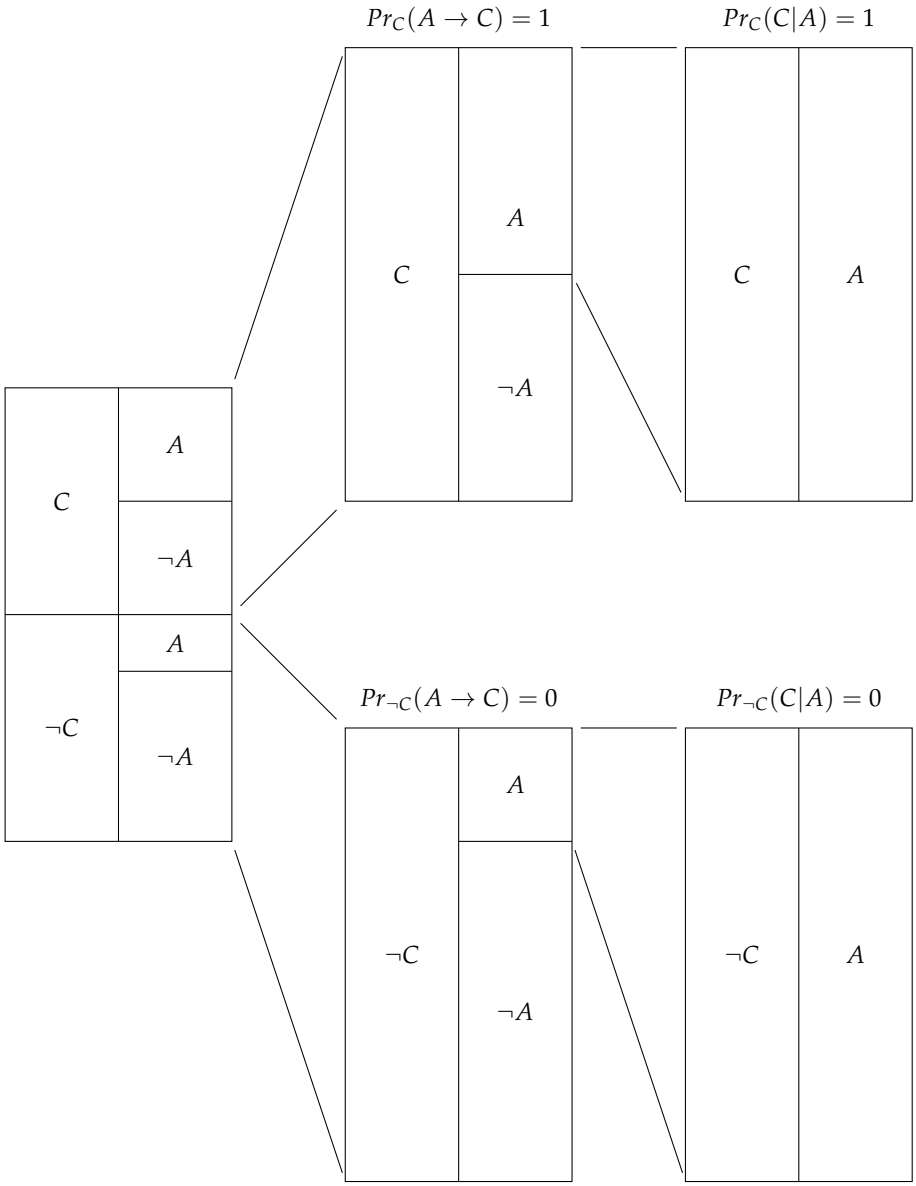


Figure 7: The Lewis triviality theorem illustrated

versions of THE THESIS that apply to counterfactuals rather than indicatives (Williams, 2012). On a slightly more optimistic note, *non*-triviality results can be obtained by adopting (sufficiently weak) non-classical logics (Morgan & Mares, 1995).

Another way out of Lewis's triviality theorem is to reject THE THESIS. Kaufmann (2004) produces examples of indicative conditionals in English that intuitively seem to violate THE THESIS, and Douven and Verbrugge (2013) provide experimental evidence that English speakers' judgments about indicative conditionals violate THE THESIS.

If probability is probability of truth, defenders of the NTV view should reject THE THESIS too. However, defenders of the NTV view typically defend versions THE THESIS, but adopt alternative interpretations of 'probability', on which the probability of a conditional is not the probability of its truth.

Calling on alternative theories of probability makes sense: probability is a versatile explanatory tool, and the NTV theory has plenty of explaining to do. In particular, the NTV theory needs to explain why conditionals seem to have the features of truth-evaluable statements. It is sometimes reasonable to believe a conditional—but ordinarily, to believe something is to believe that it is true. Likewise, it is sometimes reasonable to assert a conditional—but ordinarily, to assert something is to claim that it is true. Arguments with conditionals in their premises and conclusions are sometimes valid and sometimes invalid—but ordinarily, a valid argument is one that cannot have true premises and a false conclusion, and it's not clear how to fruitfully apply the concept of validity when a premise or conclusion lacks truth conditions altogether.

Adams (1975) and Edgington (1986) give a probabilistic account of belief in conditionals. Belief comes in degrees, which are measured by probabilities. A person's degree of belief in a conditional is simply her conditional degree of belief in its consequent on its antecedent.

Adams (1975) gives a probabilistic account of validity for conditionals. An argument is said to be probabilistically valid just in case it is impossible for its premises to be probable and its conclusion improbable. More precisely, an argument from premises $P_1, P_2, \ldots, P_n$ to conclusion C is valid just in case, for every real number $\epsilon > 0$, there is a real number $\delta > 0$ such that, if each of $P_1, P_2, \ldots, P_n$ has probability greater than $1 - \delta$, then C has probability at least $1 - \epsilon$.

Adams' definition of validity coincides with the classical definition where $P_1, P_2, \ldots, P_n$ and C are conditional-free sentences, and lets us define validity for arguments containing simple conditionals. The theory is built to handle only simple conditionals, and does not let us assess validity for arguments containing compound sentences with conditionals as parts. McGee (1989) extends Adams' theory to cover compounds of conditionals.

Edgington (1995) gives a non-probabilistic account of what it is to assert a conditional: it is to assert the consequent if the antecedent is true, and to assert nothing otherwise. She argues that her account assimilates conditional assertions to a larger class of conditional speech acts, including:

Conditional Questions 'If he phones, what shall I say?'

Conditional Commands 'If he phones, hang up.'

Conditional Promises 'If he phones, I promise not to be rude.'

Conditional Agreements 'If he phones, we're on for Sunday.'

Conditional Offers 'If you phone, you can have a 20% discount.'

Any speech act whatsoever, she claims, can be performed conditionally or unconditionally. We can think of conditionals as 'speech act bombs' primed to detonate when and only when the antecedent is true (see Egan, 2009).

8 Dynamic Semantics

So far, we've seen several accounts of conditionals that posit more to their meanings than truth conditions—either because conditionals have no truth conditions (on the NTV account) or because their truth conditions are not sufficient to determine when they can reasonably be asserted (on the material conditional account). Enter dynamic semantics, which provides new tools for modeling meaning.

Dynamic semantics explains the meanings of sentences by appeal to a *conversational context*—a set of background assumptions taken for granted by all the participants in a conversation. For instance, if a group of friends is discussing where to go for lunch, the conversational context might include the information that among the nearby restaurants are Veggie Garden and Buddha's Palace. The *context set* is a set of worlds compatible with those background assumptions (see Stalnaker, 1999, p. 84).[11]

The conversational context changes as the conversation progresses, and the context set shrinks and grows accordingly. When a participant makes an assertion, then the content of the assertion is added to the context, and all the worlds incompatible with what is asserted are eliminated from the context set. For instance, if someone asserts that Veggie Garden is open,

11 To give a complete theory of conditionals, the conversational context will need to include more information than just the context set. Other proposed parameters include a probability function or set of probability functions (Yalcin, 2007, 2012b), and a function that ranks worlds from most to least likely (Spohn, 2015). However, I focus my exposition on the context set to provide a simple illustration of the main ideas.

then the worlds where Veggie Garden is closed are eliminated from the context set.

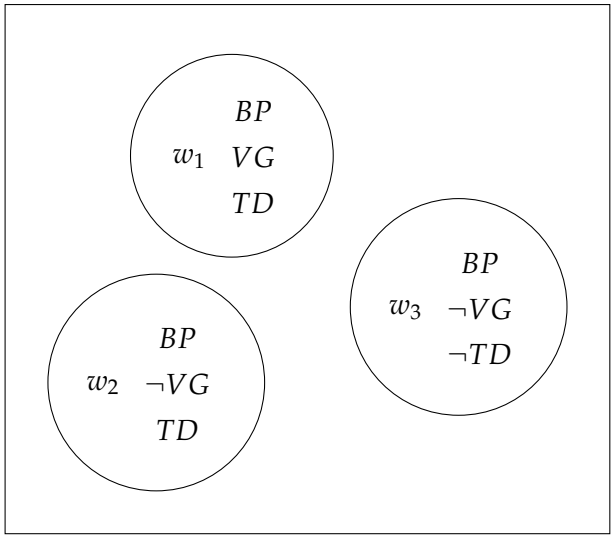
Figure 8 depicts the effect of asserting 'Veggie Garden is open today' on the context set. The original context set is shown in the rectangle at the top of the figure: the circles depict worlds. Each world is labeled with a set of propositions true at that world: 'BP' stands for 'Buddha's Palace is open'; 'VG' stands for 'Veggie Garden is open'; and 'TD' stands for 'we can get gluten-free tofu dogs'.

Notice that some assertions have no effect on the context set. If someone were to assert 'Buddha's Palace is open', none of the worlds in the context set would be eliminated. This is because 'Buddha's Palace is open' is already acceptable in the original context—it follows from what is accepted.

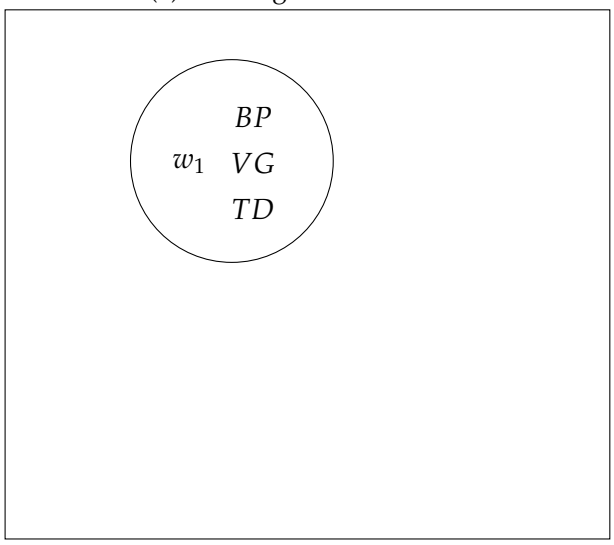
Starr (2014) proposes that within this dynamic semantics framework, conditionals can be understood as *tests*, along the lines of the Ramsey test. To determine the effect of asserting a conditional 'if A, then B' on a context c, we first suppose A, by considering the context $c[A]$ that results from adding A to c. We then check whether B is true under the supposition. If B is true at $c[A]$, then c 'passes' the test, and c remains unchanged. Otherwise, c 'fails' the test. If a conditional passes the test, it is acceptable in the original context.

This characterization of acceptability, by itself, is not enough to determine the effect of uttering a conditional in a context where it is not already acceptable. For instance, suppose you go to pet a dog, and I say 'if you pet it, it will bite.' This conditional doesn't follow from our shared background information, but you can use it to rule out possibilities—in particular, those possibilities where you pet the dog and it does not bite. What explains the relationship between my utterance and the corresponding change to the context set? In very broad terms, uttering a conditional should change the context set so that the conditional becomes acceptable, and the change involved should be the smallest one that does the job. There are multiple ways of spelling out what constitutes a minimal change to contextual information, but the part of the account that deals with acceptability can be separated from the part that deals with context change.

To illustrate the concept of a conditional test, consider the conditional 'if Veggie Garden is open, then we can get gluten-free tofu dogs', as asserted in the context depicted by Figure 8a. To perform the test, we first create a new context, by augmenting the old context with the information that Veggie Garden is open; the resulting context set is depicted in Figure 8b. We then check whether, in the new context, 'we can get gluten-free tofu dogs' is acceptable. If so, the old context passes the test, and the conditional is acceptable in the old context; otherwise, the old context fails the test, and the conditional is not acceptable in the old context. Starr extends his account to handle counterfactuals (which use a modified test



(a) The original context set



(b) The context set after an assertion of 'Veggie Garden is open'

Figure 8: The effects of an assertion on the context set

in which the context set is expanded with extra possibilities before adding the information in the antecedent).

Other theorists offer context-dependent truth conditions for conditionals using the tools of dynamic semantics. Stalnaker (1975) and Williams (2008) defend a modified closest-worlds theory of indicative conditionals, where, if w is a world in the context set, every world in the context set is stipulated to be closer to w than every world outside it. Gillies (2007) and Fintel (2001) propose strict conditional theories of counterfactuals, where a set of salient worlds is fixed by the context. New worlds are added to the set as the conversation goes on; in particular, if someone asserts a conditional whose antecedent is false at all the salient worlds, the set is expanded to include at least one world compatible with the antecedent.

9 CONDITIONALS AS MODAL RESTRICTORS

According to Kratzer (2012, p. 86), many of the above views of conditionals are 'based on a momentous syntactic mistake.' Contrary to popular opinion, she claims, 'There is no two-place *if…then* connective in the logical forms for natural languages.' Instead, conditionals restrict modal operators.

One can think of modal operators as quantifiers over possible worlds: to say that necessarily $2+2=4$ is to say that in all possible worlds, $2+2=4$; to say that possibly pigs fly is to say that in some possible world, pigs fly; and to say that it will probably rain is to say that in most possible worlds (on some suitable way of measuring 'most'), it rains. Like quantifiers, modal operators can be restricted. To say that necessarily, if the Peano axioms are true, then $2+2=4$, is to say that in all possible worlds where the Peano axioms are true, $2+2=4$. Likewise, to say that if pigs had hollow bones, then possibly pigs would fly, is to say that in some possible world where pigs have hollow bones, pigs fly, and to say that if there are cumulus clouds on the horizon, it will probably rain, is to say that in most possible worlds where there are cumulus clouds on the horizon, it will rain.

The modal restrictor view is a generalization of work by D. Lewis (1975) who notes that conditionals can be used to restrict quantifiers. Consider the following class of examples.

Sometimes Always Usually Never	if a farmer owns a donkey, she feeds it carrots.

The quantifiers 'sometimes', 'always', 'usually', and 'never' are what Lewis calls *unselective quantifiers*. To say that always, farmers feed donkeys carrots is to say that for all ways of assigning a farmer to x and a donkey to y, x feeds y carrots. To add the clause 'if a farmer owns a donkey' is to restrict the quantifier, so that it ranges only over cases where farmer x owns donkey y.

The modal restrictor view is Y-shaped: it can handle both indicatives and counterfactuals (Kratzer, 1981). To explain how this works, we need three ingredients: a modal base, the modal force of an operator, and an ordering.

According to Kratzer, the context of an utterance supplies a *modal base*, or a function f mapping each world w to a set of propositions that is 'held fixed' when we speculate about what might or must have been true at w. When we consider what is physically possible, the modal base might assign to each world the laws of physics that obtain at that world, but leave out physically contingent truths. When a detective speculates about who the burglar might be, the modal base might assign to each world the detective's evidence at that world. To determine what is possible (or necessary, or likely) at a world w, we need to quantify over the possible worlds where the all of the propositions in $f(w)$ are true.[12]

Different operators are associated with different kinds of *modal force*—roughly, different kinds of quantification over possible worlds. The operators 'necessarily', 'possibly', 'it is likely that', and 'it is a good possibility that' are all associated with different modal forces. Finally, the context of utterance supplies an *ordering source* g, which lets us map each world to an ordering over worlds.[13] (One possible interpretation of this ordering is the 'closeness' ordering from Section 4.2, but there are others. Conditional and unconditional statements about what ought to happen use an ordering source that ranks worlds from most to least ideal.)

We can then say that the conditional 'Necessarily if A, then B' is true at a world w just in case B is true at all the closest A-worlds to w (according to the ordering $g(w)$) where all the propositions in $f(w)$ are true. Likewise,

12 Kratzer's theory could be reformulated in terms of a familiar two-place accessibility relation among worlds. We might say that world x is accessible from world w (Rwx in the usual formalism) if and only if all of the propositions in $f(w)$ are true at x. A few complications arise when the modal base maps some worlds onto inconsistent sets of propositions. Kratzer wants to say that in such cases, there are non-trivial facts about what is possible; she gives the example of a modal base that assigns to each world the set of propositions that are required by a group of Maori elders in that world (Kratzer, 1981, pp. 16-20). In one world w, the elders disagree amongst themselves, and so their requirements are inconsistent. Nonetheless, there are non-trivial facts about what is necessary at w according to the elders' requirements; Kratzer claims that the structure of the set $f(w)$ of propositions plays an essential role in determining what is necessary.

13 Kratzer's ordering source officially maps worlds to sets of propositions, which are then used to create an ordering. I omit this extra step.

'Possibly if A, then B' is true at a world w just in case at some of the closest A-worlds to w (according to the ordering $g(w)$) where all the propositions in $f(w)$ are true, and similarly for other operators with other modal force. For indicative conditionals, the modal base is some piece of salient known information. For counterfactual conditionals, the modal base is empty (and thus, all possible worlds are consistent with it) while the ordering source is very rich. Kratzer's account even has the material conditional account as a special case, where the modal base maps each world w to a set of propositions true only at w, and the strict conditional as another special case, where the modal base is empty and the ordering source is completely noncommittal, invariably ranking all worlds on a par with each other.

'Bare' conditionals cause trouble for the modal restrictor view. Conditionals supposedly restrict modal operators, but where is the modal operator in a conditional like 'If the lights in his study are on, then Roger is home'? Kratzer (1979, 1981) argues that conditionals without overt modal operators nonetheless contain implicit modal operators; the underlying logical form of the example conditional is '(MUST: the lights in his study are on) Roger is home'; the epistemic 'MUST' is unspoken.

Heim (1982) provides evidence for Kratzer's modalized interpretation of bare conditionals in the form of 'donkey sentences' like 'If John owns a donkey, then he feeds it carrots'. On at least one plausible reading, our sample donkey sentence means that John feeds carrots to every donkey he owns—or in more cumbersome terms, for every x such that x is a donkey and John owns x, John feeds x carrots. If the conditional were an ordinary two-place connective, we would have trouble explaining how the same variable x, bound by the same quantifier, could occur in both the antecedent and the consequent of the donkey sentence. The conditional would have the form $A \rightarrow B$, where A contained a quantifier ranging over donkeys. But Kratzer's restrictor analysis, together with the assumption that bare conditionals contain a tacit necessity operator, gives the correct reading, while providing a uniform treatment of bare and modalized conditionals.

It is often claimed that Kratzer's modal restrictor theory allows us to escape the triviality results of Section 7. Rothschild (2013), for instance, suggests that Kratzer can escape the triviality results by denying THE THESIS. To illustrate Rothschild's argument, let's consider the conditional I originally used to motivate THE THESIS.

ACE If I draw a red ace, then it will be the ace of hearts.

I accept the conditional:

CHANCY ACE With probability 1/2, if I draw a red ace, then it will be the ace of hearts.

Rothschild suggests that on Kratzer's account, CHANCY ACE does not express the thought that ACE has probability 1/2, or the thought that the probability of ACE's being true is 1/2. When I assert CHANCY ACE, I am not asserting that ACE has probability 1/2. Furthermore, when I am 50% confident that if I draw a red ace, it will be the ace of hearts, this does not amount to my being 50% confident that ACE is true.

Charlow (2015) argues that even if Rothschild is right, Kratzer's account is still vulnerable to the triviality result, since steps Equations 5 and 8 can be motivated independently of THE THESIS. He goes on to argue that other easy ways out of the triviality result fail on the modal restrictor view.

10 CONCLUSION

Conditionals are important in both everyday reasoning and philosophical argument. There are conditional beliefs, conditional assertions, and conditional propositions, all of which can figure in arguments. The theories canvassed in this article try to systematize the broad range of data about which conditionals seem true, and which inferences seem valid. More phenomena remain to be explained: this article has focused on conditional beliefs and assertions, and on conditionals in English.

We can gather the similarities among the accounts discussed above into a sort of rake-shaped theory (a generalization of Bennett's concept of a Y-shaped theory), with a short 'handle' that captures what is common to all conditionals, which then splits into many 'tines' that capture the particularities of individual theories. All of the theories we have considered so far have the following commitments in common.

1. Conditionals are evaluated at 'points'.
2. To evaluate a conditional 'if A, then C' at a point p, one generates a new point q by adding the information in A to p.
3. The evaluation of the consequent C at q is the evaluation of the entire conditional at p.

The accounts disagree about the natures of points, what status conditionals and their consequents should be evaluated for, and what adding an antecedent amounts to. Table 1 summarizes how different views answer this question. (NB: Selection function and relevant logic accounts typically treat the initial point and the new point as belonging to different types—the initial point is a world, while the new point is a set of worlds. But we can ensure that both points are of the same type by rewriting the theory so that the initial point is a singleton set of one world; this is what I have done in Table 1.)

	Points	Status	Adding A to a Point
Strict Conditional	Sets of worlds	Truth in all worlds (original point is a singleton $\{w\}$)	Taking all worlds possible at w compatible with A
Closest Worlds and Past Predominance	Sets of worlds	Truth in all worlds (original point is a singleton $\{w\}$)	Taking all closest worlds possible at w compatible with A
Causal Modeling	Causal models with valuations	Truth in a model	Intervening to make A true
Relevant Logic	Sets of worlds	Truth in all worlds (original point is a singleton $\{w\}$)	Taking all worlds y such that $Rwxy$ for some world x compatible with A
Material Conditional	Worlds	Truth in the world	Doing nothing if A is true; moving to the 'absurd world' (where everything is true) otherwise
Probability Accounts	Probability functions	Probability $x \in [0, 1]$	Conditionalizing on A
Dynamic Test Theory	Contexts	Acceptability	Updating to accommodate an assertion of A
Modal Restrictors	Information states: modal base + ordering source	Obtaining with a given modal force	Taking all closest worlds to A in the modal base, according to the ordering source

Table 1: Theories of conditionals and their components

Within each of the accounts, there are open questions: the nature of the selection function; the correct interpretation of counterpossibles; how best to respond to the triviality theorems; what makes a conditional believable or assertable in a given context; how to handle bare modals on the restrictor account.

There are also open questions about how the accounts interact. Some accounts seem to be special cases of others: the past predominance view is a way of filling in the meaning of 'closest' on the closest-worlds account. At other times, different accounts appear to be rivals: it can't be both that indicative conditionals have the truth conditions given by the material interpretation, and that they lack truth values. At other times, they seem to be modeling different domains: as with Pearl's causal modeling theory of counterfactuals and Starr's dynamic semantics theory of indicatives. Much of the interest for future research lies in understanding the interactions between the different models of conditionals.

If conditionals are useful in a wide variety of domains, from childhood development to everyday reasoning to philosophy, then conditionals are well worth studying. I have given reasons for thinking that conditionals are useful in a wide variety of domains. You may draw your own conclusions.

REFERENCES

Adams, E. W. (1970). Subjunctive and indicative conditionals. *Foundations of Language*, *6*(1), 89–94.

Adams, E. W. (1975). *The logic of conditionals: An application of probability to deductive logic*. D. Reidel.

Adams, E. W. (1988). Modus tollens revisited. *Analysis*, *48*(3), 122–128.

Aksenov, P. (2013). Stanislav Petrov: The man who may have saved the world. Retrieved, from http://www.bbc.com/news/world-europe-24280831

Alonso-Ovalle, L. (2009). Counterfactuals, correlatives, and disjunction. *Linguistics and Philosophy*, *32*(2), 207–244.

Amsel, E. & Smalley, D. (2000). Beyond really and truly: Children's counterfactual thinking about pretend possible worlds. In P. Mitchell & K. Riggs (Eds.), *Children's reasoning and the mind* (pp. 121–147). Psychology Press Ltd.

Appiah, A. (1985). *Assertion and conditionals*. Cambridge: Cambridge University Press.

Arregui, A. (2009). On similarity in counterfactuals. *Linguistics and Philosophy*, *32*(3), 245–278.

Ayer, A. (1954). Freedom and necessity. In *Philosophical essays*. London: Macmillan.

Barwise, J. & Perry, J. (1981). Situations and attitudes. *Journal of Philosophy*, *78*(11), 668–691.

Beall, J., Brady, R., Dunn, J. M., Hazen, A. P., Mares, E., Meyer, R. K., … Sylvan, R. (2012). On the ternary relation and conditionality. *Journal of Philosophical Logic*, *41*(3), 595–612.

Bennett, J. (1988). Farewell to the phlogiston theory of conditionals. *Mind*, *97*(388), 509–27.

Bennett, J. (2001). Conditionals and explanations. In A. Byrne, R. Stalnaker, & R. Wedgwood (Eds.), *Fact and value: Essays on ethics and metaphysics for judith jarvis thomson*. Cambridge, MA: MIT Press.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford University Press.

Bernstein, S. (2016). Omission impossible. *Philosophical Studies*, *173*(10), 2575–2589.

Berto, F. (2013). Impossible worlds. *Stanford Encyclopedia of Philosophy*. Retrieved from http://plato.stanford.edu/entries/impossible-worlds/

Bjerring, J. C. (2013). On counterpossibles. *Philosophical Studies*, *168*(2), 1–27.

Bobzien, S. (2002). The development of modus ponens in antiquity: From Aristotle to the 2nd century AD. *Phronesis*, *47*(4), 359–94.

Bradley, R. (2000). Conditionals and the logic of decision. *Philosophy of Science*, *67*(3), S18–S32.

Brewka, G. (1991). *Nonmonotonic reasoning: Logical foundations of commonsense*. Cambridge: Cambridge University Press.

Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies*, *160*(1), 139–166.

Brogaard, B. & Salerno, J. (forthcoming). A counterfactual account of essence. *The Reasoner*.

Brogaard, B. & Salerno, J. (2008). Counterfactuals and context. *Analysis*, *68*(297), 39–46.

Brogaard, B. & Salerno, J. (2013). Remarks on counterpossibles. *Synthese*, *190*(4), 639–660.

Cantwell, J. (2013). Conditionals in Causal Decision Theory. *Synthese*, *190*(4), 661–679.

Charlow, N. (2015). Triviality for restrictor conditionals. *Noûs*, *49*(3), 1–32.

Choi, S. (2006). The simple vs. reformed conditional analysis of dispositions. *Synthese*, *148*(2), 369–379.

Choi, S. (2009). The conditional analysis of dispositions and the intrinsic dispositions thesis. *Philosophy and Phenomenological Research*, *78*(3), 568–590.

Collins, J., Hall, N., & Paul, L. A. (2004). Counterfactuals and causation: History, problems, and prospects. In J. Collins, N. Hall, & L. Paul

(Eds.), *Causation and counterfactuals* (pp. 1–57). Cambridge, MA: The MIT Press.

Cross, C. B. (1990). Temporal necessity and the conditional. *Studia Logica*, *49*(3), 345–363.

Darwall, S. L. (1983). *Impartial reason*. Ithaca: Cornell University Press.

Dias, M. & Harris, P. [P.L.]. (1990). The influence of the imagination on reasoning by young children. *Developmental Psychology*, *8*(4), 305–318.

Díez, J. (2015). Counterfactuals, the discrimination problem and the limit assumption. *International Journal of Philosophical Studies*, *23*(1), 85–110.

Douven, I. & Verbrugge, S. (2013). The probabilities of conditionals revisited. *Cognitive Science*, *37*(4), 711–730.

Dowell, J. J. L. (2011). A flexible contextualist account of epistemic modals. *Philosophers' Imprint*, *11*(14), 1–25.

Dudman, V. (1983). Tense and time in english verb clusters of the primary pattern. *Australian Journal of Linguistics*, *3*(1), 25–44.

Dudman, V. (1984). Parsing 'if'-sentences. *Analysis*, *44*(4), 145–53.

Edgington, D. (1986). Do conditionals have truth-conditions? *Crítica*, *18*(52), 3–30.

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329.

Edgington, D. (2004). Counterfactuals and the benefit of hindsight. In P. Dowe & P. Noordhof (Eds.), *Cause and chance: Causation in an indeterministic world* (pp. 12–27). Routledge.

Edgington, D. (2008). Counterfactuals. *Proceedings of the Aristotelian Society*, *108*(1), 1–21.

Egan, A. (2009). Billboards, bombs and shotgun weddings. *Synthese*, *166*(2), 251–279.

Fine, K. (1975). Critical notice of Lewis, counterfactuals. *Mind*, *84*(335), 451–458.

Fintel, K. v. (2001). Counterfactuals in a dynamic context. In M. Kentstowicz (Ed.), *Ken Hale: A life in language*. Cambridge, MA: MIT Press.

Fintel, K. v. (2011). Conditionals. In K. von Heusinger, C. Maienborn, & P. Portner (Eds.), *Semantics: An international handbook of meaning* (pp. 1515–1538). DeGruyter.

Galles, D. & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, *3*(1), 151–182.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge, MA: MIT Press.

Gibbard, A. (1981). Two Recent Theories of Conditionals. In W. Harper, R. C. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 211–247). Reidel.

Gibbard, A. & Harper, W. L. (1981). Counterfactuals and two kinds of expected utility. In W. Harper, R. C. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 153–190). Reidel.

Gillies, A. S. (2007). Counterfactual scorekeeping. *Linguistics and Philosophy*, *30*(3), 329–360.

Gillon, B. (2011). Logic in classical Indian philosophy. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2011). Retrieved from http://plato.stanford.edu/archives/sum2011/entries/logic-india/

Gopnik, A. (2009). *The philosophical baby: What children's minds tell us about truth, love, and the meaning of life*. Random House.

Grice, H. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Hájek, A. (manuscript). *Most counterfactuals are false*.

Hall, N. & Hájek, A. (1994). The hypothesis of the conditional construal of conditional probability. In *Probabilities and conditionals: Belief revision and rational decision* (pp. 75–110). Cambridge: Cambridge University Press.

Harris, P. [Paul]. (2000). *The work of the imagination: Understanding children's worlds*. Blackwell Publishing.

Heim, I. (1982). *The semantics of definite and indefinite noun phrases* (Doctoral dissertation, University of Massachusetts).

Hintikka, J. (1975). Impossible possible worlds vindicated. *Journal of Philosophical Logic*, *4*(4), 475–484.

Huber, F. (2013). Structural equations and beyond. *Review of Symbolic Logic*, *6*(4), 709–732.

Jackson, F. (1987). *Conditionals*. Cambridge, MA: Blackwell Publishing.

Jago, M. (2013). Recent work in relevant logic. *Analysis*, *73*(3), 526–541.

Kaufmann, S. (2004). Conditioning against the grain. *Journal of Philosophical Logic*, *33*(6), 583–606.

Kolodny, N. & MacFarlane, J. (2010). Ifs and oughts. *Journal of Philosophy*, *107*(3), 115–143.

Koons, R. (2014). Defeasible reasoning. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2014). Retrieved from http://plato.stanford.edu/entries/reasoning-defeasible/

Krakauer, B. (2012). *Counterpossibles* (Doctoral dissertation, University of Massachusetts).

Kratzer, A. (1979). Conditional necessity and possibility. In R. Bäurle, U. Egli, & A. Stechow (Eds.), *Semantics from different points of view* (pp. 117–47). Springer.

Kratzer, A. (1981). The notional category of modality. In H. Eikmeyer & H. Reiser (Eds.), *Words, worlds, and contexts*. de Gruyter.

Kratzer, A. (2012). *Modals and conditionals: New and revised perspectives*. Oxford: Oxford University Press.

Krzyzanowska, K. (2013). Belief ascription and the Ramsey test. *Synthese*, *190*(1), 21–36.

Lauer, S. & Condoravdi, C. (2014). Preference-conditioned necessities: Detachment and practical reasoning. *Pacific Philosophical Quarterly*, *95*(4), 584–621.

Lewis, C. (1918). *Survey of symbolic logic*. University of California Press.

Lewis, D. (1973a). *Counterfactuals*. Blackwell Publishing.

Lewis, D. (1973b). Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, *2*(4), 418–446.

Lewis, D. (1975). Adverbs of quantification. In E. Keenan (Ed.), *Semantics of natural language* (pp. 3–15). Cambridge: Cambridge University Press.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, *85*(3), 297–315.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*(4), 455–476.

Lewis, K. S. (2015). Elusive counterfactuals. *Noûs*, *49*(4).

Lillard, A. (2001). Pretend play as twin earth: A social-cognitive analysis. *Developmental Review*, *21*(4), 495–531.

Loewer, B. (1976). Counterfactuals with disjunctive antecedents. *Journal of Philosophy*, *73*(16), 531–537.

Mares, E. D. (1994). Why we need a relevant theory of conditionals. *Topoi*, *13*(1), 31–36.

Mares, E. D. & Fuhrmann, A. (1995). A relevant theory of conditionals. *Journal of Philosophical Logic*, *24*(6), 645–665.

McGee, V. (1985). A counterexample to modus ponens. *The Journal of Philosophy*, *82*(9), 462–471.

McGee, V. (1989). Conditional probabilities and compounds of conditionals. *Philosophical Review*, *98*(4), 485–541.

McGee, V. (2005). 24.241 logic I, fall 2005. Retrieved from http://ocw.mit.edu/courses/linguistics-and-philosophy/24-241-logic-i-fall-2005/readings/chp14.pdf

Mckay, T. & Inwagen, P. V. (1977). Counterfactuals with disjunctive antecedents. *Philosophical Studies*, *31*(5), 353–356.

Melia, J. (2001). Reducing possibilities to language. *Analysis*, *61*(1), 19–29.

Menzies, P. (2014). Counterfactual theories of causation. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2014). Retrieved from http://plato.stanford.edu/archives/spr2014/entries/causation-counterfactual/

Moore, G. (1912). *Ethics*. London: Williams and Norgate.

Morgan, C. G. & Mares, E. D. (1995). Conditionals, probability, and non-triviality. *Journal of Philosophical Logic*, *24*(5), 455–467.

Moss, S. (2012). On the pragmatics of counterfactuals. *Noûs*, *46*(3), 561–586.

Nolan, D. (1997). Impossible worlds: A modest approach. *Notre Dame Journal of Formal Logic*, *38*(4), 535–572.

Nolan, D. (2013). Why historians (and everyone else) should care about counterfactuals. *Philosophical Studies*, *163*(2), 317–335.

Nozick, R. (1981). *Philosophical explanations*. Cambridge, MA: Harvard University Press.

Nute, D. (1975). Counterfactuals and the similarity of words. *Journal of Philosophy*, *72*(21), 773–778.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd). Cambridge: Cambridge University Press.

Phillips, I. (2007). Morgenbesser cases and closet determinism. *Analysis*, *67*(293), 42–49.

Pollock, J. L. (1976). The 'possible worlds' analysis of counterfactuals. *Philosophical Studies*, *29*(6), 469–476.

Prior, E. W., Pargetter, R., & Jackson, F. (1982). Three theses about dispositions. *American Philosophical Quarterly*, *19*(3), 251–257.

Ramsey, F. (1978). Law and causality. In D. Mellor (Ed.), *Foundations* (pp. 128–51). Routledge.

Reiss, J. (2009). Counterfactuals, thought experiments, and singular causal analysis in history. *Philosophy of Science*, *76*(5), 712–723.

Rothschild, D. (2013). Do indicative conditionals express propositions? *Noûs*, *47*(1), 49–68.

Routley, R. & Meyer, R. (1972a). The semantics of entailment II. *Journal of Philosophical Logic*, *1*(1), 53–73.

Routley, R. & Meyer, R. (1972b). The semantics of entailment III. *Journal of Philosophical Logic*, *1*(2), 192–208.

Routley, R. & Meyer, R. (1973). The semantics of entailment I. In H. Leblanc (Ed.), *Truth, syntax, and semantics* (pp. 194–243). North-Holland.

Ryle, G. (1950). 'If', 'so', and 'because'. In M. Black (Ed.), *Philosophical analysis*. Ithaca, NY: Cornell University Press.

Salerno, J. & Brogaard, B. (forthcoming). Williamson on counterpossibles. *The Reasoner*.

Schulz, K. (2011). 'if you'd wiggled A, then B would've changed'. *Synthese*, *179*(2), 239–251.

Sider, T. (2002). The ersatz pluriverse. *The Journal of Philosophy*, *99*(6), 279–315.

Slote, M. (1978). Time in counterfactuals. *The Philosophical Review*, *87*(1), 3–27.

Sobel, J. H. (1970). Utilitarianisms: Simple and general. *Inquiry*, *13*(1-4), 394–449.

Sosa, E. (1999). How to defeat opposition to Moore. *Philosophical Perspectives*, *13*, 141–153.

Spohn, W. (2015). Conditionals: A unified ranking-theoretic perspective. *Philosophers' Imprint*, *15*(1).

Stalnaker, R. (1968). A theory of conditionals. *American Philosophical Quarterly*, 98–112.

Stalnaker, R. (1975). Indicative conditionals. *Philosophia*, *5*(3), 269–86.

Stalnaker, R. (1981). A defense of conditional excluded middle. In W. Harper, R. C. Stalnaker, & G. Pearce (Eds.), *Ifs* (pp. 87–104). D. Reidel.

Stalnaker, R. (1999). Assertion. In *Context and content*. Oxford University Press.

Starr, W. B. (2014). A uniform theory of conditionals. *Journal of Philosophical Logic*, *43*(6), 1019–1064.

Thomason, R. & Gupta, A. (1980). A theory of conditionals in the context of branching time. *Philosophical Review*, *89*(1), 65–90.

Vinci, T. C. (1988). Objective chance, indicative conditionals and decision theory; or, how you can be smart, rich and keep on smoking. *Synthese*, *75*(1), 83–105.

Walton, D. (2001). Are some modus ponens arguments deductively invalid? *Informal Logic*, *22*(1), 19–46.

Walton, K. (1990). *Mimesis as make-believe: On the foundations of the representational arts*. Harvard University Press.

Warmbrod, K. (1982). A defense of the limit assumption. *Philosophical Studies*, *42*(1), 53–66.

Weisberg, D. & Gopnik, A. (2013). Pretense, counterfactuals, and bayesian causal models: Why what is not real really matters. *Cognitive Science*, *37*(7), 1368–1381.

Williams, J. R. G. (2008). Conversation and conditionals. *Philosophical Studies*, *138*(2), 211–223.

Williams, J. R. G. (2012). Counterfactual triviality: A Lewis-impossibility argument for counterfactuals. *Philosophy and Phenomenological Research*, *3*(85), 648–670.

Williamson, T. (2007). *The philosophy of philosophy*. Oxford: Blackwell Publishing.

Woodward, J. (2004). Counterfactuals and causal explanation. *International Studies in the Philosophy of Science*, *18*(1), 41–72.

Yagisawa, T. (2010). *Worlds and individuals, possible and otherwise*. Oxford University Press.

Yalcin, S. (2007). Epistemic modals. *Mind*, *116*(464), 983–1026.

Yalcin, S. (2012a). A counterexample to modus tollens. *Journal of Philosophical Logic*, *41*(6), 1001–1024.

Yalcin, S. (2012b). Bayesian expressivism. *Proceedings of the Aristotelian Society*, *112*(2), 123–160.

COLOPHON

This book was typeset in LaTeX, using the `classicthesis` package developed by André Miede and Ivo Pletikosić. The style is inspired by Robert Bringhurst's seminal book, *The Elements of Typographic Style*.