THE OPEN HANDBOOK OF FORMAL EPISTEMOLOGY

RICHARD PETTIGREW & JONATHAN WEISBERG, EDS.

LIST OF CONTRIBUTORS

R. A. Briggs
Stanford University

Michael Caie
University of Pittsburgh

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

This stub is a placeholder; work on this entry hasn't begun yet.

Lewis (1981) argues that 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.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

REFERENCES

Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, *59*(1), 5–30.

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:

For all x and $y \in X$, $EU(x) \geq EU(y)$ 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 Bernouilli 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] Bernouilli 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]

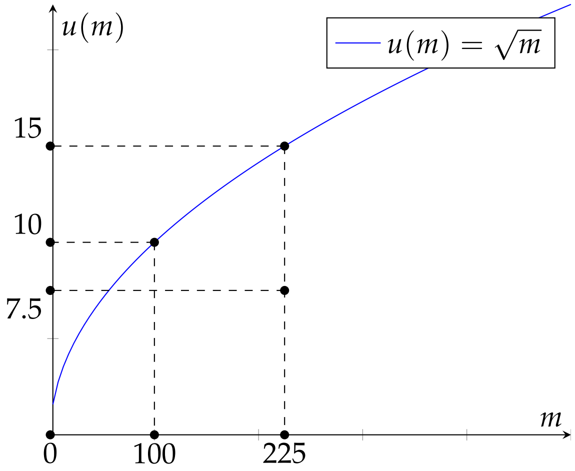


Figure 1: A concave utility function

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.

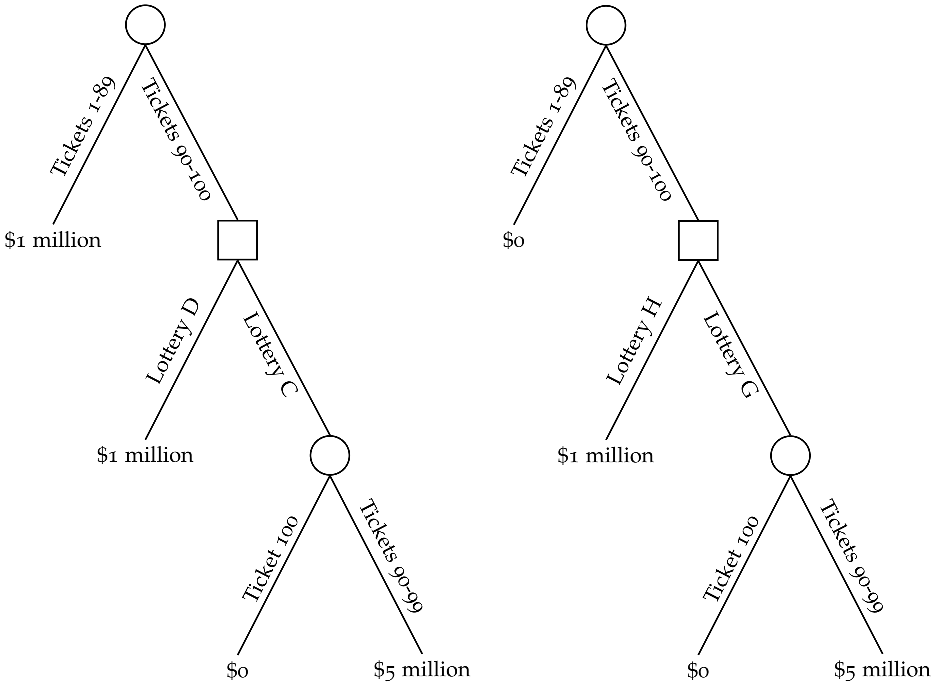
7.1 *Dynamic arguments in favour of separability*

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] Sophisticated agents treat continuation trees within dynamic choice problems

[49] See, for instance, Wakker (1988).

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

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 disprefered 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?

Since most agents prefer for good outcomes to occur earlier, and bad outcomes to occur later, Samuelson (1937) proposed the discounted utility

[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.

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 'instantenous' 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 received special philosophical attention in the case of cost-benefit analyses of the effects of climate change. Ethicists and economists have debated

[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.

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

[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).

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. *Philosopher's Imprint*, *10*.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics*, *75*, 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. (1999). *The foundations of causal decision theory*. Cambridge University Press.

Joyce, J. (2011). A defense of imprecise credence in inference and decision. *Philosophical Perspectives*, *24*, 281–323.

Joyce, J. (2012). Regret and instability in causal decision theory. *Synthese*, *187*, 123–145.

Joyce, J. & 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. *Philosopher's Imprint*, *14*(4).

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 $A_i \cap A_j = \varnothing$ 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 probablism, 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 counterintuitve. 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 bais 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 probabilsm 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 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 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 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. (2010). Subjective probabilities should be sharp. *Philosophers' Imprint*, *10*(5), 1–11.

Elga, A. (2012). Errata for subjective probabilities should be sharp.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The 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*(1), 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: 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*. New York: 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. *Annals of Statistics*, *21*(3), 1139–54.

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*(1), 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, R. (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

PRIMITIVE CONDITIONAL PROBABILITIES

Kenny Easwaran

This stub is a placeholder; work on this entry hasn't begun yet.

Lewis (1981) argues that 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.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

REFERENCES

Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, *59*(1), 5–30.

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 $0 + 0 + 0 + \ldots + 0 + \ldots = 1$ *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, non-standard 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 multiplicatory 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 dp.[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 dp indicated an infinitesimal increment of a quantity p. In contemporary standard analysis, however, there are no non-zero infinitesimals and dp 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 aplied 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. 77 by de Finetti (1974, p. 347) concerning the absurdity of $0 + 0 + 0 + \ldots + 0 + \ldots = 1$. 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*:

(N0) 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 $\forall A \in \mathcal{P}(\Omega)$, $P(A) = J(\varphi_A)$, 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:

$$\begin{aligned}\frac{\mathrm{d}y}{\mathrm{d}x} &= \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{y(x + \Delta x) - y(x)}{\Delta x},\end{aligned}$$

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 G0B8616N and G066918N.

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

This stub is a placeholder; work on this entry hasn't begun yet.

Lewis (1981) argues that 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.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

REFERENCES

Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, *59*(1), 5–30.

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, $\ldots$, 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 $B\bot\neg\phi$ denote the set of all inclusion-maximal subsets of B that do not entail $\neg\phi$. In other words, $B\bot\neg\phi$ 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 $B\bot\neg\phi$ 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 $B\bot\neg\phi$, 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 $I_v F_b F_s$ denote the possible world in which Verdi is Italian, Bizet is French, and Satie is French. In general, a possible world assigns the two nationalities (I and F) to the three composers (v, b, and s). So there are eight possible worlds total, shown in Figure 1. The arrows

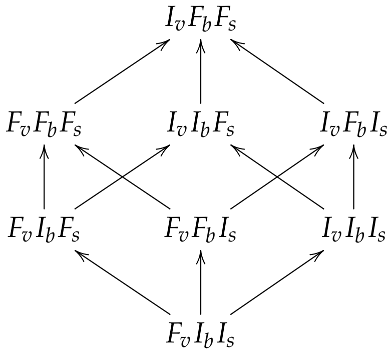


Figure 1: Hasse diagram of the Three Composers problem

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

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

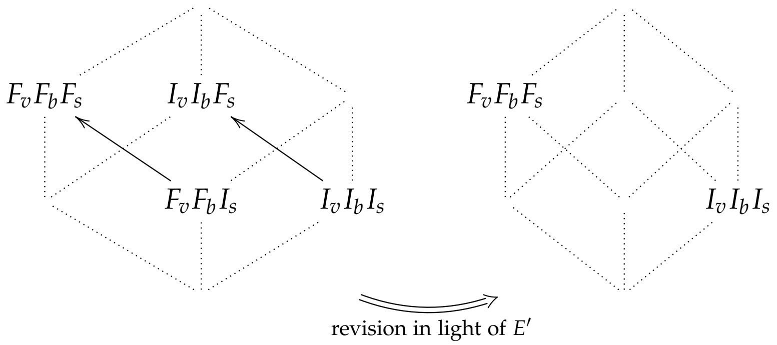


Figure 2: Revising in light of E, and then E'

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:

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

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

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

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

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

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

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

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

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

Namely, we "grab" the order $\geq$ over the whole W, "cut" the part of $\geq$ over X, and "paste" it on "top" of the other part $W \setminus X$, making any world inside X more plausible than any world outside X (condition 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	$\neg h, n_1$	$\neg h, b_1, n_2$	$\dots$	$\neg h, b_1, b_2, \dots, n_{101}$	$\dots$
$S^{\text{ock}}_{\text{non-skep}}$						
$S^{\text{non-ock}}_{\text{non-skep}}$						
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 "$\neg h, b_1, \dots, b_i, n_{i+1}$" 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:

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

where $\mathsf{Cn}\, X$ denotes the set of logical consequences of X, for any set X of sentences. So, for example, $S_{\text{skep}}(b_1, b_2, \dots, b_{i-1}, n_i)$ contains $\neg h$ because n_i entails $\neg h$. But $S_{\text{skep}}(b_1, b_2, \dots, b_i)$ 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:

$$S^{\text{ock}}_{\text{non-skep}}(\phi_1, \dots, \phi_i) =_{\text{def}} \begin{cases} \mathsf{C}n\{\phi_1, \dots, \phi_i, h\} & \text{if } i \geq 100 \text{ and } \phi_j = b_j \text{ for} \\ & \text{all } j \leq i, \\ \mathsf{C}n\{\phi_1, \dots, \phi_i\} & \text{otherwise.} \end{cases}$$

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:

$$S^{\text{non-ock}}_{\text{non-skep}}(\phi_1, \ldots, \phi_i) =_{\text{def}} \begin{cases} \mathrm{C}n\{\phi_1, \ldots, \phi_i, \neg h\} & \text{if } 50 \leq i < 100 \text{ and} \\ & \phi_j = b_j \text{ for all } j \leq i, \\ \mathrm{C}n\{\phi_1, \ldots, \phi_i, h\} & \text{if } i \geq 100 \text{ and } \phi_j = b_j \\ & \text{for all } j \leq i, \\ \mathrm{C}n\{\phi_1, \ldots, \phi_i\} & \text{otherwise.} \end{cases}$$

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 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 "$\neg h, b_1, \dots, n_{i+1}$": 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."

	h	$\neg h, n_1$	$\neg h, b_1, n_2$	$\dots$	$\neg h, b_1, b_2, \dots, n_{101}$	$\dots$
$S^{\text{ock}}_{\text{non-skep}}$						
$S^{\text{non-ock}}_{\text{non-skep}}$						
S_{skep}	No	Yes	Yes	$\dots$	Yes	$\dots$

Table 2: Decision table for the Raven Problem continued

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 "$\neg h, b_1, \dots, b_i, n_{i+1}$" 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	$\neg h, n_1$	$\neg h, b_1, n_2$	$\dots$	$\neg h, b_1, b_2, \dots, n_{101}$	$\dots$
$S^{\text{ock}}_{\text{non-skep}}$	(Yes, 0)	(Yes, 0)	(Yes, 0)	$\dots$	(Yes, 0)	$\dots$
$S^{\text{non-ock}}_{\text{non-skep}}$	(Yes, 0)	(Yes, 0)	(Yes, 0)	$\dots$	(Yes, 1)	$\dots$
S_{skep}	(No, 0)	(Yes, 0)	(Yes, 0)	$\dots$	(Yes, 0)	$\dots$

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 the non-skeptic Ockham strategy is "OK."

	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 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 Kelly, Genin, and Lin (2016), which build on Schulte (1999) and Kelly (2007). Also see 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:

> $\alpha =$ the conjunction of what Carla says and what Dora says;
> $\beta =$ what Elmer says.

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

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

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

This hypothetical exchange between Stalnaker and the defender of Axiom C2 raises a deep question. The clash between Stalnaker's counterexample and the defender's response can be taken as a debate over *what*

[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:

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

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

The question I want to discuss here is how the above line of thought can possibly be turned into an explicit argument with a clearly specified normative conclusion. Let us explore some possibilities. Suppose that the procedure (0)–(3) of partial meet contraction has been given an interpretation in line with the motivations provided in section 4.2. Suppose, further, that the normative thesis to be argued for is 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 $B \perp \neg\phi$. Gärdenfors' interpretation of $\geq$ does not cause any incoherence, but he leaves us with some unanswered questions. First, how should we understand the concept that Gärdenfors refers to as epistemic importance? Second, why 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. 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 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 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. (1) the true answer $\mathsf{Ans}(\mathcal{Q} \mid \mathcal{C})$ exists;
  2. (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. (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. (1999). Iterated belief revision, reliability, and inductive amnesia. *Erkenntnis*, *50*(1), 7–53.

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

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

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

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. (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 $\dotplus$ turns each old belief set $\mathcal{B} \subseteq \mathcal{L}$ and each sentence A from $\mathcal{L}$ into a new belief set $\mathcal{B} \dotplus A = Cn(\mathcal{B} \cup \{A\})$. 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

$*7.$ $\mathcal{B} * (A \wedge B) \subseteq Cn\,\big((\mathcal{B} * A) \cup \{B\}\big)$. Conjunction 1

$*8.$ If $\neg B \notin \mathcal{B} * A$, then
$Cn\,\big((\mathcal{B} * A) \cup \{B\}\big) \subseteq \mathcal{B} * (A \wedge B)$. 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 $\mathcal{B} \mathbin{\dot{-}} \neg A$ by adding the new information A:

$$\mathcal{B} * A = Cn\,\big((\mathcal{B} \mathbin{\dot{-}} \neg A) \cup \{A\}\big)\,.$$

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 $\mathcal{B} \mathbin{\dot{-}} A$ 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 $\mathcal{L}$:

$\dot{-}1.$ $\mathcal{B} \mathbin{\dot{-}} A = Cn\,(\mathcal{B} \mathbin{\dot{-}} A)$. 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 $B \npreceq A$.

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 1 *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 $r \subseteq q$. **s** is nested

s2. $p \in \mathbf{s}$; and: if $q \in \mathbf{s}$, then $p \subseteq q$. **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 **s**1 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 **s**2 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 **s**3 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 **s**3 requirement **s**4 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.

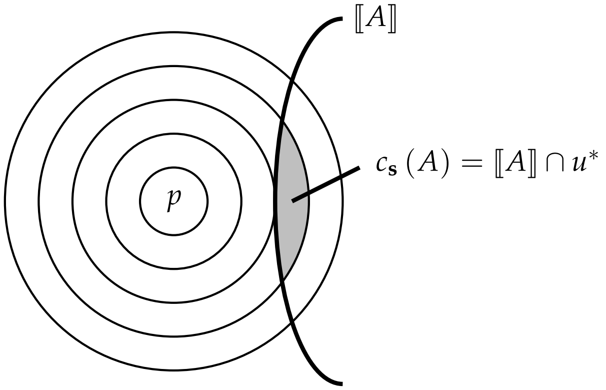


Figure 1: The possible worlds "closest" to the center p

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 $[\![\mathcal{B} * A]\!] = [\![T(c_{\mathbf{s}}(A))]\!] = u^* \cap [\![A]\!]$ 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).

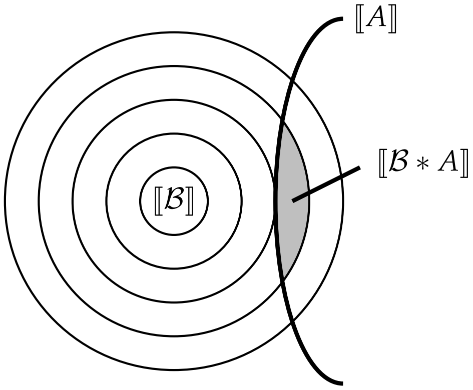


Figure 2: The strongest proposition believed after revising $\mathcal{B}$ by A

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 2 *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 $\mathbf{s}$ in $w_{\mathcal{L}}$ that is centered on $[\![\mathcal{B}]\!]$ and satisfies $\mathbf{s}1-\mathbf{s}4$ 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 $A \in \mathcal{L}$:*

$$\mathcal{B} * A = T\left(c_{\mathbf{s}}(A)\right).$$

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 $\mathbf{s}$ in $w_{\mathcal{L}}$ that is centered on $[\![\mathcal{B}]\!]$ and satisfies $\mathbf{s}1-\mathbf{s}4$ such that for all $A \in \mathcal{L}$:

$$\mathcal{B} * A = T\left(c_{\mathbf{s}}(A)\right).$$

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 1 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 **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 **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 $\{C \in \mathcal{L} : A \mathrel{\Box\!\!\rightarrow} C\}$ 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 $\{\omega \in q : \omega \leq \omega^*$ 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 $\mathcal{L}$:

≤5. If $\omega, \omega' \in [\![A]\!]$, then: $\omega \leq_{\mathcal{B}} \omega'$ just in case $\omega \leq_{\mathcal{B}*A} \omega'$.

≤6. If $\omega, \omega' \notin [\![A]\!]$, then: $\omega \leq_{\mathcal{B}} \omega'$ just in case $\omega \leq_{\mathcal{B}*A} \omega'$.

≤7. If $\omega \in [\![A]\!]$ and $\omega' \notin [\![A]\!]$ and if $\omega <_{\mathcal{B}} \omega'$, then $\omega <_{\mathcal{B}*A} \omega'$.

≤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 1 and 2 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 3 *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 $A \in \mathcal{L}$:*

$$\mathcal{B} * A = T\left(\left\{\omega \in [\![A]\!] : \omega \leq \omega^* \text{ for all } \omega^* \in [\![A]\!]\right\}\right).$$

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 $A \in \mathcal{L}$:

$$\mathcal{B} * A = T\left(\left\{\omega \in [\![A]\!] : \omega \leq \omega^* \text{ for all } \omega^* \in [\![A]\!]\right\}\right).$$

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 4 *Let $\mathcal{L}$ and $w_{\mathcal{L}}$ be as in Theorem 3. Suppose $*$ is a revision operator from $\wp(\mathcal{L}) \times \mathcal{L}$ into $\wp(\mathcal{L})$ that satisfies $*1-*8$. According to Theorem 3, 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 $A \in \mathcal{L}$, $\mathcal{B} * A = T\left(\left\{\omega \in [\![A]\!] : \omega \leq_{\mathcal{B}} \omega^* \text{ for all } \omega^* \in [\![A]\!]\right\}\right)$. 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, ms).

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, $\beta\left(A \mid C\right) > 0$. 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 $\varrho\left(\cdot\right) : \mathcal{A} \to \mathbb{N} \cup \{\infty\}$ 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 $\varrho_{E_i \uparrow z_i}\left(\cdot\right)$,*

$$\varrho_{E_i \uparrow z_i}\left(\cdot\right) = \min{}_{i \in I}\left\{\varrho\left(\cdot \cap E_i\right) + z_i - m\right\},$$

where $m = \min{}_{i \in I}\left\{z_i + \varrho\left(E_i\right)\right\}$.

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, $r(A \mid B) = r(A \mid \overline{B})$

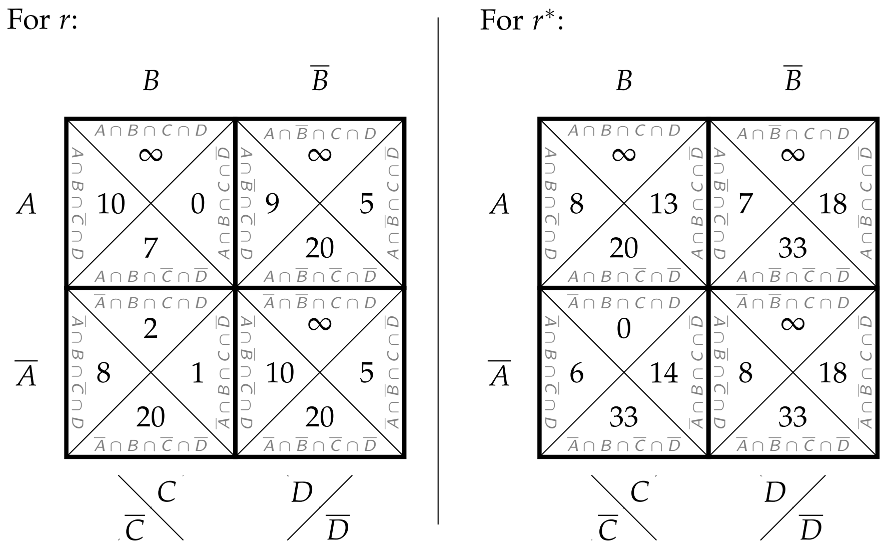

For r:
B
$\overline{B}$
A
$\overline{A}$
$A \cap B \cap C \cap D$
∞
$A \cap B \cap \overline{C} \cap D$
10
0
$A \cap B \cap C \cap \overline{D}$
7
$A \cap B \cap \overline{C} \cap \overline{D}$
$A \cap \overline{B} \cap C \cap D$
∞
$A \cap \overline{B} \cap \overline{C} \cap D$
9
5
$A \cap \overline{B} \cap C \cap \overline{D}$
20
$A \cap \overline{B} \cap \overline{C} \cap \overline{D}$
$\overline{A} \cap B \cap C \cap D$
2
$\overline{A} \cap B \cap \overline{C} \cap D$
8
1
$\overline{A} \cap B \cap C \cap \overline{D}$
20
$\overline{A} \cap B \cap \overline{C} \cap \overline{D}$
$\overline{A} \cap \overline{B} \cap C \cap D$
∞
$\overline{A} \cap \overline{B} \cap \overline{C} \cap D$
10
5
$\overline{A} \cap \overline{B} \cap C \cap \overline{D}$
20
$\overline{A} \cap \overline{B} \cap \overline{C} \cap \overline{D}$
C
$\overline{C}$
D
$\overline{D}$
For r^*:
B
$\overline{B}$
A
$\overline{A}$
$A \cap B \cap C \cap D$
∞
$A \cap B \cap \overline{C} \cap D$
8
13
$A \cap B \cap C \cap \overline{D}$
20
$A \cap B \cap \overline{C} \cap \overline{D}$
$A \cap \overline{B} \cap C \cap D$
∞
$A \cap \overline{B} \cap \overline{C} \cap D$
7
18
$A \cap \overline{B} \cap C \cap \overline{D}$
33
$A \cap \overline{B} \cap \overline{C} \cap \overline{D}$
$\overline{A} \cap B \cap C \cap D$
0
$\overline{A} \cap B \cap \overline{C} \cap D$
6
14
$\overline{A} \cap B \cap C \cap \overline{D}$
33
$\overline{A} \cap B \cap \overline{C} \cap \overline{D}$
$\overline{A} \cap \overline{B} \cap C \cap D$
∞
$\overline{A} \cap \overline{B} \cap \overline{C} \cap D$
8
18
$\overline{A} \cap \overline{B} \cap C \cap \overline{D}$
33
$\overline{A} \cap \overline{B} \cap \overline{C} \cap \overline{D}$
C
$\overline{C}$
D
$\overline{D}$

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.

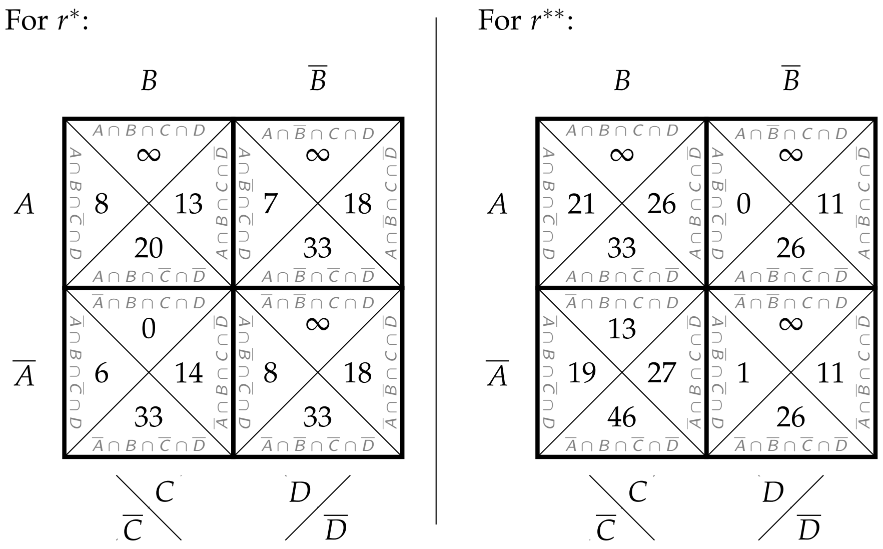


Figure 4: Sophia's new and newer ranks for various conjunctions

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,

$$\mathcal{B}_{\varrho(\cdot \mid C)} = \left\{A \in \mathcal{A} : \varrho\left(\overline{A} \mid C\right) > 0\right\}.$$

$\mathcal{B} \subseteq \mathcal{A}$ 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 5 *An ideal doxastic agent's belief set $\mathcal{B}_\varrho$ and conditional belief sets $\mathcal{B}_{\varrho(\cdot \mid C)}$ 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 (ms).

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

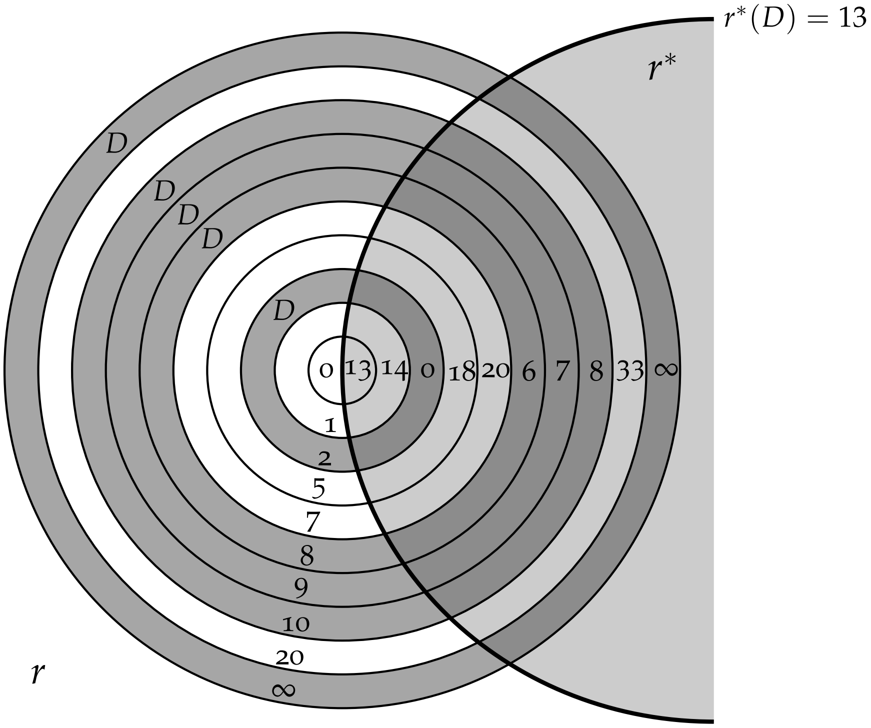


Figure 5: Sophia's ranking functions depicted as "numbered onions"

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

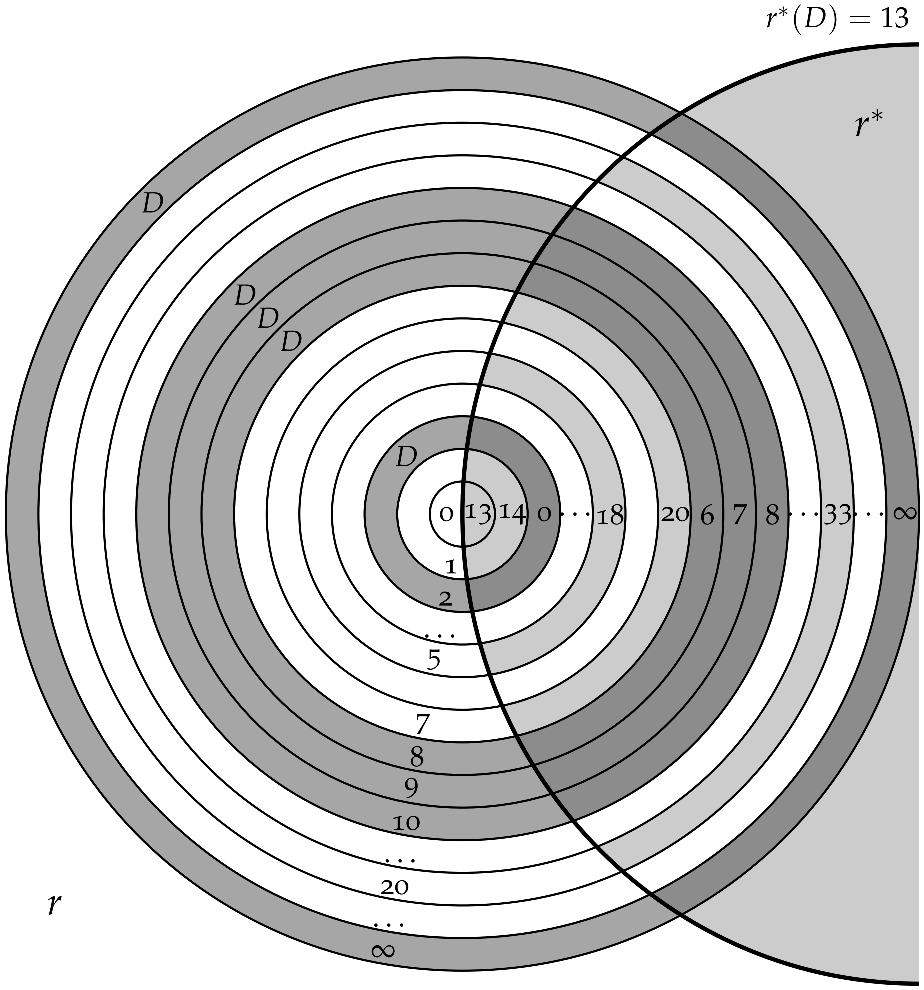


Figure 6: Sophia's ranking functions depicted as "numbered onions" with empty layers

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. (ms). 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. (ms). *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

This stub is a placeholder; work on this entry hasn't begun yet.

Lewis (1981) argues that 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.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

REFERENCES

Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, *59*(1), 5–30.

10

DOXASTIC LOGIC

Michael Caie

This stub is a placeholder; work on this entry hasn't begun yet.

Lewis (1981) argues that 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.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

REFERENCES

Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, *59*(1), 5–30.

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 Celsuis.
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*.

∴ *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 Paripatetics 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.

∴ 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.

∴ If not C, then not A.

And strengthening the antecedent is:

1. If A, then C.

∴ 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.

∴ 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.

∴ 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, 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.

∴ 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.

∴ 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 '$A \mathrel{\Box\!\!\rightarrow} C$' 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 (ms) 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.) $A \mathrel{\Box\!\!\rightarrow} C$ 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.

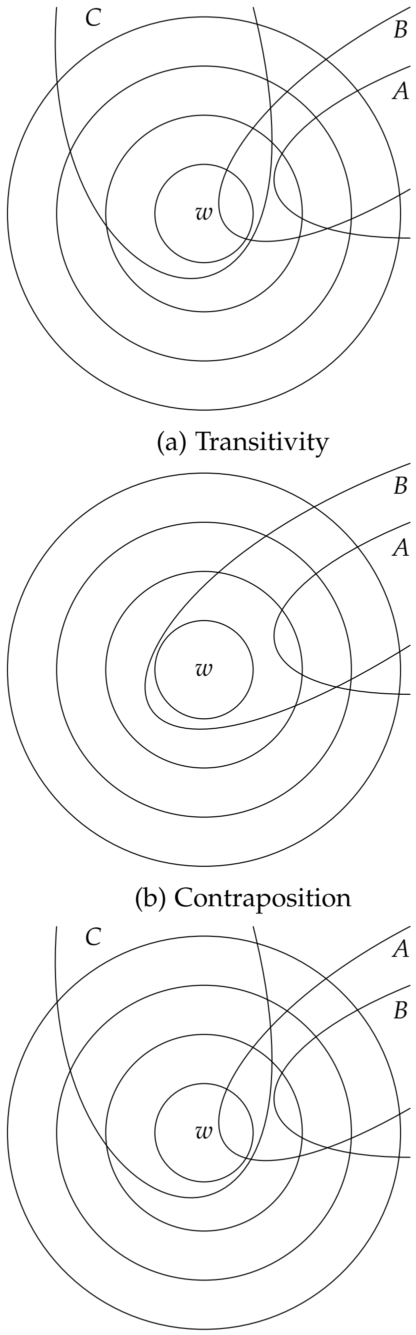
After all, Petrov's superiors were poised to launch the missiles in the event of an attack, and it seems that the phone lines and missile system were in working order. The only missing ingredient was the report from Petrov.

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

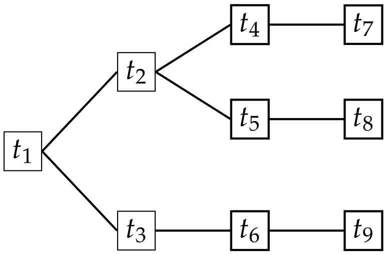


Figure 2: A model of branching time

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 dispensible; 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.

$$\textsc{catch} = \textsc{car}$$

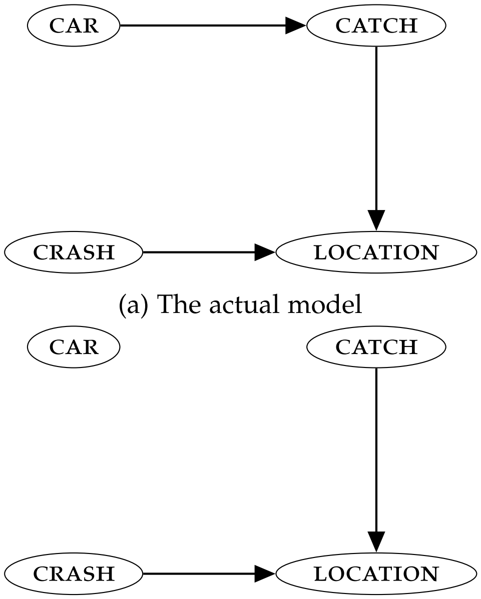
$$\textsc{location} = \begin{cases} 0 & \text{if } \textsc{catch} = 0, \\ 1 & \text{if } \textsc{catch} = 1 \text{ and } \textsc{crash} = 0, \\ 2 & \text{if } \textsc{catch} = 1 \text{ and } \textsc{crash} = 1. \end{cases}$$

(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 $\textsc{catch} = 1$. We then check whether the consequent is true (i.e., $\textsc{location} = 2$) in the resulting submodel.

First, we delete the structural equation for CATCH, turning CATCH into an exogenous variable. Our only remaining structural equation is

$$\textsc{location} = \begin{cases} 0 & \text{if } \textsc{catch} = 0, \\ 1 & \text{if } \textsc{catch} = 1 \text{ and } \textsc{crash} = 0, \\ 2 & \text{if } \textsc{catch} = 1 \text{ and } \textsc{crash} = 1. \end{cases}$$

(The graph for the resulting submodel is shown in Figure 3b.)

Second, we set the values of the variables. The antecedent requires that

$$\textsc{catch} = 1.$$

Since neither CAR nor CRASH is downstream from CATCH, we have

$$\textsc{car} = 0,$$
$$\textsc{crash} = 1.$$

Finally, the value of LOCATION is fixed by the structural equation. Since $\textsc{catch} = 1$ and $\textsc{crash} = 1$, we have

$$\textsc{location} = 2.$$

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 commmitments—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 $A \mathrel{\Box\!\!\rightarrow} B$ 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, $A \mathrel{\Box\!\!\rightarrow} C$ 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 $A \mathrel{\Box\!\!\rightarrow} C$. 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, $A \mathrel{\Box\!\!\rightarrow} C$ 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: $A \mathrel{\Box\!\!\rightarrow} B$ 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 Mondus 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.

∴ If I do my chores today, then the world will implode.

1. Dinner will be delicious.

∴ 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.

∴ 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 $A \to C$ 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-

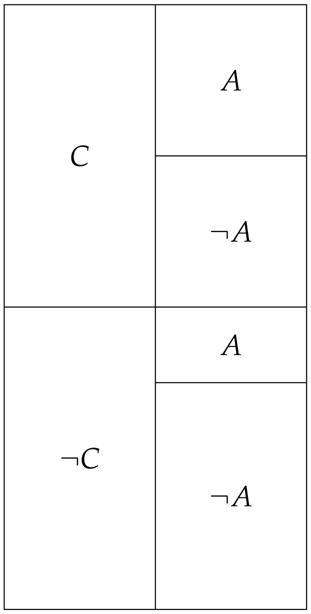


Figure 4: A probability space

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 $Pr(A|C)$.

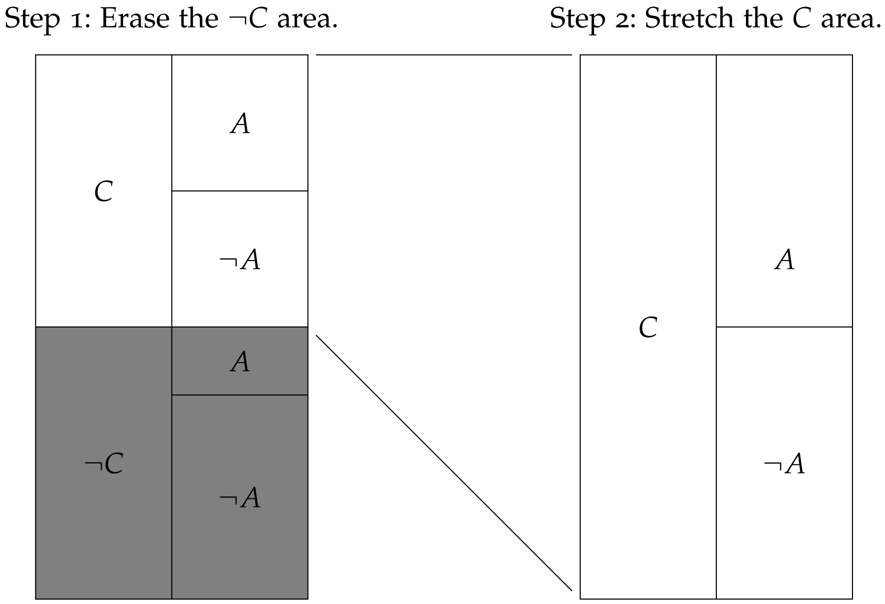


Figure 5: Calculating conditional probability

According to the Law of Total Probability (illustrated in Figure 6), for any two propositions X and Y,

$$Pr(Y) = Pr(Y|X) \times Pr(X) + Pr(Y|\neg X) \times Pr(\neg X). \tag{1}$$

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 $A \to C$ for Y in Equation 1 yields:

$$Pr(A \to C) = Pr(A \to C|C) \times Pr(C) + Pr(A \to C|\neg C) \times Pr(\neg C). \tag{2}$$

In other words, we can split the probability space into a C part and a $\neg C$ part, and figure out the probability of $A \to C$ 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

$$Pr_C(C|A) = 1. \tag{3}$$

Thus, by THE THESIS and Equation 3,

$$Pr_C(A \to C) = 1. \tag{4}$$

By the definition of Pr_C and equation 4,

$$Pr(A \to C|C) = 1. \tag{5}$$

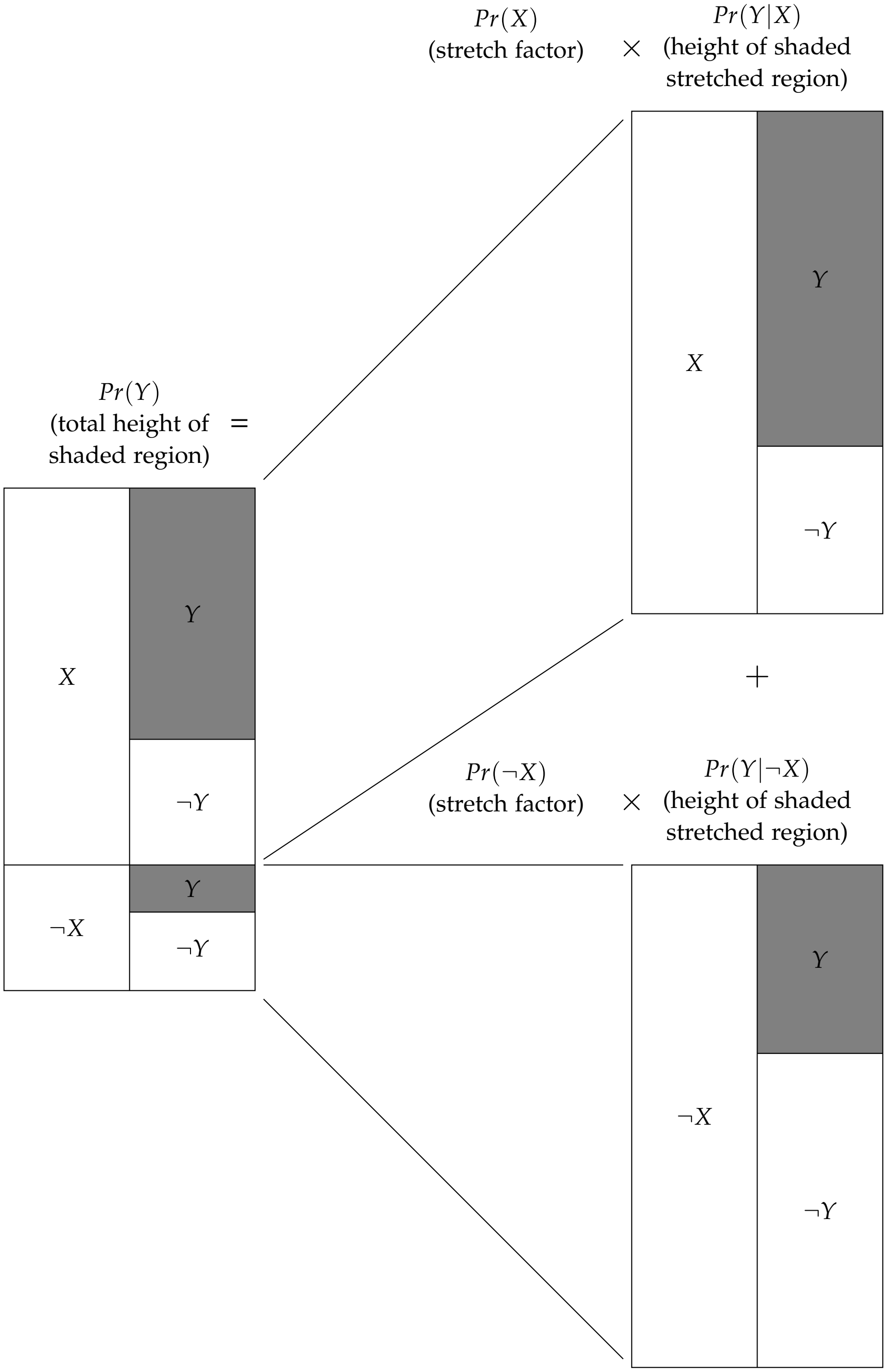


Figure 6: The Law of Total Probability

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

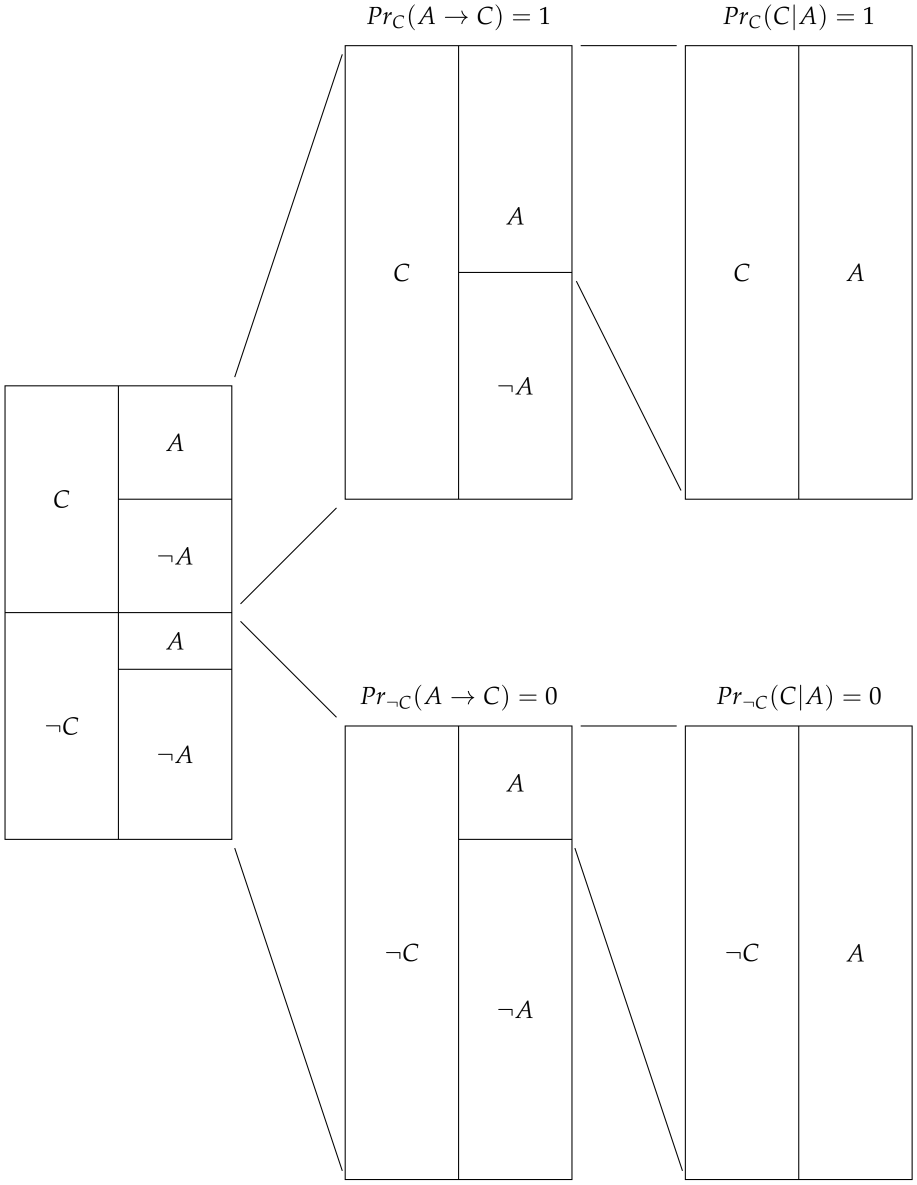


Figure 7: The Lewis triviality theorem illustrated

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.

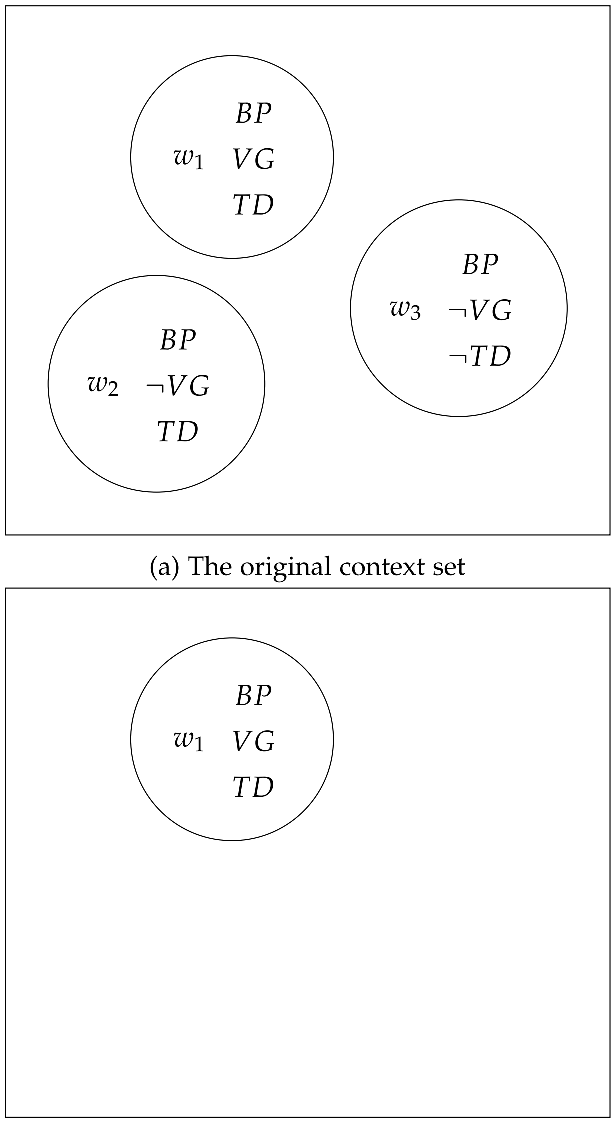
Figure 8 depicts the effect of asserting 'Veggie Garden is open today' on the context set. The original context set is shown in the rectangle at the top of the figure: the circles depict worlds. Each world is labeled with a set of propositions true at that world: 'BP' stands for 'Buddha's Palace is open'; '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 $A \to B$, 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. (ms). *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*.