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 is, what it means to take the appropriate means to one's ends, whatever

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

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

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

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

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

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

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

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

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

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

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

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

1.3 *Some Applications*

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

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

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

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

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

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

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

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

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

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

1.4 *Formulating Decision Problems*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2 REPRESENTATION THEOREMS

2.1 *The Preference Relation*

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2 *Von Neumann and Morgenstern*

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

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

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

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

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

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

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

2.3 *Savage*

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

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

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

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

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

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

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

2.4 *Jeffrey, Bolker, and Joyce*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3 INTERPRETATIONS OF EXPECTED UTILITY THEORY

3.1 *Interpretations of Utility*

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

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

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

20 Bernoulli proposed a gamble in which a coin is thrown repeatedly. If it lands heads the first time, the player gets \$2. If it lands tails, the prize is doubled, and the coin thrown again. This procedure is repeated indefinitely. The expected value of the resulting gamble is thus $\$2 \cdot \frac{1}{2} + \$4 \cdot \frac{1}{4} + \$8 \cdot \frac{1}{8} + \ldots$, 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 Section 4 and Section 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

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.

favour of probabilistic beliefs, apart from the plausibility of the axioms 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,

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 (this volume) 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 $D_0, D_1, D_2, \ldots, D_{1000}$. 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:

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

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 Section 5.2 and Section 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 Table 6 and Table 7.

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

Table 6: Allais Paradox: First Choice

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

Table 7: Allais Paradox: Second Choice

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

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

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

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

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

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

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

5.3 *Separability and Risk Aversion*

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

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

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

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

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

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

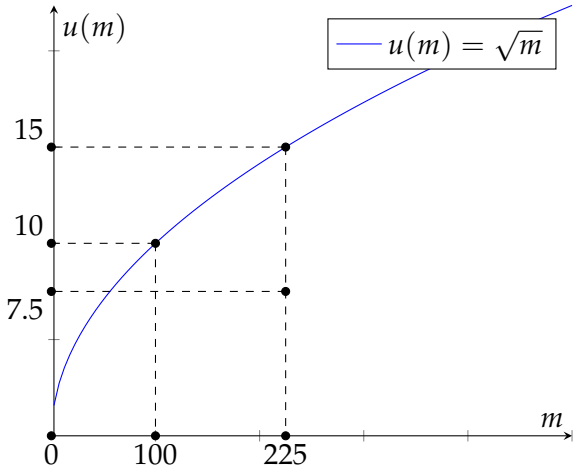


Figure 1: A concave utility function

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

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

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

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

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

6 ALTERNATIVES TO EXPECTED UTILITY THEORY

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

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

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

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

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

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

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

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

7 DYNAMIC CHOICE

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

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

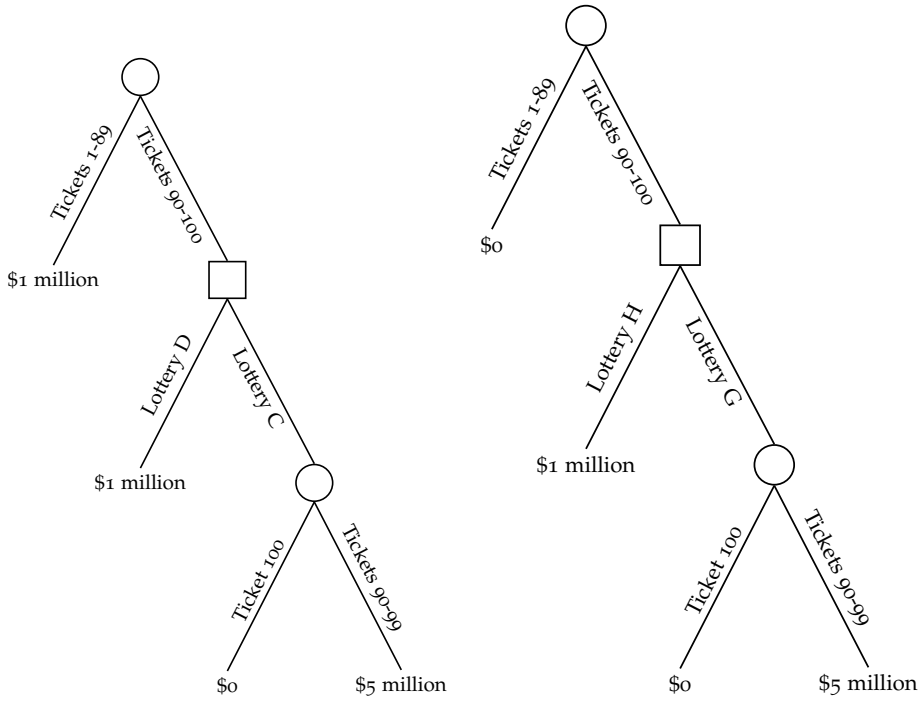
7.1 *Dynamic Arguments in Favour of Separability*

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

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

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

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



(a) First Choice (b) Second Choice

Figure 2: Dynamic Allais Problem

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

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

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

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

NEC (Normal-form/Extensive-form Coincidence). In any dynamic decision problem, the agent should choose the same as she would, were she to simply choose one course of action at the beginning of the decision problem.

SEP (Dynamic Separability). Within dynamic decision problems, the agent treats continuation trees as if they were new trees.

DS (Dynamic Consistency). The agent does not make plans she foreseeably will not execute.

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

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

49 See, for instance, Wakker (1988).

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

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

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

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

7.2 *Time Preferences and Discounting*

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

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

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

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

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

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

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

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

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

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

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

8 CONCLUDING REMARKS

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

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

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

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

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

ACKNOWLEDGEMENTS

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

REFERENCES

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

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

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

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

Andreou, C. (2012). Dynamic choice. *The 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.), *The Stanford encyclopedia of philosophy* (Summer 2015). Retrieved from http://plato.stanford.edu/archives/sum2015/entries/imprecise-probabilities/

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Elga, A. (2010). Subjective probabilities should be sharp. *Philosophers Imprint*, *10*(5), 1–11.

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

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

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

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

Fishburn, P. (1981). Subjective expected utility: A review of normative theories. *Theory and Decision*, *13*, 139–199.

Fishburn, P. (1998). Stochastic utility. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory* (Vol. 1). Kluwer.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*(2), 351–401.

Gibbard, A. & Harper, W. (1978/1981). Counterfactuals and two kinds of expected utility. In W. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs: Conditionals, belief, decision, chance, and time* (pp. 153–190). Reidel.

Gigerenzer, G., Todd, P. M., & Group, A. R. (2000). *Simple heuristics that make us smart*. Oxford University Press.

Gollier, C. (2002). Discounting an uncertain future. *Journal of Public Economics*, *85*(2), 149–166.

Hajek, A. (2008). Arguments for – or against – probabilism. *British Journal for the Philosophy of Science*, *59*(4), 793–819.

Hammond, P. (1988). Consequentialist foundations for expected utility. *Theory and Decision*, *25*, 25–78.

Hare, R. (1981). *Moral thinking*. Oxford University Press.

Hausman, D. (2000). Revealed preference, belief, and game theory. *Economics and Philosophy*, *16*(1), 99–115.

Holton, R. (2009). *Willing, wanting, waiting*. Oxford University Press.

Horgan, T. (1981/1985). Counterfactuals and newcomb's problem. In *Paradoxes of rationality and cooperation: Prisoner's dilemma and newcomb's problem* (pp. 159–182). University of British Columbia Press.

Horwich, P. (1987). *Asymmetries in time*. MIT Press.

Jackson, F. (1991). Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics*, *101*(3), 461–482.

Jeffrey, R. (1965/1983). *The logic of decision* (2nd). University of Chicago Press.

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

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

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

Joyce, J. M. & Gibbard, A. (1998). Causal decision theory. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory* (Vol. 1). Kluwer.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kaplan, M. (1983). Decision theory as philosophy. *Philosophy of Science*, *50*, 549–577.

Kaplan, M. (1996). *Decision theory as philosophy*. Cambridge University Press.

Kavka, G. (1983). The toxin puzzle. *Analysis*, *43*(1), 33–36.

Koopmans, T. (1960). Stationary ordinal utility and impatience. *Econometrica*, *28*, 287–309.

Levi, I. (1980). *The enterprise of knowledge*. Cambridge, MA: 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 probabilities. In R. Pettigrew & J. Weisberg (Eds.), *The open handbook of formal epistemology*. PhilPapers.

Mas-Colell, A., Whinston, M., & Green, J. (1995). *Microeconomic theory* (1st ed.). Oxford University Press.

McClennen, E. (1990). *Rationality and dynamic choice: Foundational explorations*. Cambridge University Press.

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

Mill, J. S. (1861/1998). *Utilitarianism* (R. Crisp, Ed.). Oxford University Press.

Morrison, D. (1967). On the consistency of preferences in allais' paradox. *Behavioral Science*, *12*(5), 373–383.

Moss, S. (2015). Time-slice epistemology and action under indeterminacy. *Oxford Studies in Epistemology*.

Nordhaus, W. (2007). A review of the stern review on the economics of global warming. *Journal of Economic Literature*, *155*, 686–702.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in honor of carl g. hempel* (pp. 114–115). Synthese Library. Reidel.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.

Pedersen, A. & Wheeler, G. (2014). Demystifying dilation. *Erkenntnis*, *79*(6), 1305–1342.

Pettigrew, R. (2011). Epistemic utility arguments for probabilism. *The Stanford Encyclopedia of Philosophy*. Retrieved from http://plato.stanford.edu/archives/win2011/entries/epistemic-utility/

Pettigrew, R. (2014). *Risk, rationality, and expected utility theory*. APA author meets critic session.

Pettit, P. (1991). Decision theory and folk psychology. In M. Bacharach & S. Hurley (Eds.), *Foundations of decision theory: Issues and advances* (pp. 147–175). Blackwell.

Price, H. (1991). Agency and probabilistic causality. *British Journal for the Philosophy of Science*, *42*(2), 157–176.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, *3*(4), 323–343.

Quinn, W. (1990). The puzzle of the self-torturer. *Philosophical Studies*, *59*(1).

Rabin, M. (2000). Risk aversion and expected utility: A calibration theorem. *Econometrica*, *68*(1281–1292).

Rabin, M. & Thaler, R. (2001). Anomalies: Risk aversion. *Journal of Economic Perspectives*, *15*, 219–232.

Rabinowicz, W. (1995). To have one's cake and eat it, too: Sequential choice and expected utility violations. *Journal of Philosophy*, *92*(11), 586–620.

Ramsey, F. P. (1926/2010). Truth and probability. In A. Eagle (Ed.), *Philosophy of probability: Contemporary readings* (pp. 52–94). Routledge.

Sahlin, N.-E. & Weirich, P. (2014). Unsharp sharpness. *Theoria*, *80*, 100–103.

Salles, M. (1998). Fuzzy utility. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory* (Vol. 1). Kluwer.

Samuelson, P. (1937). A note on measurement of utility. *Review of Economic Studies*, *4*, 155–161.

Samuelson, P. (1963). Risk and uncertainty: A fallacy of large numbers. *Scientia*, *98*(108–113).

Sartre, J.-P. (1945/2007). *Existentialism is a humanism* (A. Elkaïm-Sartre, Ed.). Yale University Press.

Savage, L. (1954). *The foundations of statistics*. Wiley.

Schmidt, U. (2004). Alternatives to expected utility theory: Formal theories. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory* (pp. 757–837). Kluwer.

Seidenfeld, T. (1988). Decision theory without "independence" or without "ordering". *Economics and Philosophy*, *4*, 267–290.

Seidenfeld, T. & Wasserman, L. (1993). Dilation for sets of probabilities. *The Annals of Statistics*, *21*(3), 1139–1154.

Sepielli, A. (2013). Moral uncertainty and the principle of equity among moral theories. *Philosophy and Phenomenological Research*, *86*(3), 580–589.

Simon, H. (1976). From substantive to procedural rationality. In T. J. Kastelein, S. K. Kuipers, W. A. Nijenhuis, & G. R. Wagneaar (Eds.), *25 years of economic theory* (Vol. 2, pp. 65–86). Springer US.

Singer, P. (1993). *Practical ethics*. Cambridge University Press.

Spohn, W. (1977). Where luce and krantz do really generalize savage's decision model. *Erkenntnis*, *11*, 113–134.

Stalnaker, R. (1972/1981). Letter to david lewis. In W. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs: Conditionals, belief, decision, chance, and time* (pp. 151–152). Reidel.

Stern, N. (2007). *The economics of climate change*. Cambridge University Press.

Strotz, R. H. (1956). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, *23*(3), 165–180.

Sugden, R. (2004). Alternatives to expected utility theory: Foundations. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory* (pp. 685–755). Kluwer.

Tenenbaum, S. & Raffman, D. (2012). Vague projects and the puzzle of the self-torturer. *Ethics*, *123*(1), 86–112.

Thaler, R. (1981). Some empirical evidence on dynamic inconsistency. *Economic Letters*, *8*(3), 351–401.

Troffaes, M. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, *45*, 17–29.

Tversky, A. & Kahneman, D. (1974). Judgements under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Velleman, D. (1993/2000). The story of rational action. In *The possibility of practical reason*. Oxford University Press.

von Neumann, J. & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Voorhoeve, A. & Binmore, K. (2006). Transitivity, the sorites paradox, and similarity-based decision-making. *Erkenntnis*, *64*(1), 101–114.

Wakker, P. (1988). Nonexpected utility as aversion to information. *Journal of Behavioral Decision Making*, *1*, 169–175.

Weber, M. [Max]. (1922/2005). *Wirtschaft und gesellschaft. grundriss der verstehenden soziologie* (A. Ulfig, Ed.). Zweitausendeins-Verlag.

Weber, M. [Michael]. (1998). The resilience of the allais paradox. *Ethics*, *109*(1), 94–118.

Weirich, P. (1986). Expected utility and risk. *British Journal for the Philosophy of Science*, *37*, 419–442.

Weitzman, M. (2009). On modeling and interpreting the economics of catastrophic climate change. *The Review of Economics and Statistics*, *91*(1), 1–19.

White, R. (2010). Evidential symmetry and mushy credence. *Oxford Studies in Epistemology*, *3*, 161–186.

Williams, J. R. G. (2014). Decision-making under indeterminacy. *Philosophers' Imprint*, *14*(4), 1–34.

Williamson, J. (2010). *In defense of objective bayesianism*. Oxford University Press.

Zynda, L. (2000). Representation theorems and realism about degrees of belief. *Philosophy of Science*, *67*(1), 45–69.