DECISION THEORY

Lara Buchak

Forthcoming in the *Oxford Handbook of Probability and Philosophy*, eds. Christopher Hitchcock and Alan Hájek. OUP.

0. Introduction

Decision theory has at its core a set of mathematical theorems that connect rational preferences to functions with certain structural properties. The components of these theorems, as well as their bearing on questions surrounding rationality, can be interpreted in a variety of ways. Philosophy's current interest in decision theory represents a convergence of two very different lines of thought, one concerned with the question of how one ought to act, and the other concerned with the question of what action consists in and what it reveals about the actor's mental states. As a result, the theory has come to have two different uses in philosophy, which we might call the *normative* use and the *interpretive* use. It also has a related use that is largely within the domain of psychology, the *descriptive* use.

The first two sections of this essay examine the historical development of normative decision theory and the range of current interpretations of the elements of the theory, while the third section explores how modern normative decision theory is supposed to capture the notion of rationality. The fourth section presents a history of interpretive decision theory, and the fifth section examines a problem that both uses of decision theory face. The sixth section explains the third use of decision theory, the descriptive use. Section seven considers the relationship between the three uses of decision theory. Finally, section eight examines some modifications to the standard theory and the conclusion makes some remarks about how we ought to think about the decision-theoretic project in light of a proliferation of theories.

1. Normative Decision Theory

The first formal decision theory was developed by Blaise Pascal in correspondence with Pierre Fermat about "the problem of the points," the problem of how to divide up the stakes of players involved in a game if the game ends prematurely.[1] Pascal proposed that each gambler

[1] See Fermat and Pascal (1654).

should be given as his share of the pot the monetary expectation of his stake, and this proposal can be generalized to other contexts: the monetary value of a risky prospect is equal to the expected value of that prospect. Formally, if L = {x_1$, p_1; x_2$, p_2; … } represents a "lottery" which yields x_i$ with probability p_i, then its value is:

$$EV(L) = \sum_{i=1}^{\infty} p_i \, x_i$$

This equivalence underlies a prescription: when faced with two lotteries, you ought to prefer the lottery with the higher expected value, and be indifferent if they have the same expected value. More generally, you ought to *maximize expected value*.

This norm is attractive for a number of reasons. For one, it enjoins you to make the choice that would be better over the long run if repeated: over the long run, repeated trials of a gamble will average out to their expected value. For another, going back to the problem of the points, it ensures that players will be indifferent between continuing the game and leaving with their share. But there are several things to be said against the prescription. One is that it is easy to generate a lottery whose expected value is infinite, as shown by the St. Petersburg Paradox (first proposed by Nicolas Bernouilli). Under the norm in question, one ought to be willing to pay any finite amount of money for the lottery {$1, ½; $2, ¼; $4, 1/8; … }, but most people think that the value of this lottery should be considerably less. A second problem is that the prescription does not seem to account for the fact that whether one should take a gamble seems to depend on what one's total fortune is: one ought not risk one's last dollar for an even chance at $2, if losing the dollar means that one will be unable to eat. Finally, the prescription doesn't seem to adequately account for the phenomenon of risk-aversion: most people would rather have a sure thing sum of $x than a gamble whose expectation is $x (for example, $100 rather than {$50, ½; $150, ½}) and don't thereby seem irrational.

In response to these problems, Daniel Bernouilli (1738) and Gabriel Cramer (see Bernouilli 1738: 33) each independently noted that the amount of satisfaction that money brings diminishes the more money one has, and proposed that the quantity whose expectation one ought to maximize is not money itself but rather the "utility" of one's total wealth. (Note that for Bernouilli, the outcomes are total amounts of wealth rather than changes in wealth, as they were for Pascal.) Bernouilli proposed that an individual's utility function of total wealth is u($x) = log($x). Therefore, the new prescription is to maximize:

$$EU(L) = \sum_{i=1}^{\infty} p_i \, u(\$x_i) = \sum_{i=1}^{\infty} p_i \, log(\$x_i)$$

This guarantees that the St. Petersburg lottery is worth a finite amount of money; that a gamble is worth a larger amount of one's money the wealthier one is; and that the expected utility of any lottery is less than the utility of its monetary expectation.

Notice that the norm associated with this proposal is objective in two ways: it takes the probabilities as given, and it assumes that everyone should maximize the same utility function. One might reasonably wonder, however, whether everyone does get the same amount of satisfaction from various amounts of money. Furthermore, non-monetary outcomes are plausibly of different value to different people, and the proposal tells us nothing about how we ought to value lotteries with non-monetary outcomes. A natural thought is to revise the norm to require that one maximize the expectation of one's own, *subjective* utility function, and to allow that the utility function take any outcome as input.

The problem with this thought is that it is not clear that individuals have access to their precise utility functions through introspection. Happily, it turns out that we can implement the proposal without such introspection: John von Neumann and Oskar Morgenstern (1944) discovered a *representation theorem* that allows us to determine whether an agent is maximizing expected utility merely from her pair-wise preferences, and, if she is, allows us to determine an agent's entire utility function from these preferences. Von Neumann and Morgenstern identified a set of axioms on preferences over lotteries such that if an individual's preferences conform to these axioms, then there exists a utility function of outcomes, unique up to positive affine transformation, that represents her as an expected utility maximizer.[2] The utility function represents her in the following sense: for all lotteries L_1 and L_2, the agent weakly prefers L_1 to L_2 if and only if L_1 has at least as high an expected utility as L_2 according to the function. Thus, we can replace expected objective utility maximization with expected subjective utility maximization as an implementable norm, even if an agent's utility function is opaque to her.

[2] I will often talk about an agent's utility function when strictly speaking I mean the family of utility functions that represents her. However, facts about the utility function that are not preserved under affine transformation, such as the zero point, will not count as "real" facts about the agent's utility values.

Leonard Savage's (1954) representation theorem took the theory one step further. Like von Neumann and Morgenstern, Savage allowed that an individual's values were up to her. But Savage was interested not primarily in how an agent should choose between lotteries when she is given the exact probabilities of outcomes, but rather in how an agent should choose between ordinary acts when she is uncertain about some feature of the world: for example, how she should choose between breaking a sixth egg into her omelet and refraining from doing so, when she does not know whether or not the egg is rotten. Savage noted that an act leads to different outcomes under different circumstances, and, taking an outcome to be specified so as to include everything an agent cares about, he defined the technical notion of an act as a function from possible states of the world to outcomes.[3] For example, the act of breaking the egg is the function {egg is good -> I eat a 6-egg omelet; egg is rotten -> I throw away the omelet}. More generally, we can represent an act f as $\{x_1, E_1; \ldots ; x_n, E_n\}$, where E_i are mutually exclusive and exhaustive events (an event being a set of states), and each state in E_i results in outcome x_i under act f.[4] Savage's representation theorem shows that an agent's preferences over these acts suffice to determine both her subjective utility function of outcomes and her subjective probability function of events, provided her pair-wise preferences conform to the axioms of his theorem.[5] Formally, *u* and *p* represent an agent's preferences if and only if she prefers the act with the highest expected utility, relative to these two functions:

$$EU(f) = \sum_{i=1}^{n} p(E_i)u(x_i)$$

Savage's theory therefore allows that both the probability function and the utility function are subjective. The accompanying prescription is to maximize expected utility, relative to these two functions.

Since Savage, other representation theorems for subjective expected utility theory have been proposed, most of which are meant to respond to some supposed philosophical problem

[3] Savage used the terminology "consequence" where I am using "outcome."

[4] Savage also treats the case in which the number of possible outcomes of an act is not finite (Savage (1954: 76-82), although his treatment requires bounding the utility function. Assuming each act has a finite number of outcomes will simplify the discussion.

[5] Again, the utility function is unique up to positive affine transformation. The probability function is fully unique.

with Savage's theory.[6] One set of issues surrounds what we should prefer when utility is unbounded and acts can have an infinite number of different outcomes, or when outcomes can have infinite utility value.[7] Another set of issues concerns exactly what entities are the relevant ones to assign utility and probability to in decision-making. The developments in this area begin with Richard Jeffrey (1965), who objected to Savage's separation between states, outcomes, and acts, and argued that the same objects ought to be the carriers of both probability and value. Jeffrey proposed a theory on which both the probability and utility function take propositions as inputs. Axiomatized by Ethan Bolker (see Jeffrey 1965: 142-3, 149), Jeffrey's theory enjoins the agent to maximize:[8]

$$u(A) = \sum_{i=1}^{\infty} p(S_i|\, A)u(S_i\&A)$$

where S_i and A both stand for arbitrary propositions (they range over the same set), but S_i is to play the role of a state and A of an act. Bolker's (1965-67) representation theorem provides axioms on a preference relation over the set of propositions that allow us to extract p and u, although the uniqueness conditions are more relaxed than in the aforementioned theories. Jeffrey proposed that we ought to interpret the items that an agent has preferences over as "news items"; so, for example, one is asked whether one would prefer the news that one breaks the egg into one's omelet or that one does not. The connection to action, of course, is that one has the ability to create the news when it comes to propositions about acts one is deciding between.

Certain features of Jeffrey's interpretation are inessential to the maximization equation. It is not necessary to follow Jeffrey in interpreting preferences as being about news items. Nor is there consensus that p and u ought to have as their domain the same set of objects.[9] For

[6] See Fishburn (1981) for a helpful catalogue of some of these.

[7] See, for example, Vallentyne (1993), Nover and Hájek (2004), Bartha (2007), Colyvan (2008), and Easwaran (2008).

[8] Jeffrey used a slightly different, but equivalent, formulation. He also used functions named *prob* and *des* rather than p and u, but the difference is terminological.

[9] Of course, while this feature is inessential to Jeffrey's maximization equation as written above, it is essential to Bolker's representation theorem.

example, while it is clear that we can assign utility values to acts under our own control, Wolfgang Spohn (1977) and Isaac Levi (1991) each argue that we cannot assign these probability.

Another issue with Jeffrey's theory has been the source of a significant development in decision theory. Because the belief component of Jeffrey's theory corresponds to conditional probabilities of states given acts, this component will have the same numerical value whether an act causes a particular outcome or is merely correlated with it. Therefore, agents will rank acts that are merely correlated with preferred outcomes the same as acts that tend to cause preferred outcomes. This is why Jeffrey's theory has come to be known as *evidential expected utility (EEU) theory*: one might prefer an act in part because it gives one evidence that one's preferred outcome obtains. Many have argued that this feature of the theory is problematic, and the problem can be brought out by a case know as *Newcomb's problem* (first discussed by Robert Nozick (1969)).

Here is the case. You are presented with two boxes, one closed and one open so that you can see its contents; and you may choose either to take only the closed box, or to take both boxes. The open box contains $1000. The contents of the closed box were determined as follows. A predictor predicted ahead of time whether you would choose to take the one box or both; if he predicted that you would take just the closed box, he's put $1M in the closed box, but if he predicted that you would take both, he's put nothing in the closed box. Furthermore, you know that many people have faced this choice and that he's predicted correctly every time.

Assuming you prefer more money to less, evidential EU theory recommends that you take only one box, since the relevant conditional probabilities are one and zero (or close thereto): p(there is $1M in the closed box | you take one box) ≈ 1, and p(there is $0 in the closed box | you take two boxes) ≈ 1. But many think that this is the wrong recommendation. After all, the closed box already contains what it contains, so your choice is between receiving whatever is in that box and receiving whatever is in that box plus an extra thousand dollars. Taking two boxes *dominates* taking one box, the argument goes: it is better in every possible world. We might diagnose the mis-recommendation of EEU theory as follows: p($1M | one box) is high because taking one box is correlated with getting $1M, but taking one box cannot cause $1M to be in the box because the contents of the box have been already determined; and so EEU gets the recommendation wrong because conditional probability does not distinguish between correlation

and causation. Not everyone accepts that two-boxing is the correct solution: those who advocate one-boxing point out that those who take only one box end up with more money, and since rationality ought to direct us to the action that will result in the outcome we prefer, it is rational to take one box. However, those who advocate two-boxing reply that even though those who take only one box end up with more money, this is a case in which they are essentially rewarded for behaving irrationally.

For those who advocate two-boxing, one way to respond to this problem is to modify evidential EU theory by adding a condition like *ratifiability* (Jeffrey 1983: 19-20), which says that one can only pick an act if it still has the highest EU on the supposition that one has chosen it. However, this does not solve the general problem of distinguishing *A*'s being evidentially correlated with *S* from *A*'s causing *S*. To yield the two-boxing recommendation in the Newcomb case, as well as to address the more general problem, Allan Gibbard and William Harper (1978) proposed *causal expected utility theory*, drawing on a suggestion of Robert Stalnaker (1972). Causal expected utility theory enjoins an agent to maximize:

$$u(A) = \sum_{i=1}^{\infty} p(A\square \longrightarrow S_i)u(S_i\&A)$$

where $p(A\square \longrightarrow S_i)$ stands for the probability of the counterfactual "If I were to do A then S_i would happen."[10] Armendt (1986) proved a representation theorem for the new theory, and Joyce (1999) provided a unified representation theorem for both evidential and causal expected utility theory.

Causal expected utility theory recommends two-boxing if the pair of counterfactuals "If I were to take one box, there would be \$0 in the closed box" and "If I were to take two boxes, there would be \$0 in the opaque box" are assigned the same credence, and similarly for the corresponding pair involving \$1M in the opaque box. This captures the idea that the contents of the closed box are independent of the agent's choices, and vindicates the reasoning that taking two boxes will result in an extra thousand dollars:

$$u(1\ box) = p(1\ box\square \longrightarrow C(\$0))u(\$0) + \ p(1\ box\square \longrightarrow C(\$1M))u(\$1M)$$

$$u(2\ boxes) = p(2\ boxes\square \longrightarrow C(\$0))u(\$1K) + \ p(2\ boxes\square \longrightarrow C(\$1M))u(\$1M + \$1K)$$

[10] Other formulations of causal decision theory include that of Lewis (1981) and Skyrms (1982).

To get this result, it is important that the counterfactuals in question are what Lewis (1981) calls "causal" counterfactuals rather than "back-tracking" counterfactuals. For there are two senses in which the counterfactuals "If I were to take one box, there would be $0 in the closed box" and "If I were to take two boxes, there would be $0 in the closed box" can be taken. In the back-tracking sense, I would reason from the supposition that I take one box *back to* the conclusion that the predictor predicted I would take one box, and I would assign a very low credence to the former counterfactual; but I would by the same reasoning assign a very high credence to the latter. In the causal sense, I would hold fixed facts about the past, since I cannot now cause past events, and the supposition that I take one box would not touch facts about what the predictor did; and by this reasoning I would assign equal credence to both counterfactuals.

It is worth considering how Savage's original theory would treat the Newcomb problem. Savage's theory uses unconditional credences, but correctly resolving the decision problem depends on specifying the states, outcomes, and acts in such a way that states are independent of acts. So, in effect, Savage's theory is a kind of causal decision theory. Indeed, Lewis (1981: 13) thought of his version of causal decision theory as returning to Savage's unconditional credences, but building the correct partition of states into the formalism itself rather than relying on an extra-theoretical principle about entity-specification.

All of the modifications mentioned here leave the basic structure of the theory intact – probability and utility are multiplied and then summed – and treat both *p* and *u* as subjective, so we can put them all under the heading of *subjective expected utility theory* (hereafter EU theory).

How should we understand the two functions, *p* and *u*, involved in EU theory? In the case of the probability function, although there is debate over whether *p* is *defined* by preferences ("betting behavior") via a representation theorem or whether preferences are merely a way to *discover p*, it is widely acknowledged that *p* is supposed to represent an agent's beliefs. In the case of the utility function, there are two philosophical disagreements. First, there is a disagreement about whether the utility function is defined by or merely discovered from preferences. If one thinks the utility function is defined by preferences, there is a further question about whether it is merely a convenient way to represent preferences or whether it refers to some pre-theoretical, psychologically real entity like strength of desire or perceived amount of satisfaction. Functionalists, for example, hold that utility is (at least partially) constituted by its role in preferences but also hold that utility is psychologically real. Since the

term "realism" is sometimes used to refer to the view that utility is independent of preferences, and sometimes used to refer to the view that utility is a psychologically real quantity, I will use the following terminology. I will call the view that utility is discovered from preferences *non-constructivist realism* and the view that utility is defined from preferences *constructivism*. I will call the view that utility does correspond to something psychologically real *psychological realism* and the view that utility does not refer to any real entity *formalism*.[11] Non-constructive realist views will be psychologically realist as well; however, functionalism counts as a constructivist, psychological realist view. Hereafter, when I am speaking of psychological realist theories, I will speak as if utility corresponds to desire, just as subjective probability corresponds to belief, though there may be other proposals about what utility corresponds to.

2. The Norm of Normative Decision Theory

Representation theorems connect preferences conforming to a set of axioms on the one hand to utilities and probabilities such that preferences maximize expected utility on the other. Thus, representation theorems give us an equivalent way to state the prescription that one ought to maximize expected utility: one ought to have preferences that accord with the axioms. The upshot of this equivalence depends on which theory of utility one adopts. For psychological realists, both formulations of the norm may have some bite: the "maximization" norm is a norm about how preferences ought to be related to beliefs and desires, and the "axiom" norm is an internal norm on preferences. For formalists, since there is really no such thing as utility, the *only* sensible formulation of the norm is as the axiom norm. But for both interpretations, an important advantage of the representation theorems is that judgments about whether an agent did what she ought, as well as arguments about whether EU theory identifies a genuine prescription, can focus on the axioms.

A point of clarification about the equivalent ways to state the norm of EU theory is needed. "Maximize expected utility" admits of two readings, one narrow-scope ("Given your utility function, maximize its expectation") and one wide-scope ("Be such that there is a utility

[11] The term constructivism comes from Dreier (1996), and the term formalism comes from Hanssen (1988). Bermúdez (2009) uses "operationalism" for what I call formalism. Zynda (2000) uses "strong realism" for what I call non-constructivist realism and "weak realism" for what I call psychological realism.

function whose expectation you maximize"). And the axiom norm is only equivalent to the wide-scope maximization norm. For the narrow-scope norm to apply in cases in which one fails to live up to it, one must be able to count as having a utility function even when one does not maximize its expectation. Clearly, this is possible according to the non-constructivist realist.[12] I will also show that in many cases, it is possible according to all psychological realists.

We can note the historical progression: in its original formulations, decision theory was narrow-scope, and the utility function (or its analogue) non-constructivist realist: money had an objective and fixed value. However, wide-scope, constructivist views are most popular nowadays. Relatedly, whereas originally a central justification of the norm was via how well someone who followed it did over the long run, such justifications have fallen out of favor and have been replaced by justification via arguments for the axioms.

One final point of clarification. So far, we have been talking about the relationship of beliefs and desires to preferences. But one might have thought that the point of a theory about decision-making was to tell individuals what to *choose*. The final piece in the history of decision theory concerns the relationship between preference and choice. In the heyday of behaviorism, Samuelson's (1938) idea of "revealed preference" was that preference can be cashed out in terms of what you would choose. However, nowadays philosophers mostly think the connection between preference and choice is not so tight. Throughout the rest of this article, I will use preference and choice interchangeably, while acknowledging that I take preference to be more basic and recognizing that the relationship between the two is not a settled question.

There are two ways to take the norm of normative decision theory: to guide to one's own actions or to assess from a third-person standpoint whether a decision-maker is doing what she

[12] However, there is an additional problem with the wide-scope norm for the non-constructivist realist: maximizing the expectation of some utility function doesn't guarantee that you've maximized the expectation of your own utility function. The connection between the utility function that is the output of a representation theorem and the decision-maker's actual utility function would need to be supplemented by some principle, such as a contingent version of Christensen's (2001) "Representational Accuracy" or by his "Informed Preference."

ought.[13] Having explained the norm of normative decision theory, I now turn to the question of what sort of "ought" it is supposed to correspond to.

3. Rationality

Decision theory is supposed to be a theory of rationality; but what concept of rationality does it analyze? Decision theory is sometimes said to be a theory of *instrumental* rationality – of taking the means to one's ends – and sometimes said to be a theory of *consistency*. But it is far from obvious that instrumental rationality and consistency are equivalent. So it is worth spending time on what each is supposed to mean and how EU theory is supposed to analyze each; and in what sense instrumental rationality and consistency come to the same thing.

Let us begin with instrumental rationality and with something else that is frequently said about decision theory: that it is "Humean." Hume distinguished sharply between reason and the passions and said that reason is concerned with abstract reasoning and with cause and effect, and while a belief can be contrary to reason, a passion (or in our terminology, a desire) is an "original existence" and cannot itself be irrational. As his famous dictum goes, "'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger."[14] Hume thinks that although we cannot pass judgment on the ends an individual adopts, we can pass judgment if she chooses means insufficient for her ends. To see how decision theory might be thought to provide this kind of assessment, consider the psychological realist version of the theory in which an individual's utility function corresponds to the strengths of her desires. This way of thinking about the theory gives rise to the natural suggestion that the utility function captures the strength of an agent's desires for various ends, and the dictum to maximize expected utility formalizes the dictum to prefer (or choose) the means to one's ends.

---

[13] Bermúdez (2009) distinguishes these as two separate uses: what I call using normative decision theory to guide one's own actions he calls the "action-guiding" use, and what I call using normative decision theory for third-person assessment he calls the "normative" use; however, he includes more in the normative use of decision theory than just assessing whether the agent has preferences that conform to the norm of EU theory, such as assessing how well she set up the decision problem and her substantive judgments of desirability.

[14] Hume (1731: 416).

The equivalence of preferring the means to one's ends and maximizing expected utility is not purely definitional. True, to prefer the means to one's ends is to prefer the act with the highest *utility*: to prefer the act that leads to the outcome one desires most strongly. However, in the situations we are concerned with, it is not clear which act will lead to which outcome – one only knows that an act will lead to a particular outcome if a particular state obtains – so one cannot simply pick the act that will lead to the more preferred outcome. Therefore, there is a real philosophical question about what preferring the means to your ends requires in these situations. EU theory answers this substantive question by claiming that you ought to maximize the *expectation* of the utility function relative to your *subjective* probability function. So if we cash out EU theory in the means-ends idiom, it requires you not precisely to prefer the means to your ends, but to prefer the means that will, *on average and by your own lights*, lead to your ends. It also requires that you *have* a consistent subjective probability function and that the structure of desires is such that a number can be assigned to each outcome. So it makes demands on three kinds of entities: beliefs, desires, and preferences given these. This formulation of the maximization norm is compatible with both the narrow-scope and the wide-scope reading: if in concert with Hume's position we think that desires cannot be changed by reason, there will be only one way to fulfill this requirement; but if we think that the agent might decide to alter her desires, there will be multiple ways to fulfill this requirement.

A more modern formulation of the idea that decision theory precisifies what it is to take the means to one's ends is that decision theory is *consequentialist*. This is to say that it is a principle of decision theory that acts must be valued only by their consequences. An important justification of the norm of EU theory as the unique consequentialist norm, and a justification that formalists and psychological realists can both avail themselves of, comes from Hammond (1988). Hammond considers sequential decision problems (decision problems in "extensive" form rather than "normal" form), where decision-makers are not choosing only once but instead can revise their plan of action as new information comes in. He argues that the assumption that decision-makers value acts only for their consequences, when cashed out in terms of some seemingly plausible principles about sequential choice, entails the substantive axioms of EU theory.[15]

[15] For further discussion of this type of argument, see Seidenfeld (1988), McClennen (1990), and Levi (1991).

Even in the case of choice at a time, I think we can think of the axioms as trying to formalize necessary conditions to preferring the means to one's ends. I don't have space to pursue the suggestion here, but here is one example of what I have in mind. Consider the requirement of state-wise dominance, which says roughly that if act *f* is weakly preferred to act *g* in every state, and strictly preferred in some state that has positive probability, then you ought to strictly prefer *f* to *g* (this is a necessary condition of being representable as an EU maximizer). One plausible way to state what's wrong with someone whose preferences don't conform to this requirement is that they fail to prefer what they believe is superior in terms of satisfying their preferences, or they fail to connect their preferences about means to their preferences about ends. Not all of the axioms can be straightforwardly argued for in this way, but this can be a helpful way to think about the relationship of the axioms to instrumental rationality.

Thus, normative EU theory may be supported by arguments to the effect that the maximization norm or the axiom norm spell out instrumental rationality (leaving aside whether these arguments are ultimately successful). The other notion of rationality that decision theory is often described as analyzing is *consistency*, and it seems that the axiom formulation of the norm coheres well with this. To understand why, it is helpful to consider the related idea that logic analyzes what it is to have consistent binary beliefs. There are two important standards at work in binary belief. First, an agent ought (roughly) to believe what is reasonable to believe, given her evidence. This is a requirement about the substance of her beliefs, or about the content of her beliefs vis-à-vis her evidence or what the world is like. Second, an agent's beliefs ought to be consistent with one another in the sense elucidated by logic. This is a requirement about the structure of her beliefs, or about the content of her beliefs vis-à-vis the content of her other beliefs. This isn't to say that everyone holds that agents must be logically perfect or omniscient, or that everyone holds that there is an external standard of adherence to the evidence, but the point is that these are two different kinds of norms and we can separately ask the questions of whether a believer conforms to each.

Similarly, in evaluating preferences over acts, there are two questions we might ask: whether an agent's preferences are reasonable, and whether they are consistent. Here, the axioms of decision theory are supposed to play a parallel role that the axioms of logic play in beliefs: without regard to the content of an agent's preferences, we can tell whether they obey the axioms. So just as the axioms of logic are supposed to spell out what it is to have consistent

binary beliefs, so too are the axioms of decision theory supposed to spell out what it is to have consistent preferences.

There are several ways in which it might be argued that the axioms correctly spell out what it is to have consistent preferences.[16] One classic argument purports to show that violating one or more of them implies that you will be the subject of a "money pump," a situation in which you will find a series or set of trades favorable but will disprefer the entire package, usually because taking all of them results in sure monetary loss for you.[17] This amounts to valuing the same thing differently under different descriptions – as individual trades on the one hand and as a package on the other – and is thought to be an internal defect rather than a practical liability.[18] A different argument, due to Peter Wakker (1988), purports to show that violating one of the axioms will entail that you will avoid certain cost-free information.

What I want to propose is that consistency of preferences is an amalgamation of consistency in three different kinds of entities: consistency in preferences over outcomes, consistency in preferences about which event to bet on, and consistency in the relationship between these two kinds of preferences and preferences over acts.[19] Or, psychological realists might say: consistency in desires, consistency in beliefs, and consistency in connecting these two things to preferences. Aside from the fact that adhering to the axioms does produce three separate functions (a utility function of outcomes, a probability function of states, and an expectational utility function of acts), which is not decisive, I offer two considerations in favor of this proposal. First, arguments for each of the axioms can focus more or less on each of these kinds of consistency. For example, an argument that transitivity is a rational requirement doesn't need to say anything about beliefs or probability functions. Second, a weaker set of axioms than

---

[16] Not all philosophers think that *arguing* for this conclusion is the right way to proceed. For example, Patrick Maher (1993: 62, 83) suggests that no knock-down intuitive argument can be given in favor of EU theory, but that we can justify it by the fruits it produces.

[17] Original versions of this argument are due to Ramsey (1926) and de Finetti (1937).

[18] See Christensen (1991), although he is mostly concerned with this type of argument as it relates to the subjective probability function.

[19] By a preference to bet on E rather than F, I mean a preference to receive a favored outcome on E rather than to receive that outcome on F.

those of EU theory will produce a consistent probability function without a utility function relative to which the agent maximizes EU; and a weaker set of axioms than those of EU theory will produce a utility function of outcomes without a probability function relative to which the agent maximizes EU; and a weaker set of axioms than those of EU theory will produce a utility function and a probability function relative to which an agent maximizes something other than EU.[20] Therefore, even if the justifications of each of the axioms are not separable into those based on each of the three kinds of consistency, the kinds of consistency are formally separable. And here is a difference, then, between logic and decision theory: logical consistency is an irreducible notion, whereas decision-theoretic consistency is a matter of being consistent in three different ways.[21]

Here, then, are the ways in which instrumental rationality and consistency are related. First, and most obviously, there are arguments that each is analyzed by EU theory; if these arguments are correct, then instrumental rationality and consistency come to the same thing. Second, given that consistency appears to involve consistency in the three kinds of entities instrumental rationality is concerned with, consistency in preferences can be seen as an internal

[20] For an axiomatization of a theory that yields a probability function for a certain kind of non-EU maximizer, see Machina and Schmeidler (1992). For an axiomatization of a theory that yields a utility function for an agent who lacks a subjective (additive) probability function, see Gilboa (1987), or any non-expected utility theory that uses subjective decision weights that do not necessarily constitute a probability function. For an axiomatization of a theory that yields a utility and probability function relative to which an agent maximizes a different functional, see Buchak (forthcoming).

[21] Note, however, that for non-constructive realists, there could be a case in which two of these things are inconsistent in the right way but preferences are still consistent. See Zynda (2000: 51-60), who provides an example of an agent whose beliefs are not governed by the probability calculus and whose norm is not expected utility maximization relative to his beliefs, but who has the same preferences as someone who maximizes EU relative to a probability function.

check on whether one really prefers the means to one's ends relative to a set of consistent beliefs and desires.[22]

It was noted that even if binary beliefs are consistent, we might ask the further question of whether they are reasonable. Can a similar question be applied to preferences? Given that consistency applies to three entities, the question of reasonableness can also be separated into three questions: whether the subjective probability function is reasonable, whether the utility function is reasonable, and whether one's norm is reasonable. The reasonableness question for subjective probability is an analogue of that for binary beliefs: are you in fact apportioning your beliefs to the evidence? For the formalist, the reasonableness question for utility, if it makes sense at all, will really be about preferences. But for the psychological realist, the reasonableness question for utility might be asked in different ways: whether the strength of your desires in fact tracks what would satisfy you, or whether they in fact track the good. In EU theory, there is only one norm consistent with taking the means to your ends – maximize expected utility – so the reasonableness question appears irrelevant; however, with the introduction of alternatives to EU theory, we might pose the question, and I will discuss this in section eight.

4. Interpretive Decision Theory

The major historical developments in normative decision theory mostly came from considering the question of what we ought to do. By contrast, another strand of decision theory was moved forward by philosophical questions about mental states and their relationship to action.

---

[22] Compare to Niko Kolodny's proposal that wide-scope requirements of formal coherence as such may be reducible to narrow-scope requirements of reason. The "error theory" in Kolodny (2007) proposes that inconsistency in beliefs reveals that one is not adopting, on some proposition, the belief that reason requires; and the error theory in Kolodny (2008) proposes that inconsistency in intentions reveals that one is not adopting the intention that reason requires. Direct application of Kolodny's proposal to the discussion here is complicated by the fact that some might see the maximization norm as wide-scope and some as narrow-scope. But those who see it as narrow-scope may take a Kolodny-inspired line and think that consistency of preferences is merely an epiphenomenon of preferring that which you have reason to prefer, given your beliefs and desires.

In 1926, Frank Ramsey was interested in a precise way to measure degrees of belief, since the prevailing view was that degrees of belief weren't appropriate candidates to use in a philosophical theory unless there was a way to measure them in terms of behavior. Ramsey noted that since degrees of belief are the basis of action, we can measure the degree of a belief by the extent to which the individual would act on the belief in hypothetical circumstances. Ramsey created a method whereby a subject's preferences in hypothetical choice situations are elicited and her degrees of belief (subjective probabilities) are inferred through these, without knowing her values ahead of time. For example, suppose a subject prefers getting a certain prize to not getting that prize, and suppose she is neutral about seeing the heads side of a coin or the tails side of a coin. Then if she is indifferent between the gamble on which she receives the prize if the coin lands heads and the gamble on which she receives the prize if the coin lands tails, it can be inferred that she believes to equal degree that the coin will land heads as that it will land tails, i.e., she believes each to degree 0.5. If she prefers getting the prize on the heads side, it can be inferred that she assigns a greater degree of belief to heads than to tails.

Generalizing the insight that both beliefs and values can be elicited through preferences, Ramsey presented a representation theorem. Ramsey's theorem was a precursor to Savage's, and like Savage's theorem, Ramsey's connects preferences to a probability function and a value function, both subjective. Thus, like the normative decision theorists that came after him, Ramsey saw that maximizing expected utility with respect to one's personal probability and utility functions is equivalent to having preferences that conform to certain structural requirements. However, Ramsey was not interested in using the equivalence to reformulate the maximization norm as a norm about preferences. Rather, he *assumed* that preferences *do* conform to the axioms, and used the equivalence to discover facts about the agent's beliefs and desires.

Related to Ramsey's question of how to measure beliefs is the more general question of attributing mental states to individuals on the basis of their actions. Donald Davidson (1973) coined the term "radical interpretation" (a play on W.V.O. Quine's "radical translation") to refer to the process of interpreting a speaker's beliefs, desires, and meanings from her behavior. For Davidson, this process is constrained by certain rules, among them a principle about the relationship between beliefs and desires on the one hand and actions on the other, which, as David Lewis (1974) made explicit, can be formalized using expected utility theory. Lewis's

formulation of the "Rationalization Principle" is precisely that rational agents act so as to maximize their expectation given their beliefs and desires. Thus, Ramsey's insight became a part of a more general theory about interpreting others. For theorists who make use of EU theory to interpret agents, maximizing EU is constitutive of (rational) action; indeed, Lewis (1974: 335) claims that the Rationalization Principle has a status akin to analyticity.

An immediate tension arises between the following three facts. First, for interpretive theorists, anyone who cannot be interpreted via the Rationalization Principle will count as unintelligible. Second, actual human beings are supposed to be intelligible; after all, the point of the project is to formalize how we make sense of another person. Third, actual human beings appear to violate EU theory; otherwise, the normative theory wouldn't identify an interesting norm.

One line to take here is to retain the assumption that it is analytic that agents maximize EU, and to explain away the apparent violations. We will see a strategy for doing this in the next section, but I will argue there that adopting this strategy in such a way as to imply that EU maximization cannot be violated leads to uninformative 'interpretations.' A more promising line starts from the observation that when we try to make sense of another person's preferences, we are trying to make sense of them as a whole, not of each considered in isolation. Consider an agent whose preferences mostly conform to the theory but fail to in a few particular instances, for example, an individual who generally gets up at 7 AM to go for a run but occasionally oversleeps her alarm. We would say that she prefers to exercise in the morning. Or consider an individual who generally brings an umbrella when the chance of rain is reported as at least 50%, but one time leaves it at home when she thinks it is almost certain to rain. We would say that she considers the burden of carrying around an umbrella only moderate in comparison to how much she does not like to get wet. In general, if a large set of an individual's preferences cohere, the natural thing to say is that she has the beliefs and desires expressed by those preferences but that her preferences occasionally fail to match up with her beliefs and desires, perhaps because she is on occasion careless or confused or weak of will.

This suggests what interpretive theorists ought to do in the case of non-ideal agents: take an agent's actual preferences, consider the closest "ideal" set of preferences – the closest set of preferences that do conform to the axioms – and infer the agent's beliefs and desires from these. Thus the theorist will interpret the agent as being as close to ideally rational as possible: we

might say, as maximizing expected utility in general, but as occasionally failing to do so. Furthermore, this allows us to interpret the agent as failing to maximize the expectation of her utility function on occasion – that is, as *having* desires on this occasion but failing to prefer in accordance with them – precisely because her obeying the axioms in a large set of her preferences or having a closest ideal counterpart points to a utility function that is genuinely hers. I note that "closest" here might be cashed out either as differing least from the agent's actual preferences, or as preserving the values that the agent would endorse in a clear-headed frame of mind, or as best according with other facts about her psychology, such as her utterances. I also note that in some cases there will be no close counterpart, and it will be precisely these cases in which the interpretive theorist will count the agent as unintelligible, as not intentionally acting.

There is one problem with this method, however. It does not allow us to interpret an agent as having genuinely inconsistent beliefs or desires, only as failing to have preferences that accord with them on occasion. While I don't have space to fully explore the possibilities here, there seem to me to be several options. First, and perhaps less plausibly, an interpretive theorist might postulate that an individual "really" has a coherent set of beliefs and desires, though these aren't always correctly translated into preferences. Second, one might postulate that an agent's degree of belief in a proposition is derived from (the closest ideal set to) some privileged set of preferences; for example, as many propose, that $p(E)$ is derived from the bets in small amounts of money one is willing to make on E. And similarly, perhaps, for desire, although it is harder to say what the privileged set might be. Finally, if some of one's preferences cluster towards one ideal counterpart and some towards another, along a natural division, and we could postulate that the agent is of two minds in a very particular way.

Decision theory appears in philosophy in two different strands. The normative theorist is interested in what your preferences ought to be given your beliefs and desires or given other of your preferences. Adopting EU maximization or conformity to the axioms as the correct norm, she says that you ought to prefer that which maximizes expected utility, and she is interested in which acts would do so; or she says that you ought to have consistent preferences, and is interested in which sets of preferences are consistent. The interpretive theorist is interested in discovering what your beliefs and desires are from your preferences. Adopting EU maximization or conformity to the axioms as the correct principle of interpretation, she says that

you do (approximately) maximize expected utility or have consistent preferences, and she is interested in what beliefs and desires make it the case that you do so.

When thus described, we can see that rationality plays a different role in each use of the theory: the interpretive theorist takes it as an *assumption* that individuals are rational in the decision-theoretic sense, and the normative theorist takes decision theory as a way to *answer* the question of whether individuals are rational. Although on the face of it this makes it seem that the two uses are in tension, I have proposed that on the best way to make sense of the interpretive project, the concept of rationality that is meant to be analyzed by EU theory is importantly different in the two projects. Specifically, the rationality norm of the normative project is "strong" in that normative theorists are interested in whether *all* of the individual's preferences adhere to it, and the rationality assumption in the interpretive project is "weak" in that interpretive theorists make the assumption that an agent more-or-less follows it but not the stronger assumption that she follows it exactly and always.

5. Outcome Descriptions

A lot has been made so far of the fact that by connecting preferences to subjective utility and probability functions, we can discover how much an agent values outcomes and how likely she takes various states to be. But one issue that has not yet been remarked upon is that just as how the agent values the outcomes and views the world are not intrinsic features of any situation she faces, neither is how the agent conceptualizes the outcomes.

An illustrative example is due to John Broome (1991: 100-101). Maurice is offered choices between various activities. If the choice is between going sightseeing in Rome and going mountaineering, Maurice prefers to go sightseeing, because mountaineering frightens him. If the choice is between staying home and going sightseeing, he prefers to stay home, because Rome bores him. However, if the choice is between mountaineering and staying home, he prefers to go mountaineering, because he doesn't want to be cowardly.

If we consider Maurice's preferences among Rome, home, and mountaineering, they appear to be intransitive: he prefers Rome to mountaineering, home to Rome, and mountaineering to home. Given that transitivity is necessary for EU maximization, the interpretive theorist is unable to make sense of him given the preferences as stated; but his motivation is perfectly comprehensible (we've just described it). In addition, the normative

theorist must automatically count Maurice's preferences as irrational, without considering whether his reasons for them make sense; but it is not clear – at least not without further argument – that there really is anything wrong with his preferences.

Here is what each theorist ought to be able to say: for Maurice, choosing mountaineering when the alternative is going to Rome is different from choosing mountaineering when the alternative is staying home. Therefore, there are really (at least) four options involved in this decision problem: Rome, home-when-the-alternative-is-Rome, home-when-the-alternative-is-mountaineering, and mountaineering. And Maurice's preferences among these options are not, so far as we know, intransitive. The lesson is that insofar as we are concerned with capturing the agent's actual beliefs and desires, we cannot assume that there is a privileged description of outcomes independent of the agent himself. Furthermore, insofar as we are interested in determining whether an agent is genuinely consistent or genuinely prefers the means to his ends, we cannot rule out his caring about certain features of outcomes out of hand. What the agent believes and desires is what we are trying to determine, and that includes what the agent believes about the choices he faces.

Thus, there is an additional "moving piece" in the interpretation of an agent or in a judgment about whether his preferences are rational: how he sees the outcomes. This poses two related challenges. The first is about how to settle on the correct interpretation of the agent's preferences. The second has to do with the extent to which individuating outcomes more finely commits the agent to having preferences in choice situations that could never even in principle be realized, and how we ought to treat these preferences. I will discuss these issues in reverse order.

To illustrate the second problem, notice that it is assumed in decision theory that preferences are complete: for any two options, a decision maker must prefer one to the other or be indifferent. This means that if "home-when-the-alternative-is-Rome" and "home-when-the-alternative-is-mountaineering" are to count as options in some of the choice problems described above, the decision-maker must prefer one to the other or be indifferent. But one could never actually face a choice between these two options, by definition. Broome (1991) refers to preferences like these as "non-practical preferences." I will not discuss the metaphysics of these preferences: although there are interesting questions here, they do not obviously bear on the issues this article has been focusing on. But the epistemology of these preferences is important,

because it will make a difference to how we resolve the interpretive problem more generally. There will be a divide among those who think that which of these options an agent prefers is up to the agent, and those who think that which of these options an agent prefers is up to the decision theorist to fill in; and where one falls on this divide will determine how much freedom a decision theorist has to interpret an agent's preferences.

The other problem, then, is how to settle on an interpretation of the agent's preferences. As we've just seen, we cannot allow that the theorist's initial presentation of the outcomes is how the agent sees them. However, if we allow that the agent makes maximally fine distinctions between outcomes, then none of the outcomes will be the subject of more than one practical preference. For example, choosing each outcome in a pair-wise choice always involves rejecting the other alternative. If the agent's non-practical preferences are up to the theorist to fill in, then this will mean that the agent can never fail to maximize expected utility or fail to satisfy the axioms, since no practical preferences will be inconsistent with each other.

If the norm of EU theory were impossible to violate, the normative theory would lose its bite, since it will be trivially true that every agent adheres to the norm. But would this also be a problem for the interpretive EU theorist? Some might say that it wouldn't be; indeed, that EU maximization is trivially satisfied would lend support to the idea that it is a good interpretive assumption that agents *actually* maximize EU. But there are at least two problems with this approach for the interpretive theorist. The first is that we will be unable to tell the difference between when an individual is trying to maximize EU (or follow the axioms) but making a mistake and when she is aiming at something else,[23] although perhaps this is okay if it is argued that to act at all is to maximize EU. The second problem is that allowing outcomes to be individuated maximally finely means that 'deriving' an agent's beliefs and desires from her preferences won't be very informative. Her practical preferences in combination with each possible filling out of her non-practical preferences will give rise to a unique (up to positive affine transformation) utility and probability function, by the representation theorems. But there may be many possible fillings out. Therefore, there will be multiple and incompatible ways to interpret her beliefs and desires. And on the level of preferences, knowing what she prefers in one particular context won't tell us anything about what she prefers in an only slightly different

[23] See Hurley (1989: 55-83)

context, so we won't get a very robust explanation of her psychology. In either case, the theory is rendered uninformative: we cannot make much sense of what the agent is doing.

Most philosophers accept that either the theorist's ability to individuate outcomes or the theorist's ability to set non-practical preferences must be constrained. To constrain them, one can either introduce a rule about when two outcomes are allowed to count as different, or allow that outcomes can be individuated as finely as possible but introduce a rule about what non-practical preferences the theorist can interpret the agent as having. For most purposes, these come to the same thing, since refusing to allow that *x* and *y* are different outcomes and requiring that the correct interpretation of the agent makes her indifferent between *x* and *y* permit the same sets of practical preferences. But there are two very different types of constraining rules that the theorist could introduce (this distinction crosscuts the distinction just mentioned). To see this, consider the following suggested rules:

> R1: Outcomes should be distinguished as different if and only if they differ in a way that makes it rational to have a preference between them. (Broome 1991: 103).
>
> R2: Outcomes should be distinguished as different if and only if the agent actually has a preference between them. (Dreier 1996: 260).
>
> R3: Outcomes should be distinguished as different if and only if they differ in regard to properties that are desired or undesired by the agent. (Pettit 2002: 212)

Maurice's preferences can be accommodated by EU theory according to rule R1 only if it is rational for Maurice to care about what option he turns down when he decides to stay at home, according to rule R2 only if he in fact does care about what option he turns down when he decides to stay at home, and according to rule R3 only if turning down an option instantiates a property he in fact cares about.

Rules R2 and R3 make the possibility of distinguishing outcomes dependent on the agent's internal state, whereas R1 makes this possibility dependent on some objective feature of the agent's situation. Rules like R1 that introduce an external constraint on interpretation might be seen as principles of charity for interpretation: we should interpret an agent as making a distinction only if it is rational to make that distinction. Since these "externalist" rules restrict preferences beyond what the agent herself values, Broome has rightly pointed out that they are against the spirit of Humeanism (Broome 1993). Of course, rules like R2 and R3 can only be applied if the theorist has access to the agent's non-practical preferences or other relevant

properties of her internal state. Therefore, using these "internalist" rules relies on the theorist knowing more about an agent's psychology than the externalist rules do.

The same strategy can be applied to ensuring that the norm of normative decision theory is not trivial. As long as there is a restriction on when two outcomes can count as different, there will be sets of preferences that violate the norm of EU theory. Which type of rule to adopt will depend on the use to which normative decision theory is being put: if the theorist is using it to assess an agent, whether the theorist can rely on an internalist rule will depend on how much she can know about the agent's internal state, and if the agent is using normative decision theory to guide her own actions, whether she can rely on an internalist rule will depend on how much introspective access she has to her own internal state.

6. Descriptive Decision Theory

Although a third type of decision theory, descriptive decision theory (which sometimes goes by the name "behavioral economics"), is largely the provenance of psychology and economics rather than philosophy, it is important to say something about it both for completeness and to make clear the contrast with interpretive decision theory.

Like interpretive decision theory, descriptive decision theory is interested in describing the behavior of individuals rather than in what they ought to do. However, there is an important difference between the two approaches, which can be seen in how they have responded to findings that actual agents fail in reliable ways to maximize expected utility. Whereas interpretive decision theory has retained EU maximization as the guiding principle of interpretation, and in many cases pushed for a more complex interpretation of outcomes (as described in the previous section), descriptive decision theory has by and large abandoned expected utility maximization as an unrealistic assumption of agents and proposed alternatives.

I do not have space to go into the alternatives to EU theory that descriptive theorists have proposed (see Sugden (2004) and Schmidt (2004) for helpful surveys), but it is worth saying two ways in which these alternatives tend to differ from EU theory. First, while they generally include a function that plays the role of utility and a function that plays the role of probability, they either subject these to different constraints (e.g. the 'probability' function needn't be additive) or else combine them in a non-expectational way. Second, at least one notable alternative, Kahneman and Tversky's (1979) *prospect theory*, posits an "editing phase" during

which the decision-maker simplifies the alternatives using various heuristics before subjecting them to the maximization schema.

The differing responses of the two types of theorists to purported violation reveals two important differences between the aims of descriptive decision theory and the aims of interpretive decision theory. First, descriptive theorists are generally interested in building parsimonious models of preferences, and they are less concerned than interpretive theorists with interpreting the utility and probability functions as desires and beliefs. Interpretive theorists, by contrast, are primarily interested in extracting desires and beliefs with few or no initial assumptions, including assumptions about how an agent views the outcomes; and in doing so need be only as parsimonious about outcomes as an agent's actual psychology is. Therefore, descriptive decision theorists are more inclined to treat the outcomes (for them, generally, monetary values) as theoretical bedrock, and interpretive decision theorists are more inclined to treat the rationalization principle as theoretical bedrock. It is worth noting that for the same reasons that economists concerned merely with modeling behavior will be uninterested in the interpretive project, formalists will also not be interested in the interpretive project, since for them, there aren't any interesting entities worth discovering.

The other difference, which I will discuss further in the next section, concerns predictable deviation from rationality. Roughly, if agents predictably have preferences against the dictates of rationality, the descriptive theorist will want to include this as part of her model, since it is an accurate characterization of what the agent does, but the interpretive theorist will not, since, recalling the discussion in section four, those preferences do not accurately reflect her beliefs and desires (though predictable deviations may be included somewhere in the theory of action). Interpretive theorists are interested in characterizing the preferences of an idealized version of the agent, and descriptive theorists in those of the actual, non-ideally-rational agent.

We might put these two points succinctly, although this is certainly too coarse: descriptive theorists are concerned with *prediction*, and interpretive theorists are concerned with *explanation* in terms of beliefs and desires and with discovering something about an agent's mental states.[24]

[24] I should note that while I separate explanation from prediction, Bermúdez (2009) thinks they ought to be considered a single dimension of decision theory, and thus that the same formal theory must play both roles.

How does the descriptive project bear on the interpretive project? If the analogues of *u* and *p* in the descriptive project should be taken in a formalist vein, then the descriptive project does not have a clear bearing on the interpretive project. But insofar as the entities involved in the descriptive project can be thought of as beliefs and desires, rather than convenient ways to represent preferences, I think there is a way in which the descriptive project aids the interpretive project, and in another way it cuts against it. On the one hand, the descriptive project can help illuminate the relationship between an agent's actual choices and desires and those of her ideal counterpart. For example, one of the findings of prospect theory (Kahneman and Tversky 1979: 273) is that when a new frame of reference is experimentally induced, e.g., the agent believes she will receive $2000, her preferences over total amounts of money are altered in the sense that receiving less than (e.g.) $2000 will be treated as a "loss." If what this shows is that inducing a reference point causes people to underestimate their actual (subjective) utility below the reference point, then we can expect that the ideal counterpart will assign higher utility below the reference point than the actual agent in the grip of framing effects. On the other hand, if what the descriptive project reveals is that an agent cannot be interpreted as having stable beliefs and desires – beliefs and desires that are independent of the ways in which choices are presented – then the descriptive project undermines the interpretive project.

7. The Mutual Dependence of the Normative and Interpretive Project

In this section, I will explain how the normative and interpretive project depend on each other. Recall that the rationality assumption in interpretive decision theory is that agents are approximately expected utility maximizers; and an agent's beliefs and desires are the *p* and *u* extracted from the preferences of her ideal counterpart. But why should we think that the beliefs and desires of an agent's ideal counterpart are *her* beliefs and desires? After all, the *preferences* of her ideal counterpart aren't her actual preferences. The crucial idea is that acting consists not in actually taking the means to your ends, but in *aiming at* doing so. Therefore, the preferences of an agent's ideal counterpart are the preferences that she ought to be thought of as aiming at satisfying when she acts. This doesn't mean that she consciously aims at satisfying these preferences, or even that she consciously takes herself to be maximizing expected utility; rather, she participates in an activity (acting) which is constituted by aiming at being an EU maximizer. In the means-ends idiom, to act is to aim to take the means to your ends (or more precisely the

means that will on average and by your own lights lead to your ends), even though you might sometimes fail to do so. That aiming is not the same as succeeding explains the fact that the rationality assumption in interpretive theory is not that agents are perfectly rational but rather that they are approximately rational.[25]

Now that it is clear what the rationality assumption in interpretive decision theory amounts to, it should also be clear how the interpretive project depends on the normative project. Interpretive EU theory rests on two claims. First, on the claim that action aims at conforming to the norm that analyzes what it is to take the means to ones ends or to be consistent. Second, on the claim that this norm is captured by EU theory, either in the maximization formulation or the axiom formulation. If we were to conclude that a different norm holds of rational preferences, then interpretive decision theory would have to follow suit in adopting that norm as the one action aims at. The interpretive project depends on the correctness of the normative theory's norm, i.e., on the normative theorist being correct about which sets of preferences are candidates for those of an agent's ideal counterpart. (Note that this is another difference between interpretive and descriptive decision theory: the latter is not at all governed by what the correct norm is.)

The normative project, if it is able to say anything interesting about agents who fall short of the norm, also depends on the interpretive project. This is because identifying how an agent is falling short of the norm depends on correctly interpreting what her beliefs and desires are. Clearly, a simple yes or no answer to the question of whether the agent is doing what she ought doesn't rely on discovering the agent's beliefs and desires: if her preferences don't conform to the axioms, then she fails to do what she ought. The formalist will say that normative decision theory ends here. However, there are two additional questions the psychological realist might be interested in. First, what is the source of the agent's irrationality? And second, where should the agent go from here?

---

[25] We should untangle the question of whether postulating that agents aim at EU maximization allows the preferences of an agent's ideal counterpart to reveal her beliefs and desires from the more general question of whether postulating that agents aim at whatever the correct norm of rationality is allows this. Meacham and Weisberg (2011) argue against the former claim on empirical grounds, but we might still uphold the latter claim by arguing that rationality is best analyzed by a different norm which people in fact come closer to adhering to.

Recall that preference inconsistency could come from one (or more) of three sources: inconsistency in beliefs, inconsistency in desires, or inconsistency in the norm connecting beliefs and desires. But we need to be able understand the agent as having beliefs and desires even when she is inconsistent if we want to claim that her beliefs or her desires are inconsistent. And so if we adopt the interpretive idea that an agent's beliefs and desires can be discovered even when she is not fully in accord with the axioms by working backwards from the preferences of her ideal counterpart(s), we can say in what respect she is falling short of the normative ideal. Unless one can introspect one's beliefs and desires to a precise degree, diagnosing where the irrationality comes from depends on the possibility of interpreting agents as having beliefs and desires even when they are not obeying the axioms.

Furthermore, since the interpretive use of the theory allows us to discover an agent's beliefs and desires, it allows us to say what sort of underlying change moving from the agent's actual preference to a set of preferences that conform to the theory involves. For example, consider an individual who is willing to add $200 to the purchase price of a car he is buying if it has a radio but would not pay $200 to have a radio installed if the car came without one at the cheaper price.[26] Let us assume the closest ideal agent prefers $200 to the radio, so that the actual agent desires $200 more than she desires the radio. The narrow-scope norm says she ought to alter the preference concerning the initial purchase. The wide-scope norm of decision theory is more permissive. It says that she can resolve the irrationality by adopting any consistent set of preferences: so she can alter the preference concerning the initial purchase and retain the rest of her preferences, or she can keep that preference and alter the rest of her preferences. But even if we adopt the wide-scope norm, interpreting the agent is crucial because it allows us to say what each resolution involves: the former resolution involves conforming her preferences over acts to her underlying desires; the latter involves bringing her underlying desires in line with the preference about purchasing a new car. This doesn't by itself show how she ought to resolve the decision, since in principle it may be that the one preference is more important than her desires, but it does tie different ways of resolving the decision to preserving specific different features of his situation.

[26] Example adapted from Savage (1954: 103).

In sum, if the normative theorist wants to say more than that the agent is not doing what she ought, or that she ought to bring her preferences in line with the axioms somehow or other but with no guidance on what considerations are involved in potential resolutions, she will have to interpret the agent.

The assumption of rationality in interpretive decision theory is that the agent aims at maximizing EU, and so approximates an EU maximizer. And the goal of rationality in normative decision theory is that the agent maximizes EU in every instance. This, then, is how the two projects are mutually dependent: that agents are approximately EU maximizers depends on EU maximization being the aim of rational action, and that agents bring their preferences into line with EU maximization in a way that is governed by reasons depends on locating the source of their current deviation, which depends on understanding what their beliefs and desires are.

Descriptive decision theory also bears on these projects. As I alluded to in section six, one thing that descriptive decision theory could reveal is that it would be seriously misguided to think of action as aiming at maximizing expected utility. This would undermine interpretive EU theory. But what would it say about rational action more generally? As mentioned, interpretive decision theory makes two assumptions, one that action aims at adhering to the norm of rationality and one about what the norm is. If action doesn't aim at the maximization of EU, then we must drop either the assumption that action aims at the norm or the assumption that EU is the correct norm. If we keep the latter assumption, then it may be possible to use a descriptive theory to extract beliefs and desires and a normative theory that takes these as the real beliefs and desires and enjoins you to maximize expected utility.[27] On the other hand, we might keep the former assumption and propose a different norm, one that coheres closely enough with actual behavior that interpretive decision theory can use the norm to backwards engineer beliefs and desires despite some deviations that the correct descriptive theory predicts. Which of these positions to take will not be determined by empirical findings but by arguments about what the correct norm is, although the knowledge that humans diverge wildly from EU theory might give

[27] Bermúdez (2009: 165-167) considers but rejects a possibility like this in his discussion of whether decision theory could play multiple roles at once.

us reason to examine more closely whether EU is the correct norm, given how successful human behavior is in general.[28]

8. Challenges and Extensions

In fact, philosophers have challenged the idea that EU theory is the correct theory of rationality. Recall that in the expected utility equation, the utility value of each outcome is weighted by the probability the decision-maker assigns to the state in which she receives that outcome, and this probability is supposed to reflect her belief about that state. This assumes two things: first, that the norm relating beliefs and desires to preferences is indeed that of EU maximization, in which the average value of a gamble suffices to determine its position in the agent's preference ranking (whether this is derived from the axioms or not); second, that rational beliefs are "sharp" in that they can be measured by point-probabilities. Challenges to each of these points have been around for at least 50 years, but they have resurfaced recently. One might similarly challenge that the structure of desire is that posited by EU theory, though I don't have space to discuss such a challenge here. Each of these challenges can be posed directly about the functions *p*, *u*, or the maximization norm, but each can also take the form of criticizing one or more of the axioms, so the challenges do not rest on a particular interpretation of the utility function.

The first challenge is to the idea that we ought to care only about the expectation of utility, and not other "global" features of a gamble, such as its minimum utility value, its maximum utility value, or the spread or variance of utility. Again, since utility is derived or discovered via a representation theorem, this point must take the form of or be accompanied by a challenge to one or more axioms of EU theory. Maurice Allais (1953), taking what appears to be a non-constructivist realist view of the utility function, argued that agents might care not just about the mean utility value of a gamble, but also about its variance and skewness. But his famous counterexample to EU theory (which has since become known as the Allais Paradox) poses a challenge even for constructivists, since it shows that most decision-makers violate one of the axioms of EU theory.[29] I have recently defended axioms that give rise to other

[28] For argument that humans did not evolve to be EU maximizers, see Okasha (2007).

[29] At least under the assumption that outcomes cannot be individuated more finely.

maximization norms than that of EU theory (Buchak, forthcoming): my view is that EU maximization is one of a more general class of norms any of which an agent may adopt. In the same spirit as the idea that agents have subjective *u* and *p* function, I propose that the norm that connects these to preferences is also up to the agent. Just as *u* and *p* are subject to structural constraints, so too is the norm; and, furthermore, the question mentioned in section three about whether a particular norm (such as EU maximization) is reasonable in addition to rational can be posed.

The second challenge is to the idea that we ought to be "probabilistically sophisticated" in the sense of assigning to every event a point probability (or acting as if we do). Daniel Ellsberg (1961) proposed a pair of choice problems that have become known as the Ellsberg Paradox, purporting to show that when individuals lack precise information about objective probabilities, they don't act as if they make choices based on a single probability function. In recent years, the challenge has come from the side of epistemology rather than observed decision-making behavior, the idea being that our evidence is often imprecise or incomplete, so requiring precise degrees of belief would mean requiring degrees of belief that outrun the evidence.[30] Denying that we have sharp degrees of belief and that we need them in order to make rational decisions requires stating both what non-sharp (or "imprecise") degrees of belief are and how to make decisions with them.[31]

9. Conclusion: Decision Theor*ies*

Given both the historical progression and the issues that are currently under discussion, we ought not think of decision theory as a single theory, but rather as a collection of theories that each contains both a structural and interpretive element. The structural element describes both the internal structure of several functions and the formal relationship between these functions on the one hand and preferences on the other; a formal relationship which holds just in case a particular set of axioms is satisfied. This internal structure and relationship are argued to be those that hold for rational agents. In EU theory, the posited functions are a numerically valued utility function and a point-probability function that obeys the probability calculus; and the

[30] See the discussion found in White (2009), Elga (2010), Joyce (2010).

[31] For some examples, see Levi (1974), Sahlin and Gärdenfors (1982), and Joyce (2010).

posited relationship is that of EU maximization. The interpretive element concerns how these functions are to be interpreted: as psychologically real and in principle separate from preferences; as psychologically real and tightly connected to preferences; or merely as a representation of preferences. Whichever combination of structural element and interpretation we adopt, the underlying issues discussed in the previous few sections – the relationship between decision theory and rationality, how to individuate outcomes, and the relationship between normative decision theory and interpretive decision theory – remain the same.

BIBLIOGRAPHY

Allais, M. (1953), 'The Foundations of a Positive Theory of Choice involving Risk and a Criticism of the Postulates and Axioms of the American School', in M. Allais and O. Hagen (eds.), *Expected Utility Hypothesis and the Allais Paradox* (D. Reidel Publishing Company, 1979), 27-145.

Armendt, B. (1986), 'A Foundation for Causal Decision Theory', in *Topoi* 5/1: 3-19.

Bartha, P. (2007), 'Taking Stock of Infinite Value: Pascal's Wager and Relative Utilities', in *Synthese* 154/1: 5-52.

Bermúdez, J. (2009), *Decision Theory and Rationality* (Oxford University Press).

Bernoulli, D. (1738), 'Exposition of a New Theory on the Measurement of Risk', in *Econometrica* 22/1 (1954): 23-36.

Bolker, E. (1965), *Functions Resembling Quotients of Measures* (doctoral dissertation for Harvard University).

Bolker, E. (1966), 'Functions Resembling Quotients of Measures', in *Transactions of the American Mathematical Society* 124: 292-312.

Bolker, E. (1967), 'A Simultaneous Axiomatization of Utility and Subjective Probability ' in *Philosophy of Science* 34ethan b: 333-340.

Broome, J. (1991), *Weighing Goods: Equality, Uncertainty, and Time* (Blackwell Publishers Ltd.).

Broome, J. (1993), 'Can a Humean be moderate?', in G. Frey and C. Morris (eds.), *Value, Welfare and Morality* (Cambridge University Press), 279-86.

Buchak, L. (forthcoming), *Risk and Rationality* (Oxford University Press).

Christensen, D. (1991), 'Clever Bookies and Coherent Belief', in *Philosophical Review* 100/2: 229-47.

Christensen, D. (2001), 'Preference-Based Arguments for Probabilism', in *Philosophy of Science* 68/3: 356-376

Colyvan, M. (2008), 'Relative Expectation Theory', in *Journal of Philosophy* 105/1: 37-44.

Davidson, D. (1973), 'Radical Interpretation', in *Dialectica* 27/3-4: 313-328.

Dreier, J. (1996), 'Rational Preference: Decision Theory as a Theory of Practical Rationality', in *Theory and Decision* 40: 249-276.

Easwaran, K. (2008), 'Strong and Weak Expectations', in *Mind* 117: 633-641.

Elga, A. (2010), 'Subjective Probabilities Should be Sharp', in *Philosophers' Imprint* 10/5.

Ellsberg, D. (1961), 'Risk, Ambiguity, and the Savage Axioms', in *Quarterly Journal of Economics* 75/4: 643-669.

Fermat, P., and B. Pascal (1654), letters, collected as 'Fermat and Pascal on Probability' and translated from the French by Professor Vera Sanford, in D. Smith (ed.), *A Source Book in Mathematics* (McGraw-Hill Book Co., 1929): 546-565.

de Finetti, B. (1937), 'Foresight: Its Logical Laws, Its Subjective Sources', in H. Kyburg and H. Smokler (eds.), *Studies in Subjective Probability* (Robert E. Kreiger Publishing Co, 1980).

Fishburn, P. (1981), 'Subjective Expected Utility: A Review of Normative Theories', in *Theory and Decision* 13: 139-199.

Gärdenfors, P., and Sahlin, N. (1982), 'Unreliable Probabilities, Risk Taking, and Decision Making', in *Synthese* 53: 361-386.

Gibbard, A., and Harper, W. (1978), 'Counterfactuals and Two Kinds of Expected Utility', in W. Harper, R. Stalnaker, and G. Pearce (eds.), *Ifs: Conditionals, Belief, Decision, Chance, and Time* (Reidel, 1981), 153-190.

Gilboa, I. (1987), 'Expected Utility with Purely Subjective Non-Additive Probabilities', in *Journal of Mathematical Economics* 16: 65-88.

Hammond, P. (1988), 'Consequentialist Foundations for Expected Utility', in *Theory and Decision* 25: 25-78.

Hansson, B. (1988), 'Risk Aversion as a Problem of Conjoint Measurement', in P. Gärdenfors and N. Sahlin (eds.), *Decision, Probability, and Utility* (Cambridge University Press), 136-158.

Hume, D. (1731), *A Treatise of Human Nature* (Oxford University Press, P. Nidditch (ed.), 1978).

Hurley, S. (1989), *Natural Reasons, Personality and Polity* (Oxford University Press).

Jeffrey, R. (1965), *The Logic of Decision* (McGraw-Hill Inc.)

Jeffrey, R. (1990), *The Logic of Decision*, second edition (University of Chicago Press).

Joyce, J. (1999), *The Foundations of Causal Decision Theory* (Cambridge University Press).

Joyce, J. (2010), 'A Defense of Imprecise Credences in Inference and Decision Making', in *Philosophical Perspectives* 24/1: 281-323.

Kahneman, D., and Tversky, A. (1979), 'Prospect Theory: An Analysis of Decision under Risk', in *Econometrica* 47: 263-291.

Kolodny, N. (2007), 'How Does Coherence Matter?', in *Proceedings of the Aristotelian Society* 107/3.

Kolodny, N. (2008), 'The Myth of Practical Consistency', in *European Journal of Philosophy* 16/3: 366-402.

Lewis, D. (1974), 'Radical Interpretation', in *Synthese* 23: 331-344.

Levi, I. (1974), 'On Indeterminate Probabilities', in *Journal of Philosophy* 71: 391-418.

Lewis, D. (1981), 'Causal Decision Theory,' in *Australasian Journal of Philosophy*, 59: 5–30.

Levi, I. (1991), 'Consequentialism and Sequential Choice', in M. Bacharach and S. Hurley (eds.), *Foundations of Decision Theory* (Basil Blackwell Ltd), 92-122.

Machina, M., and Schmeidler, D. (1992), 'A More Robust Definition of Subjective Probability', in *Econometrica* 60/4: 745-780.

Maher, P. (1993), *Betting on Theories* (Cambridge University Press).

McClennen, E. (1990), *Rationality and Dynamic Choice: Foundational Explorations* (Cambridge University Press).

von Neumann, J., and Morgenstern, O. (1944), *Theory of Games and Economic Behavior* (Princeton University Press).

Nover, H., and Hájek, A. (2004), 'Vexing Expectations', in *Mind* 113: 237-249.

Nozick, Robert (1969), 'Newcomb's Problem and Two principles of Choice,' in Nicholas Rescher (ed.), *Essays in Honor of Carl G. Hempel*, Synthese Library (Reidel), 114-115.

Okasha, S. (2007), 'Rational Choice, Risk Aversion and Evolution', in *Journal of Philosophy* 104/5: 217-235.

Pettit, P. (1991), 'Decision Theory and Folk Psychology', in M. Bacharach and S. Hurley (eds.), *Foundations of Decision Theory* (Basil Blackwell Ltd), 147-175.

Ramsey, F. (1926/1931), 'Truth and Probability,' in *The Foundations of Mathematics and other Logical Essays* (Kegan, Paul, Trench, Trubner & Co., R. Braithwaite (ed.)).

Samuelson, P. (1938), 'A Note on the Pure Theory of Consumer's Behaviour', in *Economica* 5/17: 61-71.

Savage, L. (1954/1974), *The Foundations of Statistics* (John Wiley and Sons, Inc. / Dover).

Schmidt, U. (2004), 'Alternatives to Expected Utility: Formal Theories', in S. Barberà, P. Hammond, and C. Seidl (eds.), *Handbook of Utility Theory* (Kluwer Academic Publishers), 757-837.

Seidenfeld, T. (1988), 'Decision Theory without 'Independence' or without 'Ordering'', in *Economics and Philosophy* 4: 267-290.

Skyrms, B. (1982), 'Causal Decision Theory,' in *The Journal of Philosophy* 79/11: 695-711.

Spohn, W. (1977), 'Where Luce and Krantz do really generalize Savage's decision model,' in *Erkenntnis* 11: 113-134.

Stalnaker, R. (1972/1981), 'Letter to David Lewis,' in W. Harper, R. Stalnaker, and G. Pearce (eds.), *Ifs: Conditionals, Belief, Decision, Chance, and Time* (Reidel), 151-152.

Sugden, R. (2004), 'Alternatives to Expected Utility: Foundations', in S. Barberà, P. Hammond, and C. Seidl (eds.), *Handbook of Utility Theory* (Kluwer Academic Publishers), 685-755.

Vallentyne, P. (1993), 'Utilitarianism and Infinite Utility', in *Australasian Journal of Philosophy* 71/2:212- 217.

Wakker, P. (1988), 'Nonexpected utility as aversion to information', in *Journal of Behavioral Decision Making* 1: 169-75.

White, R. (2009), 'Evidential Symmetry and Mushy Credence', in T. Gendler and J. Hawthorne (eds.), *Oxford Studies in Epistemology* (Oxford University Press).

Zynda, L. (2000), 'Representation Theorems and Realism about Degrees of Belief', in *Philosophy of Science* 67/1: 45-69.