Philosophical Books Vol. 51 No. 1 January 2010 pp. 53–62

REVIEW OF *DECISION THEORY AND RATIONALITY* BY JOSÉ LUIS BERMÚDEZ

MATTHEW KOTZEN
University of North Carolina at Chapel Hill

1. Introduction

Decision Theory and Rationality[1] by José Luis Bermúdez is a terrific book. It should be read by anyone who is interested in the aims, scope, and normativity of contemporary decision theory.

The goal of the book is to examine three dimensions of decision theory. First, there is the action-guiding dimension: we might hope that decision theory will give us tools for solving decision problems, so that we can reason our ways through those decision problems that confront us in everyday life, and thereby come to a rational resolution of them. Second, there is the normative assessment dimension: we would like for decision theory to be a tool for normatively assessing both our own deliberative practices and those of others, and for characterizing some reactions to decision problems as reasonable and others as unreasonable. And third, there is the explanatory/predictive dimension: we would like for decision theory to enable us to both predict and explain the behavior of agents who are confronted with decision problems, thereby making sense of their deliberative practices and their behavior. The question that drives the book is: is there a coherent and unified way of understanding decision theory so that it can play all three roles?

The chapters of the book come at this question from different angles. In chapter 2, the question is how to understand the core notions of *utility* and *preference* so as to enable them to play their crucial roles in all three dimensions. In chapter 3, the question is how to resolve an apparent tension between the predictive/explanatory dimension and the normative dimension of decision theory that arises from the fact that people often conceive of a decision problem differently than they ought to. And in chapter 4, the question is how to extend decision theory to account for rationality *over* time, in addition to rationality *at a* time.

2. Utility and Preference

The primary task of chapter 2 is to explain two understandings of utility and to assess which understanding will enable decision theory to play the roles that were set out in chapter 1.

[1] José Luis Bermúdez, *Decision Theory and Rationality* (Oxford University Press, 2009, 176 pp $50.00 hardcover).

On the *substantive* understanding of utility, "utility is an independently specifiable quantity that is not simply a redescription of the agent's preferences."[2] The substantive conception of utility is compatible with various *theories* of utility; perhaps, for example, utility is pleasure, or good, or well-being, or happiness, etc. But whatever utility is, it is some quantity that the agent assigns different amounts of to various states of affairs, and then acts so as to maximize the expected amount of that quantity in her life. On the substantive understanding of utility, utility is conceptually prior to preference. The agent first assigns utilities to outcomes, and her preferences are to be understood as somehow derivative of that assignment; an agent prefers A to B *in virtue of* her assigning a higher quantity of utility to a world in which A is true than to a world in which B is true.

By contrast, on the *operational* understanding of utility, "utility is simply a representation of preference, which is itself to be understood in terms of choice."[3] In other words, an operational understanding of utility takes an agent's preferences to be conceptually prior to her utilities; we start with preferences that satisfy basic consistency requirements (such as transitivity and substitution), and then we derive (via a representation theorem) a utility function and probability function such that the agent can be understood as preferring scenarios with higher expected utilities (calculated with respect to that utility function and probability function).

In chapter 2, Bermúdez examines how well the substantive and the operational understandings of utility suit the three dimensions of decision theory.

With regard to the deliberation/action-guiding dimension, Bermúdez claims that "the substantive conception of utility sits very naturally with an intuitive understanding of the prescription to maximize expected utility."[4] The idea here is that if utility is some independent quantity that we attach to outcomes, then decision theory gives us some guidance about how to deliberate about what to do; we take our utilities and probabilities as given, and then we choose the course of action that we expect to maximize that independent quantity. By contrast, Bermúdez argues that the operational understanding of utility is ill-suited to ground decision theory's deliberative dimension; if utilities just are representations or summaries of the preferences that the agent already has, then decision theory has very limited practical usefulness, as it seems to recommend simply that the agent act consistently with the preferences that she already has.

Bermúdez argues that the operational understanding of utility "restricts itself to prescribing consistency with past choices" and so "is of no use in situations where there are no past choices with which to be consistent."[5] But this leads to two problems. First, it makes sense to think about maximizing expected utility even in situations that are completely novel, and we would like decision theory to be able to give us some deliberative guidance in such situations. And second,

[2] p. 47.
[3] p. 47.
[4] p. 48.
[5] p. 49.

we sometimes act inconsistently with our past preferences because we have *changed our minds* about what is valuable, and there is nothing irrational or inappropriate about such changes. But from the standpoint of the operational understanding of utility, these changes "can only be completely arbitrary."[6] Since the substantive understanding of utility avoids both of these problems, it is to be preferred to the operational understanding.

I agree that a conception of utility that takes utility to be a representation of preference *which is itself to be understood in terms of choice* suffers from the problems that Bermúdez points to. If my utilities are just numerical representations of the choices that I have already made, then my utilities are not going to be of much practical usefulness in guiding my future choices. But it is not completely clear to me that the operational understanding of utility is without resources to address these problems. Consider a conception of deliberation on which an agent's preference between any two options is most basic, though not to be understood in terms of the choices that he is already made. He simply prefers some options to others at any particular time, and we assign probabilities and utilities to him at that time (via a representation theorem) that make sense of those preferences. Now, he is confronted with two novel options, and he comes to have a preference between them; this new preference gets added to his stock of preferences from which we derive his probabilities and utilities, which may or may not be different from his old ones. And if he changes his mind about what is valuable, then the new preferences that this change brings about will force us (via a representation theorem) to assign new utilities to him, which his new preferences will maximize expected utility with respect to. In other words, I agree with Bermúdez that an operational conception of utility that ties preferences exclusively to past actual choices is ill-suited to ground the action-guiding dimension of decision theory, but it seems to me that there is room for an operational theory of utility that still takes preferences to be more theoretically basic than utilities, but allows new (theoretically basic) preferences to arise in an agent as a result of his changing his mind or of his being presented with a novel choice.

None of this is to disagree with anything that Bermúdez actually says; he is explicit about the operational understanding of utility involving a commitment to preference being understood solely in terms of past actual choice, and he suggests that this is the way that many social scientists deploy the operational understanding of utility. But because the problems that Bermúdez raises seem to derive from this additional commitment, I am hesitant to accept the conclusion that utilities need to be understood as more theoretically basic than preferences. In short, I think that Bermúdez's characterizations of the substantive and the operation understandings of utility are not exhaustive. There is a third option—an operational understanding of utility that severs the necessary connection between preference and past choice—and the considerations that Bermúdez raises against the operational understanding of utility are not fully successful in supporting the substantive understanding of utility, since it is not at all clear that they also count against this third option.

[6] p. 50.

Next, Bermúdez argues that in addition to the two problems raised for the operational understanding of utility in the context of deliberation (which he thinks also affect the normative assessment dimension of decision theory), there is an additional problem for the operational understanding that is specific to the normative assessment dimension. If decision theory is to be a tool for normatively assessing agents, then it needs to be the case that its requirements can be reflectively grounded; the mere fact that some unmotivated axioms entail some requirements is of little interest unless the axioms are plausible when applied to the intended model (in this case, decision problems faced by an agent). Bermúdez claims that this is the case with Peano Arithmetic; the Peano axioms are independently plausible, and that independent plausibility gives rise to the normative force of theorems that follow from those axioms. But in the case of decision theory, Bermúdez thinks that "we do not have the sort of clear intuitive grasp on what rationality demands that would bestow authority on axioms such as the substitution axiom in the way that our intuitive grasp on the natural numbers bestows authority on the axioms of Peano arithmetic."[7] As a result, we need the representation theorem to reflectively ground the axioms. The representation theorem shows that agents who obey the axioms are (or can be understood to be) maximizers of expected utility, and the independent desirability of being a maximizer of expected utility provides normative support for the axioms from which the representation theorem is derived. But, Bermúdez continues, this sort of support for the axioms which comes from the desirability of maximizing expected utility can only be underwritten by the substantive understanding of utility. It is only if we have independent purchase on the desirability of maximizing the expected quantity of utility that we can use the representation theorem to support the axioms; on the operational understanding, maximizing expected utility *just is* following the axioms, and so there is no distance between the two in virtue of which the latter can receive normative support from the former.

But again, I am not sure I find this a compelling reason to prefer the substantive understanding of utility to the operational one. When we know that the intended model of the Peano axioms is the natural numbers, those axioms do have considerable intuitive force. But so too, I think, do the axioms of decision theory have considerable intuitive force when we know of their intended model. Take, for example, the substitution axiom, which says that if I prefer X to Y, then I also prefer any lottery in which X is embedded to the same lottery in which Y is identically embedded. So, for example, if I prefer chocolate to vanilla, then I also prefer a 90% chance of chocolate (and, say, a 10% chance of strawberry) to a 90% chance of vanilla (and a 10% chance of strawberry). This strikes me as overwhelmingly plausible; I just would not know how to understand someone who claimed to prefer chocolate to vanilla, but then preferred a 90% chance of vanilla to a 90% chance of chocolate when all else was equal. The representation theorem shows that if an agent's preferences obey plausible consistency constraints such as the substitution axiom, then she can be assigned a probability and a utility function such that she prefers options

[7] pp. 51–2.

with the highest expected utility, calculated via those functions. Put less cautiously, it shows that if your preferences are consistent, then you are a maximizer of expected utility. But I would have thought that the requirement to be consistent in your preferences was at least as plausible a normative requirement as the requirement to multiply utilities by probabilities and sum for each option, and then to prefer the option with the highest such sum; the interest of the representation theorems is precisely that they claim to show that the latter requirement follows from the former. And I do not see why the operational understanding of utility is at a significant disadvantage here. Even if my utilities are just representations of my theoretically more basic preferences, it is still interesting that plausible consistency constraints on my theoretically basic preferences give rise to the requirement that I treat the representations of my preferences (i.e., my utilities) in the way required by decision theory. And I do not see any reason why the plausibility of the consistency constraints on my preferences cannot normatively support the requirement that I maximize expected utility, even if it is the preferences rather than the utilities that are most theoretically basic.

Finally, with respect to the psychological prediction/explanation dimension of decision theory, Bermúdez claims that only the substantive understanding of utility can capture "the full force of thinking about decision theory as a regimentation of commonsense psychological explanation."[8] The worry here is that if utilities and probabilities are going to be able to provide genuine explanations of behavior, they have to be formalizations of the folk-psychological notions of belief and desire. But since belief and desire are real, theoretically basic, psychological notions in terms of which behavior is explained, utilities and probabilities must be similarly real and basic in order to do the explanatory work of the concepts that they are surrogates for.

Put this way, the argument seems to beg the question against someone who thinks that some folk-psychological notion of preference (rather than of belief and desire) is sufficient to explain behavior ("she chose A over B because she prefers A to B"); someone who thought this would presumably be perfectly happy with the explanations provided by the formal notion of preference that appears in decision theory, since they would be analogous to the explanations provided by the folk-psychological notion. Again, I agree with Bermúdez that if preferences are theoretically basic *and to be understood in terms of past actual choice*, then my utilities and probabilities would not be able to explain my behavior, since my utilities and probabilities will be "simply redescriptions of the behavior being explained."[9] But a conception of preference as theoretically basic seems, at least in principle, to be able to do all of the explanatory work that a conception of utility as theoretically basic is able to do. Admittedly, we do give folk-psychological explanations in terms of beliefs and desires, but I am not yet persuaded that this practice is legitimate only if desire is more theoretically basic than preference.

[8] p. 53.

[9] p. 53.

3. Individuating Outcomes

Chapter 3 is devoted to a tension at the heart of decision theory between its normative assessment dimension and its explanatory/predictive dimension. To the extent that we're interested in explaining and predicting people's behavior, it seems like we need to take account of the way that they *actually* conceive of a decision problem. And to the extent that we are interesting in normatively assessing their deliberative behavior, it seems like we need to take account of the way they *ought to* conceive of a decision problem. Of course, these two are not completely divorced from one another, since people often do conceive of decision problems in the way that they ought to. But there are also cases in which an agent conceives of a decision problem in a different way than he ought to, and it might seem as though there could be no one theory that gives us the resources both to predict/explain his behavior, and also to normatively assess his deliberation about what to do.

There are various ways in which an agent might conceive of a decision problem differently than she ought to. One way this might happen is if an agent fails to assign a sufficiently high probability to a comparatively likely outcome; two of Bermúdez's examples here are people who do not purchase sufficient insurance against natural disaster (because they assign too low a probability to a natural disaster affecting them), and people who do not wear helmets while bicycling (because they assign too low a probability to their getting into a serious accident). Another way is if an agent distinguishes two outcomes that are in fact one; this arguably happens in famous cases of "framing effects," such as ones where people express a preference for a scenario described in terms of the number of lives saved over the identical scenario described in terms of the number of lives lost.

Bermúdez identifies two different reactions that we might have to cases where an agent conceives of a decision problem differently than he ought to. The *incompatibilist* view is that since decision theory is based on certain axioms characterizing the ways that agents ought to conceive of decision problems, it is useful only as a tool for normative assessment, and it is of absolutely no use in predicting or explaining the behavior of agents who violate those axioms. By contrast, the *compatibilist* view is that we can apply the principle of expected utility to an agent's way of conceiving of a decision problem, and thereby explain/predict the behavior that his conception of the decision problem will lead him to, without normatively endorsing his conception of the decision problem. So, even when one of the axioms of decision theory is breached, decision theory can still be a useful tool for the prediction and explanation of behavior.

The worry that Bermúdez identifies for the compatibilist approach is that it

> rests on the possibility of insulating the efficacy of the expected utility principle from the parsing of the decision problem and the assignment of utilities and probabilities. But this is hardly something that can be taken for granted. It could well be that the requirements of psychological explanation

> and prediction lead us to assignments of probabilities and utilities that are incompatible with applying the expected utility principle.[10]

Bermúdez spends the remainder of chapter 2 articulating two possible ways of modifying decision theory (due to Schick and Broome) that offer a hope of allowing decision theory to play the dual role of explanatory/predictive theory and theory of normative assessment. Bermúdez ends up rejecting these proposals, arguing (completely convincingly, in my view) that neither of them successfully delivers a single theory that is capable of playing both roles.

This is a useful exercise, but, in a sense, I think it supports an entirely unsurprising conclusion. Who would have ever thought that a single theory could be both an explanatory/predictive theory of human behavior and a theory of normative assessment of that behavior? In what other domain is anyone even looking for a theory to play such a dual role? Does anyone think that a theory of epistemic rationality will also be a theory of actual human reasoning? Does anyone think that an ethical theory will also be a theory of moral psychology? In almost any domain, the axioms that we formulate from which we derive normative results are axioms that real human reasoners frequently violate. People reason in ways that are different from how epistemologists think they ought to, people act in ways that are different from how ethicists think they ought to, and people deliberate about how to act in ways that are different from how decision theorists think they ought to.

Of course, if we assume that the relevant agent is epistemically rational, or moral, or deliberatively rational, then we can do a fair amount of prediction and explanation. Knowing that the moral thing to do is to φ *and that John is moral (or at least acting morally on this occasion)* gives us some basis from which to predict/explain his φ-ing. And knowing that the rational resolution of Maria's decision problem is to φ *and that Maria is deliberatively rational (or at least deliberating rationally on this occasion)* allows us to predict/explain her φ-ing. But in such cases, we have two theories, not one. We have one axiomatic theory that delivers the normative results, and then we have another descriptive theory that takes as input premises about the agent's rationality (among other things) and delivers descriptive results about the agent's behavior.

Moreover, there seem to be some breaches of rationality that make the task of predicting/explaining an agent's behavior close to impossible. One of the axioms of decision theory is that the agent has a complete and transitive weak preference ordering over outcomes. But, of course, very few (if any) of us have such an ordering, and there are various cases where agents can be shown to have intransitive preferences; that is, they prefer A to B, and B to C, but also prefer C to A. If such an agent is presented with options A, B, and C, how could we possibly predict what she is going to choose, or explain what she chose once she chooses? There are, of course, less dramatic failures of rationality that might allow us to do some prediction and explaining. If Jerome reliably prefers outcomes which are characterized in terms of the number of lives saved over identical outcomes characterized in terms of the number of lives lost, then we

[10] p. 86.

might be able to predict future preferences between such identical options characterized differently. And we may even be able figure out by precisely how much Jerome discounts a life that fails to be saved against a life that is lost, so that we could even predict his preferences among nonidentical options. But again, I just do not see that any of this has anything to do with how Jerome ought to conceive of the decision problem, or with any normative results in the neighborhood.

4. Rationality Over Time

In chapter 4, Bermúdez addresses a tension that arises when we apply decision theory in the diachronic, rather than just the synchronic, case. Bermúdez identifies two types of what he calls "sequential inconsistency," which are cases where "an agent makes a plan to choose in a particular way at a later time and then, when that time comes, chooses differently."[11]

In the first type of sequential inconsistency, constant preference sequential inconsistency, an agent has preferences that violate the substitution axiom; perhaps, for example, she prefers chocolate to vanilla, but prefers a 90% chance of vanilla (and a 10% chance of strawberry) to a 90% chance of chocolate (and a 10% chance of strawberry). Consider some event E, which the agent regards to be 90% likely. Since she prefers a 90% chance of vanilla (and a 10% chance of strawberry) to a 90% chance of chocolate (and a 10% chance of strawberry), she prefers the option of vanilla if E occurs (and strawberry otherwise) to the option of chocolate if E occurs (and strawberry otherwise). Thus, her preferences commit her to making a plan to choose vanilla over chocolate if E occurs. However, if E actually occurs, the agent is faced with a choice between chocolate and vanilla, and her preference for chocolate over vanilla commits her to choosing chocolate over vanilla, contradicting the earlier plan that she made.

In the second type of sequential inconsistency, preference reversal sequential inconsistency, the agent simply changes her mind between the earlier time and the later time. She prefers vanilla to chocolate at t_1, which commits her to planning to choose vanilla over chocolate when she is presented with that choice later on. But then, between t_1 and t_2, something happens to her that makes her come to prefer chocolate over vanilla; thus, when she is actually confronted with the choice at t_2, she chooses chocolate, again contradicting her earlier plan.

Bermúdez argues that from the perspective of the action-guiding dimension of decision theory, decision theory seems to "tacitly involve" a separability principle, which says that only the agent's preferences *at t* are relevant to what she should do *at t*; the preferences that she had earlier and the plans that she made earlier are irrelevant to what she ought to do *now*. If that is right, then since separability mandates both types of sequential inconsistency, decision theory (in its action-guiding dimension) mandates sequential inconsistency.

[11] p. 114.

However, from the standpoint of the normative assessment dimension of decision theory, Bermúdez argues that things look different. After all, Bermúdez claims, we sometimes want to criticize an agent for choosing in a sequentially inconsistent manner; there certainly seems to be *something* wrong with our agent above who makes a plan to choose vanilla over chocolate if E occurs, but then chooses chocolate over vanilla once E actually does occur.

Bermúdez spends a good part of chapter 4 explaining and discussing strategies for resolving this tension, but I must confess that I am confused over the nature of the alleged tension.

In the case of preference reversal sequential inconsistency, in at least a lot of cases, there just does not seem to be anything normatively criticizable about the agent's choices. She preferred vanilla to chocolate and t_1, and then came to have a different preference at t_2, but she had perfectly consistent preferences *at each particular time*, and surely agents are allowed to have their preferences change over time. Now, of course, there is an interesting question about whether all such changes in an agent's preferences are rational (or at least not normatively criticizable); perhaps normatively acceptable changes in preference can come about only through certain rational changes in my probability or utility function. Perhaps, for example, a change in preference that comes about due to an epistemically irrational response to new evidence is itself irrational. So, we may need to do some work in order to distinguish the rational cases of preference change from the irrational cases, but I do not see any special reason to think that this project is going to introduce a tension between the action-guiding and the normative assessment dimensions of decision theory.

In the case of constant preference sequential inconsistency, there certainly does seem to be something normatively criticizable about the agent's preferences and choice behavior. But I do not see that there is anything essentially diachronic about this defect. The agent was irrational at t_1 for having preferences that violated the substitution axiom, and since her preferences have not changed from t_1 to t_2 (or else this would be a case of preference reversal sequential inconsistency), she still has preferences that violate the substitution axiom at t_2. Granted, it took a change (namely, the agent learning between t_1 and t_2 that E occurred) to make this violation of the substitution axiom manifest, but the agent was already normatively criticizable, even back at t_1 (before she learned that E occurred), for having preferences that violated substitution. And if those preferences have not changed, then she is just as normatively criticizable at t_2.

So, if there is normative force to the synchronic axioms of decision theory, then we already have the resources to criticize an agent who is constant preference sequentially inconsistent, and there does not seem to be a need to appeal to anything diachronic in order to underwrite that criticism.

Of course, there is still the question of what an agent should do when she violates one of the synchronic axioms of decision theory and conceives of a decision problem differently than she ought to. When an agent does violate substitution, should she choose in accordance with her earlier plan (to choose vanilla over chocolate if E occurs), or should she choose in accordance with her current preference for chocolate over vanilla? Normatively speaking, decision theory just does not seem capable of giving a univocal answer; she should not have violated substitution, but given that she has, she should choose in accordance with her relevant current preference. The same problem came up in the context of individuating outcomes in chapter 3: when an agent conceives of a decision problem differently than he ought to (say, by distinguishing two outcomes that are in fact one), how should he choose? Again, no univocal answer is possible: he should not have conceived of the decision problem that way, but given that he did, he should choose the outcome which, on his (normatively criticizable) way of conceiving of the decision problem, looks to have the highest expected utility. Bermúdez helpfully introduces the distinction between hypothetical rationality and *all-things-considered* rationality here, but I find it hard to understand his claim that there is some special tension between the action-guiding and *all-things-considered* normative assessment dimensions of decision theory that is generated by the case of sequential inconsistency.

5. Conclusion

In summary, *Decision Theory and Rationality* is an interesting and admirably comprehensive book that addresses issues that need to be untangled if we are to have a clear conception of the theoretical scope of modern decision theory. I have my reservations about some of the arguments in the book, and I am not sure that I agree with Bermúdez about all of the sources of tension between the three dimensions of decision theory that he identifies, but it is certainly a careful and sophisticated analysis of deep issues in decision theory than anyone with a serious interest in decision theory should read.