Securing Decision Theory’s Foundation: Why the Theory of Practical Rationality Needs to Include a Substantive Notion of Utility

ABSTRACT. In a certain sense, decision theory is one of the most precisely worked-out domains of philosophy. Its central theorem, the “representation theorem,” has been proven with mathematical rigor. But, as some have pointed out in the recent literature, the technical aspects of decision theory have been explored far more carefully than have its philosophical foundations (see, e.g., Bermúdez (2009), Broome (2002), Dreier (1996), Sen (1973)). In this vein, I argue that the dominant interpretation of decision theory’s notion of *utility* (which I label *operationalism*) leaves the axioms at its foundation without their needed mooring. If decision theory is to serve as a theory of practical rationality, its axioms must be defensible as rational requirements. And, I argue, operationalism makes such a defense impossible: only given a *substantive* understanding of utility can decision theory be a theory of practical rationality.

Introduction

Decision theory is sometimes described as reducing practical rationality to a simple directive, applicable whenever one is faced with a decision problem: “Maximize expected utility!” The suggestion is that the proper approach to a decision problem is to maximize some quantity, “expected utility,” which, grammatically at least, seems to be composed of two parts: *utility* and *expectation*. Here, *expectation* roughly corresponds to something like your *degree of belief*; it is the likelihood you attach to various outcomes’ obtaining, conditional on your performing a given action, expressed in probabilistic terms. There are many complications involved in deciding just how to spell out this notion of degree of belief, or *credence*, but I will be focusing on questions that arise in relation to the other basic element of decision theory: *utility*.

Utility is supposed to be related in some way to an agent’s preferences; decision theory instructs the agent to maximize expected utility because it seems reasonable to think that rationality consists, at least in part, in acting in ways that satisfy one’s preferences. But how exactly to think of decision theory’s notion of utility, and its relation to preference, is no settled matter. José Luis Bermúdez, in Chapter 2 of his *Decision Theory and Rationality* (2009), contrasts two different understandings of utility. On the *operational* understanding, “utility is simply a representation of preference, which is itself to be understood in terms of choice.” On the *substantive* understanding, by contrast, “utility is an independently specifiable quantity that is not just a redescription of the agent's preferences."[1] The *substantive* view says that utility is some real entity: either some mental quantity like "strength of desire" or "degree of subjective good"; or else some objective quantity, like "amount of (actual) good." This quantity is, at least in principle, separable from an agent's preferences (and might in some way *cause* or *explain* those preferences). The operational view, or *operationalism* (about utility), on the other hand, says that utility is a mere representation of the agent's preferences: it is a theorist's tool, constructed mathematically out of the agent's preferences, and it is not (even notionally) separable from those preferences.

The operational view is dominant within economics, and it is embraced by many philosophers (particularly those, such as James Dreier, who describe themselves as "Humean" decision theorists). The substantive view is more closely linked to the historical use of the term "utility" in philosophy, dating back to Bentham.[2] Some philosophers have, in recent years, explicitly addressed the question of how these two concepts, the substantive and operational notions of utility, might be related. But, as John Broome has pointed out, decision theorists in general have had a tendency simply to conflate the two conceptions. Broome quotes John Harsanyi, who claims that the primary purpose of the von Neumann-Morgenstern (vNM) utility function (the original "operationalized" utility-construct) is *not* to represent a subject's choice behavior; instead, "it is to indicate how much utility, i.e. how much subjective importance, he assigns to various goals." Here we have a substantive notion of utility (the quantity of "subjective importance" an agent attaches to an outcome) equated with an operational one (the numerical value used to represent the "choice behavior under risk" from which vNM utilities are derived).

[1] Bermudez (2009), p. 47.

[2] See Broome (2002) for a discussion of the historical usage of the term "utility," in both its guises.

It's possible that an argument could be made for identifying the constructed value we get out of an operational procedure with some substantive notion of utility.[3] But the version of *operationalism* that I have in mind explicitly denies that such an identification should be made; indeed, it denies that there is any substantive notion of utility to be had. This view, which is closely associated with the co-called "theory of revealed preference," denies that the utility function of modern axiomatic decision theory is anything more than a mathematical construct (one used to represent a subject's choice behavior). The "utilities" the function assigns to outcomes are not intended to capture any underlying motivational states that cause or explain the subject's behavior.

The operationalist's position is stated most clearly by Luce and Raiffa (1953):

> [T]here is no need to assume, or to philosophize about, the existence of an underlying subjective utility function, for we are not attempting to account for the preferences or the rules of consistency [i.e., the axioms]. We only wish to devise a convenient way *to represent* them.

Luce and Raiffa go on to warn that a substantive understanding of utility, on which utility logically precedes (and explains) preference, amounts to "a fallacy":

> In [axiomatic decision] theory it is extremely important to accept the fact that the subject's preferences among alternatives and lotteries came prior to our numerical characterization of them. We do not want to slip into saying that he preferred *A* to *B* because *A* has the higher utility; rather, because *A* is preferred to *B*, we assign *A* the higher utility.[4]

The key point here is the *order of explanation*. According to the operationalist, a subject's preferences and the mathematical machinery of the theory explain why a certain outcome is assigned a numerical "utility" value. To suggest that a subject prefers one outcome to another *because* that outcome has a higher utility would be to get things backwards.

[3] Harsanyi himself certainly does not supply such an argument. In the passage Broome quotes, Harsanyi justifies his identification of the vNM utility function with a substantive notion of utility simply by saying that, "as its name shows, it is a utility function." But the *name* given to the constructed vNM function certainly doesn't, by itself, license us to equate it with an entirely distinct concept, "subjective importance," which also sometimes goes by the name "utility." So an argument for the identification is wanting.

[4] Luce and Raiffa (1957): pp. 22-32.

This operational view is, more or less, the prevailing orthodoxy among decision theorists.[5] Below, I present a challenge to operationalism, arguing that it cannot provide a suitable foundation for modern decision theory. The problem is that, in order for decision theory to serve as a *theory of practical rationality*, its axioms must be intelligible as *rational constraints*.[6] And, I will argue, given an operational understanding of utility, on which the order of explanation runs from preferences to "constructed" utility assignments, decision theory's axioms cannot be defended as rational constraints.

The paper proceeds as follows. In the next two sections, I will consider first a "strong" and then a "weak" version of operationalism. I will argue, following Amartya Sen, that *strong operationalism* is untenable. I will then go on to consider the *weak operationalism* endorsed by Dreier. Though the weaker version of operationalism avoids the worry presented by Sen, I will argue that it, too, fails to provide a suitable grounding for decision theory. Weak operationalism's picture of the relation between preferences and utilities is not a realistic one; and, even more importantly, the weak operationalist view deprives decision theory's axioms of their needed grounding in real, rationally constrainable mental states. In Section 3, I take a brief look at a more specific problem for the operational view: I argue that operationalism cannot supply a defense of the Continuity axiom as a rational constraint (a defense that an opponent is entitled to demand from any theorist who claims that decision theory is a theory of

[5] When I say "prevailing orthodoxy," I do not mean to suggest that most decision theorists explicitly endorse the kind of position laid out in the Luce and Raiffa quotations above. Instead, I suspect, decision theorists are more likely to be following Harsanyi in simply conflating the operational and substantive views of utility. When the question is posed explicitly, however, most decision theorists do accept the Luce and Raiffa view of the order of explanation between utility and preference (i.e., that preferences come first, allowing us to construct "utilities," which are not antecedently-specifiable real entities).

[6] As Bermudez (2009) and Dreier (1996) note, there are many uses to which decision theory might be put. It might be used as a (normative) theory of practical rationality; as an explanatory tool for analyzing human (or animal) behavior; or as a predictive tool for modeling various human and natural phenomena. Throughout this paper, I will be treating decision theory as a candidate theory of practical rationality; that is, I will be considering whether, and how, decision theory might impose demands of rational coherence on our actions, and on the patterns of choices and preferences that underlie those actions. I think that, if decision theory is to do any real work at all, it must be via this route; by my lights, decision theory seems rather ill-suited to the more descriptive tasks for which it is sometimes employed. Dreier likewise addresses the question of how to spell out decision theory as a normative theory of practical rationality, while Bermudez raises worries about whether decision theory's different "roles" can really be pulled apart at all (see esp. Chapter 5; I won't have space here to address Bermudez's worries).

practical rationality). Finally, in Section 4, I turn to what I take to be the underlying motivation for the operational view: a desire to respect the Humean idea that preferences are “original existences,” not subject to rational assessment. I suggest that the operational view does not in fact provide a satisfying way to understand Hume’s insight. In the end, I will suggest that a substantive understanding of utility provides a more promising basis for decision theory as a theory of practical rationality, even if, as the operationalist insists, the kind of theory we want is a Humean one.

1. Strong Operationalism

Sen (1973) argues that there is a flaw in the foundation of operationalism. His discussion focuses on the “theory of revealed preference” (henceforth, “TRP”), originally developed by the economist Paul Samuelson in the late 1930s. The idea of TRP is to use an agent’s *behavior* to determine her preferences, and, from those preferences, her utilities. Sen quotes Samuelson: "The individual guinea-pig, by his market behaviour, reveals his preference pattern.” As Sen notes, this method allows the scientific observer of behavior to determine an agent’s preferences from her choices, even though we normally think of preferences as *driving* choices (and the attendant behavior). According to TRP, “choices are observed first and preferences are then presumed from these observations.” But merely reversing the order of explanation between preferences and choices is not all that TRP seeks to achieve; according to Sen, economists in the grip of TRP frequently claim that “the theory of revealed preference ‘frees’ demand theory from the concept of preference and *a fortiori* from the concept of utility on which traditional demand theory was based.”

Sen argues against the idea that TRP can explain “behaviour without reference to anything other than behaviour" (I.M.D. Little). Sen’s argument is that strongly behaviorist claims like these are incompatible with decision theory's use of the concept of "inconsistency," and with the axioms at its foundation. The argument goes something like this:[7]

(1) TRP tries to spell out decision theory in terms of "choice-behavior," without making reference to any "hidden" mental states underlying a subject's behavior.
(2) Decision theory makes claims about whether two or more choices are "inconsistent" (for example: "It is inconsistent to choose *A* over *B* and also choose *B* over *A*").
(3) But choices themselves, construed as mere behavior with no mental states underlying them, cannot be inconsistent (for example: a tree that bends to the left rather than the right and then to the right rather than the left is not being inconsistent).
(4) It is only underlying *preferences*, construed as mental states, that make the concept of "inconsistent choices" intelligible (for example: preferring *A* to *B* and also preferring *B* to *A is* inconsistent).
(5) So decision theory, in labeling certain sets of choices "inconsistent," must be making reference to a set of preferences *underlying* an agent's choices in order for the theory to be intelligible.
(6) So decision theory cannot be spelled out in terms of a notion of choice that makes no reference to any underlying mental states.

Sen's thought is this: Decision theory is built on a set of axioms about, more or less, "consistency." This notion of consistency gets its intuitive grounding from our pre-theoretic concept of "preferences" as (rational) mental states. Without that grounding, we'd have no understanding of what the "consistency" of an agent's choices amounted to.[8]

Sen's argument is couched in terms of preferences and their relation to behavior (choice behavior, specifically). But the question I began with was whether we should have a substantive or operational understanding of *utility*. The reason that preferences seemed a relevant place to begin the discussion stems from the way *operationalism* is typically defined. Bermúdez, for example, describes *operationalism* as the view that "utility is simply a representation of preference*, which is itself to be understood in terms of choice*" (my emphasis). This definition represents what I will call *strong*

[7] Sen (1973), pp. 241-243.

[8] Sen also considers the possibility of removing the notion of "consistency" from the theory and making it a purely empirical hypothesis about patterns of human behavior. He argues that the axioms of decision theory are not principles we could have arrived at by observing human behavior, because the behavior itself is too impoverished a data set, and too hard to interpret without relying on the very notion of consistency at issue. So: decision theory cannot help itself to a notion of consistency if (in a behaviorist mode) it cuts itself off from the pre-theoretic concept of a preference as a mental state; and we cannot even generate decision theory's axioms by looking at behavior if we *don't* have the notion of consistency in play. Thus, we must, after all, look inside the head in order for decision theory to make any sense.

operationalism (about utility). *Strong operationalism* seeks to provide a thoroughly behaviorist definition of utility by reducing not only utility, but also preference, to observable behavior. The appeal of TRP may in large part stem from this motivation: if utility can be reduced to preference, which can in turn be reduced to observable choices, then we can make decision theory scientifically respectable (or at least, "scientifically respectable" as it was defined in the heyday of behaviorism) by removing *all* reference to any "hidden" underlying mental states. Sen's argument shows us that this ideal of scientific respectability is not attainable for decision theory (this is no great loss; the behaviorist ideal of scientific respectability has fallen out of favor, for good reasons). But Sen makes his case against *strong operationalism* by pointing out the impossibility of construing *preferences* behavioristically; his argument does not yet say anything about the question of whether *utility* should be understood substantively or operationally. We need *some* substantive mental states to serve as bedrock for decision theory; but it is important to note that, for Sen, the bedrock is the notion of *preference*, not *utility*.

2. Weak Operationalism

The previous section suggested that we cannot make sense of decision theory if its notion of "utility" is given a *strong operationalist* definition. We need the term "preference" to maintain its pre-theoretic meaning, which includes (in Sen's phrase) a "peep into the head"; a purely behaviorist decision theory can't get off the ground. But there is a weaker version of operationalism about utility that is not behaviorist "all the way down." This version, which I label *weak operationalism*, allows that preferences are real mental states; but it suggests that *utility*, in the role it plays in axiomatic decision theory, is a mere representation of these preferences, with no *further* significance of its own. On this view, there is no independent "degree of desire" or "level of subjective importance," no real mental *quantity*, that we are tracking by representing preferences with a utility function. There are preferences, and there is a mathematical method for representing them. That is all.

Dreier (1996), in defending *weak operationalism*, makes essential use of a distinction between "ordinal" and "cardinal" preferences. Ordinal preferences are simple (non-quantitative) preferences for one thing over another. There is no "strength parameter" in ordinal preferences; they say that a subject S prefers *A* to *B* to *C*, but nothing about "how much" S prefers *A* to *B* or *B* to *C*. Cardinal preferences, on the other hand, include a quantification of the "strength" of a preference. A cardinal preference might say, for example, that S prefers *A five times as much as B* (or, more precisely: S prefers *A* to the status quo five times as much as she prefers *B* to the status quo). The *weak operationalism* endorsed by Dreier holds that there really are *ordinal* preferences, substantive mental entities that might require a "peep into the head" to determine or define; but that there are no real *cardinal* preferences, no independently specifiable mental entities that we might call "strengths of preferences." Decision theory does allow us to *construct* a measure of the strength of a preference; using an agent's real ordinal preferences and the mathematical machinery of the theory, we can represent those preferences in a way that allows us to talk "as though" they came in degrees of strength. But we are not tracking any real mental entity, or any pre-theoretic notion of strength of preference, with this constructed value (which we label "utility")—there is no such pre-theoretic notion, no such mental entity.[9]

Dreier argues against the *substantive* view (that is, the idea that there are real cardinal preferences) by noting that we cannot in fact introspect preference "scores." He gives the following example. Suppose you prefer a pomegranate to an apple. Dreier claims you cannot, through

[9] Dreier frames his discussion as an argument against a version of the substantive view on which utility is to be understood as "strength of preference," and I will follow him (for the most part) in considering this kind of substantive view. My argument against weak operationalism is not, however, meant to be an argument for this (or any) specific *version* of the substantive view. My claim will be that we need to understand utility as corresponding to *some* real, substantive *quantity*; that quantity could be "strength of preference," or it could be "degree of desire," or it could be "amount of perceived good." Any of these substantive notions of utility could be substituted, *mutatis mutandis*, for "strength of preference" in what follows, without affecting the thrust of the argument.

introspection, figure out "how much" you prefer the pomegranate. Even if you assign a unit value by setting it as "the amount by which you prefer getting an apple to the status quo," Dreier claims there is no hope in the enterprise of trying to introspectively ascertain how much *more* you'd like the pomegranate. According to Dreier, "It is just not plausible that we can introspect utilities, or their ratios, directly."[10]

Below, I present three objections to Dreier's *weak operationalism*. The first offers intuitive considerations that tell against Dreier's (equally intuition-based) dismissal of the idea that we have a pre-theoretic notion of "strength of preference." The second questions where the normative force of decision theory's motto ("Maximize expected utility!") might come from on an operational view. Dreier's reply to this objection will bring us to the third objection, which is that the *axioms* of decision theory cannot be justified as principles of practical rationality if, as Dreier advises, we take an operational view of utility.

Objection 1: We do have a pre-theoretic notion of the strength of a preference

Dreier writes, "There is no particular reason to think, in advance of theorizing, that our preferences have any particular quantifiable dimension called 'strength'."[11] This line of thought seems to me to be based on an inaccurate portrayal of our intuitive notion of preference. While it seems true that we can't identify exact numerical "scores" for our preferences, we do have some (perhaps not very precise) sense of the strength of our preferences. For example, I prefer a pomegranate to an apple, and an apple to the status quo; but I also prefer world peace to a pomegranate. On Dreier's picture, this *ordering* (world peace > pomegranate > apple > status quo[12]) is all that is psychologically real about the structure of my preferences. But that seems patently false; another psychological fact that is true of me

[10] Dreier (1996) pp. 269-270.

[11] Ibid., p. 252.

[12] A note on notation: I'll be using the simple ">" symbol to represent "is strictly preferred to."

is that I prefer world peace to a pomegranate *much more* (or *much more strongly*) than I prefer a pomegranate to an apple. And this is not some outgrowth of my knowledge of the mathematical machinery of decision theory; it was an obvious intuitive fact about the structure of my mental states before I started studying that machinery. So it seems we do indeed have a pre-theoretic notion of “strength of preference.”

Objection 2: Why should we maximize expected utility, if utility is merely a theoretical construct?

As noted earlier, decision theory advises agents to maximize expected utility; it says that practical rationality consists in maximizing this quantity. But if utility is merely a mathematical construct with no pre-theoretically specifiable content, it seems mysterious why this directive would have any normative force. If utility is taken to represent a real quantity that corresponds to an agent’s good, or “the subjective importance” she attaches to various outcomes, or the strength of her preferences (in other words: if utility is given a “substantive” reading), then it might be plausible to think that rationality consists in trying to maximize expected utility. But if we *don’t* attach the value we get out of axiomatic decision theory to some such pre-theoretically intuitive entity, we would be hard-pressed to say why it is that a rational agent ought to maximize this “constructed” value.

A quick response is on hand for Dreier, one that he himself emphasizes. Dreier notes that “Maximize expected utility!” is not really the best way to spell out the advice given by decision theory, on an operational view. The theory says that, as long as you are obeying some basic axioms about how to structure your (real) *ordinal* preferences, you will *thereby* be maximizing utility. So the idea is not to try to maximize some pre-existing quantity, utility. Instead, the idea is that, in order to be rational, you should try to *follow the axioms*; doing so will mean you are “describable” by a properly behaved utility function. Dreier says we should abandon “Maximize expected utility!” in favor of “Attend to the axioms!” as the motto of decision theory.[13]

Objection 3: Why should we attend to the axioms?

Dreier’s response to Objection 2 seems reasonable. But it leaves us with a further question, one that I think poses real problems for *weak operationalism*. The question is this: *Why* should we attend to the axioms? Or, more fully: Why should we think that being rational amounts to following the axioms; why should we think that the axioms represent rational requirements?

This question itself is not, perhaps, devastating to weak operationalism: the operationalist could probably provide some intuitive considerations in favor of obeying the axioms, and in favor of seeing them as rational requirements (such considerations in favor of the axioms are presumably something that *any* proponent of decision theory should be prepared to provide). But the problem emerges when we consider the special set of “restrictions” the operationalist would be operating under when trying to provide her defense of the axioms. Because she has declared “strengths of preferences” (and all related notions of substantive utility) to be mere *outgrowths* of the theory, without any pre-theoretical grounding, she cannot point to any such notion in defending the axioms of the theory themselves. And, I claim, there is simply no way to make sense of the axioms of decision theory as rational requirements if we cannot make use of some kind of substantive, pre-theoretic notion of “strength of preference” (or something similar).

My worry is this. There are certain axioms of decision theory, those that mention *betting behavior* (which Dreier calls the “Cardinal Axioms”), which seem to bring in intuitive notions not licensed by weak operationalism’s project of explaining decision theory on the basis of a purely “ordinal” notion of preference. These Cardinal Axioms, which are needed to generate a utility function (a function Dreier

[13] Ibid., p. 253.

says *can* eventually be seen as the basis of *post hoc* "strength-of-preference" judgments), make no sense (as purported rational requirements) if we don't *already* have on board a notion of the "strength" of a preference. The weak operationalist's highly-circumscribed look inside the head doesn't provide the materials for seeing an agent's betting behavior as anything more than *mere* behavior.

I am suggesting that the Cardinal Axioms present a problem for *weak operationalism* analogous to Sen's problem for *strong operationalism*. As Sen points out, the idea that choices could even be sensibly seen as "consistent" or "inconsistent" presupposes that those choices are not mere behavior, but rather are driven by underlying *preferences*, in the intuitive, pre-theoretic sense of *real mental states*. Similarly, the idea that a specific pattern of *betting* behavior is "rational," and that an agent is "irrational" if she fails to structure her betting behavior in the way demanded by the Cardinal Axioms, presupposes that that betting behavior is driven by a real mental state and is not mere behavior. The weak operationalist allows in the mental states that save decision theory from Sen's worry; weak operationalism says that ordinal preferences are genuine mental states. But the mental states needed to ground claims about the Cardinal Axioms' being *rational norms* cannot be simple ordinal preferences.

To see why, consider two "sets" of my preferences:

S_1: world peace (*W*) > an apple (*A*) > the status quo (*S*)
S_2: a banana (*B*) > an apple (*A*) > the status quo (*S*)

The *ordinal structure* of my **S_1** preferences will match that of my **S_2** preferences. But if we start talking about my preferences among various *gambles* involving the outcomes in **S_1** and those in **S_2**, there is equally clearly going to be a *difference*. One of the Cardinal Axioms, called "Continuity," says that, if I'm rational, then I must be indifferent between *A* and some *lottery* involving *B* and *S*; that is, there is some probability *p* at which I'd be indifferent about trading a sure apple for a *p* chance of getting a banana (and a *1-p* chance of staying at the status quo). Similarly, there must be a lottery involving *W* and *S* at some probability *p'* such that I'd be indifferent between that lottery and getting a sure apple. And, surely, there will be a big difference between the respective probabilities, *p* and *p'*. I'd require pretty high odds of getting that banana before I'd be willing to gamble my sure apple (I do, after all, want to ensure that I get *some* piece of fruit); but I'd be willing to give up my sure apple for even a *miniscule* chance of achieving world peace.

The problem with weak operationalism is that it does not allow us to understand my betting behavior, the behavior that is used to construct my utility function and that is (according to weak operationalism) rationally constrained by the Cardinal Axioms, as anything other than *mere* behavior.[14] To put some numbers on it, let's say I'd gamble my apple for a 75% chance of getting a banana, and that I'd gamble that same apple for a 0.00001% chance of world peace. On the weak operationalist's view, we know that this is my behavior, and we know *something* about what's "in my head": we know that I prefer *W* to *A* to *S*, and that I prefer *B* to *A* to *S* (where my "preferences" are real mental states). But we cannot *explain*, we cannot *rationally make sense of*, my behavior, given only these tools. That is, we cannot point to anything in the "catalogue" of real mental states that weak operationalism provides us, which is limited to ordinal preferences, that could *distinguish* between my $\mathbf{S_1}$ and $\mathbf{S_2}$ preferences. The $\mathbf{S_1}$ outcomes and the $\mathbf{S_2}$ outcomes are exactly the same so far as my real-mental-state preferences are concerned, according to weak operationalism; so we are left with no tools for explaining why $p = 0.75$ while $p' = .0000001$.

There is, to be sure, an obvious intuitive explanation of what's going on with my preferences. I require a high probability for the lottery that gives me a shot at *B* because, even though I prefer *B* to *A*, I don't prefer it by *very much*. On the other hand, my preference for *W* over *A* is not like this: I prefer *W* to *A* by *quite a lot*. And because of this difference, I'm willing to take a bigger chance of not getting anything at all in order to have a shot at *W*. The obvious, intuitive idea here is the following: my betting

[14] Or perhaps it would be better to say that weak operationalism can only see the behavior as driven by *brute* preferences: weak operationalism does, after all, allow that betting behavior, like all choice behavior, is driven by preferences. But the problem is that weak operationalism does not give us a way of understanding why *betting behavior*, specifically, would be subject to any special rational constraints. Betting preferences themselves would be like behavior on the strong operationalist view Sen attacked: applying special consistency requirements to them would seem unmotivated. This issue will, I hope, become clearer below.

behavior, my preferences and "indifference points" over the different lotteries, are driven by the different *strengths* of my preferences for the outcomes involved. But weak operationalism gives us no way to mark this difference, *the difference that explains my betting behavior*, until we have derived a utility function *from that very behavior*.

The solution, it seems to me, is simply to accept that we *do* have a pre-theoretic notion of "utility" in the substantive sense; there is some degree of strength to our preferences, some amount of desire we have for various outcomes or properties, some quantity of good we perceive various outcomes to have. Some such notion is real, and it is at work in driving our behavior, in structuring our preferences. The question of whether decision theory is really a theory of practical rationality, then, comes down to this: Is our *substantive*, pre-theoretic notion of utility helpfully "systematized" by the mathematical quantity we get out of the theory? And are the axioms themselves, on a "substantive" understanding of utility, really defensible as requirements of rationality? These are the questions we should be looking at in evaluating decision theory's place within our philosophical views.

3. A Closer Look at the Continuity Axiom

In the previous section, I suggested that weak operationalism doesn't have the tools to make a subject's betting behavior (the behavior needed to generate the utility function) intelligible as rationally assessable—as subject to rational constraint. But this claim is a bit vague; it isn't yet entirely clear just how this poses a problem for the operational view of utility. In this section, I want to raise a more specific worry about operationalism: I will argue that operationalism can't provide a defense of the Continuity Axiom as a rational constraint.

Showing that this is a problem will be a bit delicate. Surprisingly, there are very few explicit defenses of decision theory's axioms in the literature. The focus is invariably on the mathematical project of showing that a utility function (unique up to positive linear transformation) can be derived from the axioms, using a suitably rich set of "revealed preferences." The lack of justification for the axioms themselves might be fairly innocuous in most cases; the "Better Prizes" axiom, for instance, says something intuitive enough that it might be thought not to need further justification.[15] But Continuity is not obviously a rational requirement; indeed, it is very hard even to say just what it is requiring, at an intuitive level. "Continuity" is a rather abstract mathematical notion, used to describe a feature of a function that can be understood easily in graphical representation, but which requires some high-level mathematical notions to spell out precisely.

Luce and Raiffa make the most serious attempt to explain and justify Continuity at an intuitive level:

> Suppose that our subject prefers alternative *A* to *B*, *B* to *C*, and *A* to *C*…. Suppose we ask his preference between (i) obtaining *B* for certain, and (ii) a gamble with *A* or *C* as the outcome, where the probability that it is *A* is *p* and the probability that it is *C* is *1-p*. We refer to these as the "certain option" and the "lottery option." It seems plausible that if *p* is sufficiently near to 1, so that the outcome of the lottery option is very likely to be *A*, the lottery will be preferred. But, if *p* is near 0, then the certain option will be preferred. As *p* changes continuously from 1 to 0 the preference for the lottery must change into preference for the certain option. We idealize this preference pattern by supposing that there is one and only one point of change and that at this point the two options are indifferent.[16]

Luce and Raiffa can be seen as describing the two "roles" that Continuity is typically thought to play in decision theory. First, Continuity rules out an agent's assigning "infinite utility" to any one outcome; that is, it requires that there be *some* odds at which an agent will, given a set of three outcomes, trade a certain middle outcome for a gamble between the most- and least-preferred outcomes. We may call this role of Continuity the *No Infinite Utilities* (NIU) role. Luce and Raiffa suggest that NIU is plausible because we can make *p* arbitrarily close to 1, so that the most-preferred outcome is essentially

[15] Better Prizes says, in Resnik's refreshingly direct terms: You should see to it that, "other things being equal, [you] prefer one lottery to another just in case the former involves better prizes"; or, more formally: "For any lotteries *x*, *y*, *z* and any number *a* (0≤ *a* ≤1), *xPy* if and only if *L*(*a*, *z*, *x*) *P L*(*a*, *x*, *z*)," where "*P*" means "is weakly preferred to" (Resnik (1987) 91-92).

[16] Luce and Raiffa (1957), p. 22. Dreier (1996) also provides a kind of defense of Continuity, in the form of a reply to a specific objection to the Axiom (pp. 255-260). The objection he considers does not strike me as a particularly strong one, however, and his reply to it does not provide a general defense of the axiom itself.

guaranteed: in that scenario, it seems likely that the agent would take the all-but-certain best outcome over the certain middle outcome.

Continuity's second role is generally thought of more abstractly. We might call it the *No Jumps* (NJ) role. It ensures that the "graph" of an agent's preferences has no "gaps": that is, if we chart an agent's preference between the "certain option" and the "lottery option," while steadily increasing the favorability of the odds, there will be no "sudden leaps" in the curve. Luce and Raiffa describe this requirement as an "idealization" whereby we assume that the agent's preference will "switch over" at a single point (a specific probability *p*), and that, at that point, the subject will be indifferent between the lottery option and the certain option (rather than having a preference for either one over the other).

It is important to note at this stage that, if decision theory is to provide a theory of practical *rationality* (in the sense I've been considering), the above claims about Continuity will have to be given a *normative*, rather than a merely *descriptive*, reading. Luce and Raiffa defend Continuity by noting that it seems plausible their subject *will* have preferences fitting its structure. But, strictly speaking, in order to provide an intuitive defense of Continuity as a rational requirement, what Luce and Raiffa would need to show is that it's plausible that the subject's preferences *should* match the structure described by Continuity. It's not immediately obvious that we *couldn't* just change the relevant descriptive terms to normative ones in Luce and Raiffa's explanation of Continuity, while maintaining the plausibility of the axiom. I just want to flag here that it is this latter, normative reading that the operationalist needs to defend, if she wants to defend decision theory as a theory of practical rationality.[17]

I want to present a possible objection to the intuitive defense of Continuity sketched above. Before doing so, I want to pause to consider the shape of the dialectic. My aim is to show that an operationalist defense of decision theory as a theory of practical rationality faces particular challenges.

[17] For the most part, Luce and Raiffa themselves present the axioms in purely descriptive terms; but they do occasionally suggest that their version of the theory can be given a normative interpretation (one on which the axioms are taken to be "rules of consistency" that serve as "a guide to consistent action"). See, e.g., p. 32.

So, in considering the intuitive defense of Continuity, my concern will not be to decide whether objections to the Axiom are decisive against decision theory itself, or if Continuity is plausible as a rational requirement. Instead, my focus will be on whether the defense of Continuity as a rational requirement can be made *without reference to strengths of preferences*. I will not be trying to reach an absolute judgment about possible arguments for Continuity. I will just be assessing whether they implicitly rely on the notion of *strengths of preferences*, a notion that is not available to the operationalist because she claims that the notion is empty (or, that it is an outgrowth of, rather than a foundation for, the axioms).

The objection I want to consider concerns the NIU aspect of Continuity, which says, essentially, that a rational subject cannot value any given outcome to an infinite degree, either positive or negative. An example of the kind of preference pattern NIU is supposed to rule out is that of a subject who thinks, "There are no odds at which I'd bet my life for a quarter." We can suppose that this subject, S, has the following preference ordering:

S_3: gaining a quarter (*A*) > status quo (*B*) > death (*C*)

But, according to the thought expressed above, there is no lottery [*p*, *A*, *C*], no matter how close to 1 *p* is, such that S would rather have the lottery than *B*.[18] That is, it is not true for S that "if *p* is sufficiently near to 1, so that the outcome of the lottery option is very likely to be *A*, the lottery will be preferred." S violates Continuity.[19]

[18] A note on notation: I will write "[*p*, *A*, *B*]" to denote a lottery with possible outcomes *A* and *B*, where the probability of getting *A* is *p*, and the probability of getting *B* is *1-p*.

[19] It might seem strange that I am focusing on the NIU aspect of Continuity, rather than NJ. After all, NIU is really just a special case of NJ: it says that there are no "jumps" specifically *at the end points* of the graph. So, in investigating Continuity, it might seem to make more sense to look at the more general principle, NJ, rather than NIU. I offer two justifications for looking at NIU rather than NJ. First is the nature of the project I'm engaged in. I'm considering whether Continuity can be defended as a rational requirement at an intuitive level; that is, I'm asking whether reasons can be given in favor of it from *outside* the mathematical formalism (and whether those justifications bring in substantive notions of utility). Since, as noted above, the NJ aspect of Continuity is itself very abstract, it seems as though it would be very hard to offer intuitive considerations for (or against) NJ. NIU, on the other hand, is usually presented in a more intuitive way; and, indeed, the objections made against Continuity tend to be focused on the NIU aspect. So, in a sense, NIU is the only place where we *can* investigate whether Continuity

The objection is that, in spite of the fact that S violates Continuity, she does *not* seem to be irrational. And if S can violate Continuity without being irrational, conforming to Continuity can't be a rational requirement.[20]

Responses to this kind of objection tend to note that the subject in question, despite what she says, often *does* choose to bet her life for a quarter. For example, there is some non-zero (though exceedingly small) chance that a given quarter has a deadly virus on it, and that you will die if you pick it up. So if S has ever picked up a quarter, she has demonstrated that she does *not* in fact violate Continuity.

There does not seem to be any particular reference to strengths of preferences in this response, but it should be noted that the response is couched in descriptive, rather than normative, terms. The normative version of decision theory we are assessing says that Continuity is a rational requirement, so a response to the objection needs to show not just that the subject described does in fact conform to the Axiom, but that an imagined subject who doesn't (one who says "I'd never bet my life for a quarter" and *means* it) is thereby violating a norm.

I suspect that a reply to the objection *can* be given in normative terms. But again, that is not the relevant question. The question is whether the reply would make implicit use of the idea of strengths of preferences. I want to suggest that any good reply to the objection would indeed have to do so. My

is defensible as a rational requirement. My second justification for focusing on NIU rather than NJ is that, combined with a few other axioms of decision theory, NIU can actually be used to *derive* NJ. I present an informal proof for this claim in the Appendix.

[20] This objection, though widely discussed in the literature, might be most strongly associated with John Searle (see his *Rationality in Action* (2001). Interestingly, Dreier (1996) and Luce and Raiffa (1957) both consider this kind of objection, and they allow that it might *not* be an irrational preference pattern to have. They both go on to cite Hausner's "Multidimensional Utilities" (1954), and they suggest that allowing for this exception to Continuity does not undermine decision theory as a whole because there are alternate versions of the theory (like Hausner's) that can accommodate such preference patterns. I won't evaluate that particular move here, but it seems to me that it could, at best, only push the problem one step further down the road. The worry is that Continuity stands without a justification on an operational view, and that objections to it cannot be adequately addressed. My feeling is that whatever move is made to fix this problem will either (1) bring in a substantive notion of utility (which might be the case with Hausner's theory); or (2) fail to provide an adequate framework for justifying whatever axioms stand in for Continuity in the modified theory.

claim is not so much that the reply itself would have to make use of the notion of strengths of preferences. Rather, it is that in order to even understand the force of the objection in the first place, some intuitive notion of strengths of preferences has to be invoked.

The objector's imagined scenario is one in which S refuses to bet her life for a quarter, at any odds. It is important to note that, on the "only ordinal preferences" picture that the weak operationalist is working with, there is no real difference between S's $\mathbf{S_3}$ preferences (the ones involving death and a quarter), and this set:

$\mathbf{S_4}$: gaining a quarter (*A*) > status quo (*B*) > loss of a quarter (C')

Here's why this is a worry for the operationalist's account. A good defense of the norm is going to need to account for the *plausibility* of the suggestion that S is not irrational. And, it seems to me, in order to do that, the defense will have to acknowledge the difference between S's $\mathbf{S_3}$ and $\mathbf{S_4}$ preferences.

To see why, consider how an account that *does* admit that there is a meaningful pre-theoretic notion of "strength of preference" can respond to the objection. It can note that, though they are structurally similar, S's $\mathbf{S_3}$ and $\mathbf{S_4}$ preferences are also substantially different. S's *B* > *C* preference is a much *stronger* preference than her *B* > *C'* preference. S prefers the status quo to losing a quarter, but only *very mildly*; on the other hand, she prefers the status quo to death *very strongly*. It is the overwhelming *strength* of S's preference for life over death that makes it seem plausibly rational for her to be unwilling to bet her life for a quarter, even though it might seem *obviously* irrational for her to be unwilling to bet *a quarter* for a quarter, no matter the odds. But, the response goes, S must bet her life for small benefits every day; so long as the odds are miniscule enough, it is not only rational, but rationally *required* for her to make those bets. The strength of her preference for life over death, even though it really is quite large, doesn't license her to refuse all bets with death as an outcome; doing so would mean a life of paralysis, and that would be an (intuitively) irrational way to live. The strength of S's preference for life over death simply needs to be "counterbalanced" by the exceedingly small odds of actually dying in performing a given action (like picking up a quarter). The fact that we are thinking about S's preferences in *cardinal* terms, as having *strengths*, is what allows us to argue that any strength of preference can be "counterbalanced" by a suitably small value for *p*.

The problem for the operationalist is that, in failing to acknowledge a notion of strength of preference, she cuts herself off from two crucial elements of the above response. First, the operationalist will be unable even to distinguish the cases where NIU is obviously an intuitive requirement of rationality (those cases involving relatively "mild" preferences, such as the preference for a quarter over the status quo) from the cases where it takes a genuine argument to show that NIU should be regarded as a rational requirement (cases involving "very strong" preferences, such as the preference for life over death). The objector who thinks S is being perfectly rational (in the case involving death) can rightly point out that the operationalist is not giving her objection a charitable reading; indeed, the objector claims, the operationalist's picture simply rules out (by fiat) that there could be any sense to her objection. According to operationalism, there is no sense in which refusing to bet one's life for a quarter is more rational (or less obviously irrational) than refusing to bet a quarter for a quarter. And, without the ability to mark that difference, operationalism can't provide a charitable interpretation of the objection, or an adequate defense of Continuity. The second problem for the operationalist involves the content of the response itself: the operationalist can't appeal to the idea of "counterbalancing" a very strong preference with a suitably miniscule value of *p* because she can't appeal to strengths of preferences *at all*.[21]

[21] The situation is actually even worse than I've been suggesting for the operationalist. If S refuses to bet her life against a quarter at any odds, she will be violating the Continuity Axiom. But, on the operational view, if a subject violates one of the axioms, she is not "describable" by a utility function, which means we can't assign utilities to outcomes for her at all. So any hope of explaining the (seemingly rather normal) preference pattern displayed by S is lost: we can't even "back out" a post-hoc "strength of preference" measure for S (i.e., a utility function) that will distinguish between her $\mathbf{S_3}$ and $\mathbf{S_4}$ preferences. We are simply left with the brute fact that she will take the bet when it's a quarter against a quarter, but not when it's her life against a quarter—even though the only real mental states we're attributing to her, the ordinal preferences, don't distinguish between those scenarios at all, and even though we have no hope of "constructing" a representation of the nature of the difference in those preferences.

4. “Simple” Preferences and Humeanism

The difficulties I outlined for weak operationalism in Sections 2 and 3 stem from two features of the view that, it seems to me, sit uneasily with each other. The first feature is the claim that an agent has ordinal preferences over probabilistic lotteries, and that these “gambling preferences” are subject to certain rational constraints. The second feature is the claim that an agent does *not* have cardinal preferences, that her real mental states are limited to a preference for *A* over *B*, and do not include such things as “a very strong preference for *A* over *B*” or “a mild preference for *C* over *D*.”

The tension I see in these two claims is that the existence of preferences over gambles seems, intuitively, to *stem from* the strength of the agent’s preferences over simple outcomes. That is, it seems wildly implausible that an agent *just has* some particular set of preferences over probabilistic outcomes. If you offer to sell me a $10 lottery ticket with a 9% chance of getting me $100, I might deliberate for a minute and decide that I prefer to keep my $10. If you offer me another $10 lottery ticket that has a 9% chance at creating world peace, my preference will surely be different. But there is equally surely a *reason* for this difference; namely, the much greater *strength* of my preference for world peace (over the status quo) compared to my preference for $100. My preferences about numerically-specified lotteries don’t just come from nowhere; and, if it somehow turned out that they *did*, it seems mysterious why practical rationality should be thought to consist in conforming those “out of nowhere” lottery-preferences to a particular set of patterns.

A motivating argument for this “out of nowhere” view of lottery-preferences is sometimes given along Humean lines. For Hume, preferences are “original existences.” They are “simple,” not composed out of other mental states, and (crucially) they are not subject to rational assessment. But if preferences over lotteries stem from the strength of preferences over simple outcomes, in the way I’ve suggested, then they *are* subject to rational constraint: they are the product of *reasoning* about how strongly we prefer *A* to *B*, and how much the odds of the lottery make us "discount" the strength of this preference. And this picture violates the Humean insight that preferences are "simple."

The appeal of a Humean decision theory is that it lets us construct norms for making complex decisions (for coming to have preferences in complex situations) from some basic rules and some simple preferences.[22] The idea that the simple preferences at the heart of this story aren't subject to rational assessment is crucial to the "Humean-ness" of the Humean account. We can't call Hume's imagined preference for the destruction of the whole world to the scratching of his finger *irrational* because it is *basic*. It is not arrived at through reasoning, so questions of rationality are simply misplaced.

This Humean picture seems to have something to it, intuitively. We find ourselves "saddled with" our basic desires; we don't choose them or reason our way to them. The weak operationalist insists that we build a theory without "looking under the hood" of these basic Humean desires, the "original existences" that the view associates with *simple* preferences. And, intuitively, Hume's "world destruction (*D*) > scratching of finger (*F*)" preference might be such a simple preference. But Hume's preference is simple in *two* senses. It is an *ordinal*, rather than a cardinal, preference (it doesn't say "how much" Hume prefers *D* over *F*); but it is also a preference over simple, rather than probabilistic, *outcomes*. Weak operationalism latches onto the first sense of simplicity, claiming that only ordinal preferences are really "original existences," and that the cardinal preferences we generate out of decision theory are not. In focusing on just the first sense in which Hume's imagined preference is "simple," however, weak operationalism ignores the possibility that there is anything significant about "simplicity" in the second sense.

Weak operationalism lumps Hume's ordinal preference for the simple *outcome D* over *F* into the same category as his ordinal preferences over *probabilistic* outcomes. Consider the following lotteries:

[22] Luce and Raiffa are quite explicit about the appeal of this kind of picture (p. 32); Dreier repeatedly invokes the Humean notion of simple preferences as "original existences," and his picture is likewise one on which "complex" (i.e., cardinal) preferences are built out of "simple" (i.e., ordinal) preferences, using the axioms as guides.

L_1: [0.23, an apple, a pomegranate]
L_2: [0.88, $10, catching the flu]

Suppose Hume prefers L_1 to L_2. On the weak operationalist picture, both of these ordinal preferences, the one with simple outcomes and the one with probabilistic outcomes, get to be "original existences," not subject to rational constraint. And, in the case of the L_1-over-L_2 preference, that seems wrong.

Here is the idea I'm getting at. The operationalist's Humeanism feels intuitive, in part, because it talks about constructing rationality constraints for complex choices out of two elements: (1) basic axioms of consistency; and (2) basic "original existence" simple preferences. But this tends to obscure the fact that some of the "original existences" on the operationalist picture, the preferences involving *probabilistic lotteries*, are not really so simple.

Dreier describes the process of deriving a utility function from ordinal lottery-preferences with the following example. You prefer an orange to an apple. But, when asked, you just don't know "to what degree" or "with what relative strength." So, Dreier imagines, we could ask how you feel about various probabilistic lotteries involving the fruits. And, he suggests, we can "suppose you find that you are indifferent between" a 25% chance of getting an orange and a sure thing apple. He then shows how we could use this information to *generate* a utility function, and a "constructed strength of preference scale," for you. But this story seems a bit odd, by my lights. What does it mean to say that I could "find" that I am indifferent between a specific probability of getting an orange and a sure thing apple? Does it mean that, when I am confronted with such a situation, I find my arm hovering in the middle of the two fruits (the apple sitting out in the open, the orange hidden behind one—I know not which—of four doors)?

Dreier addresses, in another context, the epistemic question of how we can know our own preferences; he admits that it may require "abstraction and careful consideration." In the case of the apple and the orange, though, he suggests we just "find" our indifference point. But the "careful consideration" model seems more apt: we have to think about the probabilities involved and our feelings about the different fruits in deciding on an indifference point. And, crucially, it seems that what we are "carefully considering" is, in part, something along the lines of the strength of our preferences: I need to figure out *how much* I want an orange, and *how much more* I want an apple. But these notions are equivalent, or at least very closely related, to my *cardinal* preferences for each fruit over the status quo. Strengths of preferences factor crucially in the *generation* of the ordinal preferences over probabilistic outcomes that weak operationalism takes to be basic. So the claim that such ordinal preferences are logically prior to the notion of the strength of a preference looks untenable.

To see just how bad things get for Dreier's view, consider what he says about *impractical preferences*: preferences involving choices that we couldn't possibly be faced with. In responding to a challenge to operationalism from Broome, Dreier claims that the list of "simple" preferences at decision theory's foundation must include impractical preferences: that is, we must have (real) preferences over outcomes that couldn't possibly be presented together in an actual choice.[23]

So consider the following outcomes (one probabilistic, one impractical):

L_3: [0.7415, a pomegranate, an apple]
L_4: an orange when the other choice is a banana

Suppose Sally prefers L_3 to L_4. On Dreier's view, Sally's preference is, like Hume's $D > F$ preference, not subject to rational constraints; it is an "original existence." But, clearly, Sally has to put a lot of "work" into figuring out her preference. And intuitively, part of that work involves Sally's figuring out how *strongly* she prefers pomegranates to oranges, and how much the probabilities involved "discount" the strength of that preference (and what the heck the "when the other choice is a banana" aspect of the choice is doing). We have moved a good distance from Hume's $D > F$ preference; any intuitions about

[23] Dreier (1996), pp. 260-270. Broome introduced the idea of impractical preferences in an earlier version of his (2002) Chapter 5. There, he argues that an operational view *needs* to include real impractical preferences among its basic elements, in order to avoid becoming completely vacuous (vacuous because it would license *any* set of preferences as rational). I won't go into Broome's argument for this claim here; but the argument seems sound to me, and Dreier, in defending operationalism, accepts Broome's conclusion.

“simple” preferences’ not being subject to rational constraints (because they are original existences, because they are not arrived at through reasoning) have been left far behind.

Conclusion

I have argued that on the dominant, *operational* understanding of utility, decision theory cannot be defended as a theory of practical rationality. The axioms at its foundation need to be viewed through the lens of a *substantive* notion of utility if they are to serve as rational constraints on our preferences.

I have also suggested that the operational view cannot really do justice to the Humean insight that is supposed to motivate it. Some of the “simple” (that is, *ordinal*) preferences at the base of an operational version of decision theory bear little resemblance to Hume’s “original existences,” the basic preferences that he thought were outside the scope of rational assessment. A better way to flesh out Hume’s idea about “original existences” might be to say that preferences for simple *outcomes* are the basic ones, while allowing that these preferences can include *strengths* (where the strengths would also not be subject to rational assessment). Or perhaps we should think of the original existences as preferences over (or desires for) simple *properties* of outcomes. But, however we spell out the details, it seems much more plausible to invoke Hume’s “desires are original existences” mantra in support of a theory that says we have some basic, not-arrived-at-through-reasoning, *strengths* to our preferences for simple outcomes than it does to invoke that same mantra in support of a theory that regards complex, impractical preferences over lotteries as basic inputs. Operationalism has become the dominant view among decision theorists in part by waving a Humean banner. But if, going against the prevailing orthodoxy, we build decision theory on the foundation of a *substantive* understanding of utility—one that provides the tools for defending the Cardinal Axioms as rational constraints, and one that is more plausible in its own right—we might just produce a more truly Humean theory of practical rationality.

Appendix: Deriving *No Jumps* from *No Infinite Utilities*

In Section 3, I described two "roles" that the Continuity Axiom is generally thought to play in decision theory. The first, which I labeled *No Infinite Utilities* (NIU), says the following:

Suppose S has the following preference ordering: $A > B > C$. Call the option of getting B for certain "the certain option" (or simply "B") and a lottery of the form [p, A, C] "the lottery option" (or "L"). Then there must be some p close to 0 such that $B > L$, and some p close to 1 such that $L > B$.

The second role of Continuity is *No Jumps* (NJ), which says:

There will be a single value of p such that $B \approx L$ (where "$x \approx y$" denotes "S is indifferent between x and y").

Claim: NJ can be derived from NIU, if we also assume the following commonly-accepted axioms of decision theory:[24]

<u>Ordering (O)</u>: S has a complete and transitive preference ordering.
<u>Better Chances (BC)</u>: For any lotteries x and y and any numbers a and b (both between 0 and 1, inclusively), if $x > y$, then $a > b$ iff $[a, x, y] > [b, x, y]$.[25]
<u>Reduction of Compound Lotteries (RCL)</u>: For any lotteries x and y and any numbers *a, b*, c, d (all between 0 and 1, inclusively), if $d = a*b + (1 – a)*c$, then $[a, [b, x, y], [c, x, y]] \approx [d, x, y]$.

What follows is an (extremely) informal proof of this claim.

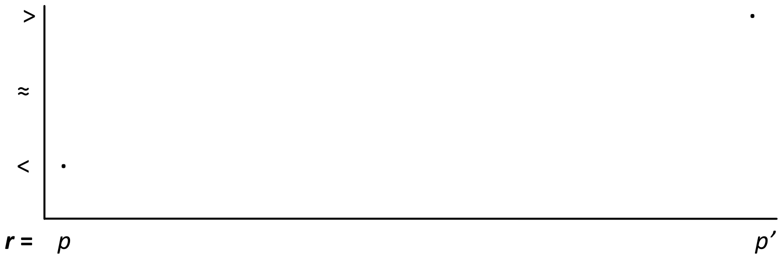
Suppose S has preferences that obey NIU, O, and BC. We want to show that S must also obey NJ. That is, for an arbitrary set of outcomes A, B, and C, where $A > B > C$, there will be a single value of p such that $B \approx [p, A, C]$. So we need to establish the *existence* and *uniqueness* of an "indifference point" (where S is indifferent between the certain and lottery options).

First, uniqueness: Suppose that $B \approx [p, A, C]$ for some p. Now assume, for *reductio*, that $B \approx [p', A, C]$ for some p', where $p' \neq p$. Then, by transitivity (O), we know that $[p, A, C] \approx [p', A, C]$. Since we also know that $p' \neq p$, we know that either $p > p'$ or $p' > p$. So, by BC, we know that either $[p, A, C] > [p', A, C]$, or $[p', A, C] > [p, A, C]$. But then we know (by (O)) that it is not the case that $[p, A, C] \approx [p', A, C]$. We have reached a contradiction: S both is and is not indifferent between $[p, A, C]$ and $[p', A, C]$. Thus, we can reject the assumption that there could be more than one value of p such that $B \approx [p, A, C]$.

[24] Adapted from Resnik (1987), pp. 22, 90-92.

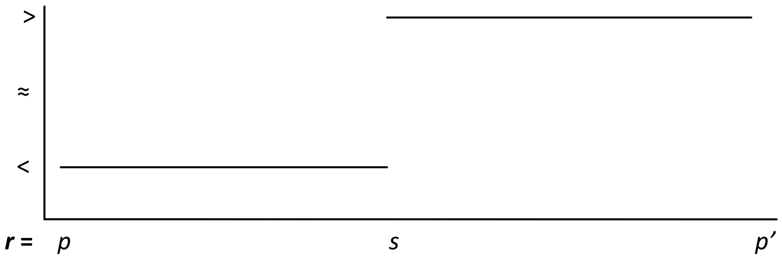
[25] A brief note on my formulation of BC: Resnik spells out this axiom in terms of lotteries over lotteries; that is, he allows that the outcomes themselves might be further lotteries. This is fairly standard, and just about all decision theories will, at least in their developed forms, endorse such a framework. But some ways of spelling out the axioms, such as that found in Luce and Raiffa (1957), originally develop the axioms in terms of lotteries over simple outcomes only. I don't think any special worries arise from going with the more permissive version found in Resnik, but, since I will be relying on the possibility of including lottery outcomes within lotteries in the proof that follows, I wanted to flag the point here.

Now, existence: We begin by noting that, by NIU, there is some p close to 0 such that $B > [p, A, C]$, and some p' close to 1 such that $[p', A, C] > B$. By completeness (O), we know that, for any r between p and p' (inclusive), either $B > [r, A, C]$ or $[r, A, C] > B$ or $B \approx [r, A, C]$. That is, S's preference between B and $[r, A, C]$ is everywhere defined. We will model this fact with a graphic depiction of S's preference for the certain and lottery options as r increases from p to p'. Along the horizontal axis, we plot the values of r; along the vertical axis, we plot the three possible relations between the lottery option and B. At the far left of the graph, $r = p$, and we know that, at this point, S prefers B to the lottery. We depict this by plotting a point over that value of r parallel to the "<" symbol (which indicates that the lottery with this value of r is dispreferred to B). Likewise, at the far right of the graph, we plot a point parallel to the ">" symbol, in order to indicate that, for that value of r (namely, $r = p'$), S prefers the lottery option to B. Here is the graph so far:



We now imagine filling in the rest of the graph for the intermediate values of r. We know that we'll have an output for every value of r (by completeness). So, take some value of r such that $[r, A, C] < B$ (we know that we'll have at least one such value, namely p). By BC, we know that, for any $r' < r$, $[r', A, C] < [r, A, C]$. And, by transitivity, we also know that $[r', A, C] < B$. Graphically, this means that, for any point plotted on the bottom "<" line, every point to the left of it will also be plotted on that line. Similarly, we can use BC and transitivity to conclude that, given any value of r such that $[r, A, C] > B$, and any $r' > r$, $[r', A, C] > B$. Graphically, this will mean that, for any point plotted on the top ">" line, every point to the right of it will also be plotted on that line.

Now, suppose that there is some value of r is such that $B \approx [r, A, C]$. In that case, we have established the existence of an indifference point, and our proof is complete. So suppose instead (for *reductio*) that there is no such r. Then we know that the graph will have a point above each value of r parallel either to the top ">" line or the bottom "<" line. And, by what was established in the previous paragraph, we know that, for any point on the top line, every point to the right of it will also be on the top line, while, for every point on the bottom line, every point to the left of it will also be on the bottom line. Our graph will thus look something like this:

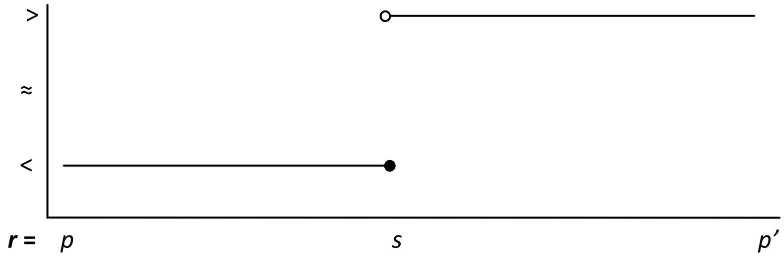


Note that there can be no "gap" between the end of the bottom line and the start of the top line; if there were, we'd violate completeness. Now, consider the value of r at the endpoint of the bottom line; call this value s. We have not yet established that s should be plotted as *part of* the bottom line (that is, we don't yet know whether

$[s, A, C] < B$). Graphically, s could be a kind of “limit,” an “open bubble,” where every point to the left of it was plotted on the bottom line, but the s point itself was not on that line. What we do know is that the preference relation must be defined for $[s, A, C]$ (by completeness). So there are three possibilities: $[s, A, C]$ could be preferred, dispreferred, or indifferent with respect to B.

We will consider the case where $[s, A, C] < B$. We will show that this scenario contradicts NIU; and we assert (without further argument) that an analogous result can be shown for the $[s, A, C] > B$ case. So the only possibility remaining will be that $B \approx [s, A, C]$; this will establish the existence claim, and complete the proof.

So, suppose (for *reductio*) that $[s, A, C] < B$ (call this lottery J). This means that it is not the case that $[s, A, C] > B$; graphically, it means that the s point should be plotted on the bottom line, and an open bubble should be plotted above s on the top line:



We also know that, for any arbitrarily small ε, $[s+\varepsilon, A, C] > B$; if that were not the case, then, contrary to our definition of s, s would not be the endpoint of the bottom “<” line. We can also see this graphically by noting the open bubble at the left end of the top “>” line; the bubble indicates that s is the *limit* for the “>” line.

Now, consider the following set of outcomes: A, B, and J. We know that $A > B > J$ (by the supposition of the previous paragraph, along with the fact that $A > B > C$, and transitivity). Consider S’s preferences over B and the following lottery (call it I): $[p, A, J]$. By NIU, we know that, for some value of p close to 0, S will prefer B to I (B is the certain option, I the lottery option). But I is just the compound lottery $[p, A, [s, A, C]]$: it gives S a p chance of getting A, and a $1\text{-}p$ chance of getting a lottery with an s chance of getting A and a $1\text{-}s$ chance of getting C. This compound lottery has two final outcomes, A and C; and the overall chance of getting A is greater than s. So that means, when S prefers B to I, she is preferring B to a compound lottery that is equivalent, by RCL, to a lottery of the form $[s+\varepsilon, A, C]$. And that contradicts the fact, mentioned above, that for any arbitrarily small ε, $[s+\varepsilon, A, C] > B$.

Spelling out this line of thought more formally, we can apply RCL as follows:

Let $t = p*1 + (1 - p)*s$. Then, by RCL: $[p, [1, A, C], [s, A, C]] \approx [t, A, C]$.

We now simplify:

$$\begin{aligned} t &= p*1 + (1 - p)*s \\ &= p + s - p*s \\ &= s + (p - p*s) \\ &= s + p*(1 - s) \end{aligned}$$

And we substitute: $[p, [1, A, C], [s, A, C]] \approx [s + p*(1\text{-}s), A, C]$
We also know that $s < 1$, and $p > 0$; this implies that $(s + p*(1\text{-}s)) > s$.

So we can rewrite:

$[p, [1, A, C], [s, A, C]] \approx [s + p*(1-s), A, C]$

as

$[p, A, J] \approx [s+\varepsilon, A, C]$ (i.e., $I \approx [s+\varepsilon, A, C]$ for some value of ε)

But we know that $B > I$. So, by transitivity, we know that $B > [s+\varepsilon, A, C]$ for some value of ε. This contradicts the fact that, for any arbitrarily small ε, $[s+\varepsilon, A, C] > B$. So the supposition that $[s, A, C] < B$ must be rejected. An analogous proof can be constructed to show that it cannot be the case that $[s, A, C] > B$. Thus, it must be the case that $[s, A, C] \approx B$. We have now established the existence of an indifference point, and the proof is complete. ■

References

Bermúdez, José Luis. *Decision Theory and Rationality*. Oxford: Oxford University Press, 2009.

Broome, John. *Ethics out of Economics*. Cambridge: Cambridge University Press, 1999.

Dreier, James. "Rational Preference: Decision Theory as a Theory of Practical Rationality." In *Theory and Decision*, 40: pp. 249-276 (1996).

Hume, David. *A Treatise of Human Nature* (1739). Book II, Part III, Sect. III: "Of the influencing motives of the will." L. A. Selby-Bigge (ed.). Oxford: Clarendon Press, 1978.

Luce, R. Duncan and Howard Raiffa. *Games and Decisions: Introduction and Critical Survey*. New York: John Wiley & Sons, Inc., 1957.

Resnik, Michael D. *Choices: An Introduction to Decision Theory*. Minneapolis: University of Minnesota Press, 1987.

Sen, Amartya. "Behavior and the Concept of Preference." *Economica*, New Series (1973), 40(159): 241-259.