

## The Story of Rational Action<sup>1</sup>

---

Decision theory comprises, first, a mathematical formalization of the relations among value, belief, and preference; and second, a set of prescriptions for rational preference. Both aspects of the theory are embodied in a single mathematical proof. The problem in the foundations of decision theory is to explain how elements of one and the same proof can serve both functions.

I hope to solve this problem in a way that anchors the decision-theoretic norms of rational preference in fundamental intuitions about rationality in general. I will thus depart from the tradition of anchoring those norms in intuitions about gambling strategies or preference structures of the sort that are the special concern of the theory itself. Although my interpretation is meant to capture what is right about the decision-theoretic conception of rational preference, it will lead me to argue that there is also something fundamentally wrong about that conception. In my view, decision theory tells us how to be rational in our preferences because it tells us how to have preferences that make sense; but there are ways of making sense that outrun, and may in fact conflict with, the prescriptions of decision theory.

The mathematical proof at the heart of decision theory concerns an agent and a set of *options*, some of which are chancy, in the sense that they can yield various possible *outcomes*, depending on whether particular *contingencies* obtain. The agent is imagined to have binary *preferences* among these options—that is, preferences for one option over another in the situation that only those two options are available. What the proof demonstrates is that if the agent's binary preferences satisfy various formal requirements, then we can construct a pair of functions assigning quantities called *utilities* to the possible outcomes of his taking an option, and quantities called *probabilities* to the contingencies on which those outcomes depend, in such a way that he always prefers the option

<sup>1</sup> This chapter originally appeared in *Philosophical Topics* 21 (1993) 229–54 and is reprinted by permission of the Board of Trustees of the University of Arkansas. I am indebted to Jim Joyce for extensive comments on several drafts of this chapter; I also received valuable suggestions from John Broome, Jean Hampton, and John Devlin. Some of the material in this chapter was presented in the “Philosopher’s Holiday” lecture series at Vassar College, thanks to Stephanie Spalding, Tim Horvath, and Jennifer Church. Work on this chapter was supported by the Edna Balz Lacy Faculty Fellowship at the Institute for the Humanities, University of Michigan.

with the higher *expected utility*—that is, the higher sum of possible utilities, after each has been discounted by its probability.

This proof is called a representation theorem, because it demonstrates the possibility of representing the agent's preferences in a particular way—namely, as maximizing the actuarial product of two functions. The formal requirements that an agent's preferences must satisfy, in order to be assured of such a representation, serve as the axioms of the theory.<sup>2</sup> The axioms state that the agent's preferences satisfy conditions such as transitivity, independence, monotonicity, and the like.

The transitivity axiom states that, for any A, B, and C, if the agent prefers A to B and B to C, then he prefers A to C. The independence axiom states that if he prefers A to B, and his preference for A is unaffected by whether p is true, then he prefers the chancy option [A if p, C if not-p] to the option [B if p, C if not-p]—that is, he prefers a chance of getting A to an equal chance of getting B, provided that the “consolation prize” is the same in either case.

The monotonicity axiom states that if the agent prefers betting for the truth of p to betting against it with a given pair of payoffs, then he will prefer betting for p to betting against it with any other pair of payoffs. Here a bet for the truth of p is a chancy option in which the preferred payoff is obtained if p is true. Thus, preferring to bet for the truth of p with the payoffs A and B consists in preferring the option [A if p, B if not-p] to [B if p, A if not-p], given that A is also preferred to B. The axiom states that under these circumstances, the agent prefers [C if p, D if not-p] to [D if p, C if not-p] for any C and D where C is preferred to D.

A fourth axiom, the axiom of continuity, states that if the agent prefers A to B, and B to C, then there will be some contingency p such that the agent is indifferent between B and the gamble [A if p, C if not-p]. That is, there will be some contingency on which the agent is willing to risk losing B for the chance of winning A instead.

Even without rehearsing the proof of the representation theorem, we can see intuitively how preferences obedient to these requirements (among others, perhaps) provide the basis for constructing utility and probability functions. If the agent prefers A to B, then we infer that his expected utility for A is greater than that for B. If the agent prefers betting for p to betting against it, then we infer that his probability for p is greater than that for not-p. And suppose that the agent prefers A to B, and B to C, but is indifferent between B and a bet for p with payoffs A and C. In that case, we infer that he regards the gain he

<sup>2</sup> My summary of the theorem omits axioms that are required to guarantee the uniqueness of the utility and probability functions constructed from an agent's preferences. I have chosen to focus on the axioms that are required to guarantee the existence of such functions.

stands to realize if he gets A instead of B, and the loss he stands to suffer if he gets C instead of B, as canceling each other out in the expected value of the gamble, so that the gamble is equivalent in value to B itself. These gains and losses will cancel each other out in the gamble's expected value only if the probabilities by which they're discounted succeed in erasing any difference in size between them. We therefore infer that the proportion between the agent's gain for getting A instead of B, and his loss for getting C instead of B, is the inverse of the proportion between his probability for p and that for not-p.<sup>3</sup> By following these and similar principles, we can reason from the agent's preferences to assignments of utility and probability that represent those preferences as maximizing expected utility.

These axioms express conditions that are not only jointly sufficient (when appropriately supplemented) but also individually necessary for the agent's preferences to be representable in this fashion. Thus, for example, if the agent prefers A to B, and B to C, then we shall have to infer a higher value for A than for B, and a higher value for B than for C—in which case we shall already have inferred a higher value for A than for C. The resulting values won't represent the agent's preferences unless he obeys the requirement of transitivity, by preferring A to C. Secondly, if the agent prefers A to B, we shall have to infer a higher value for A than for B, which will cause the expected value of [A if p, C if not-p] to be higher than that of [B if p, C if not-p], since these gambles are similar in every other respect. These values won't represent the agent's preferences, then, unless he obeys the requirement of independence, by preferring the former gamble to the latter. Thirdly, if the agent prefers to bet for p with payoffs A and B, but prefers to bet against p with payoffs C and D—thereby violating the requirement of monotonicity—then there will be no probability that we can infer for p in such a way as to represent both preferences. And, finally, if the agent prefers A to B, and B to C, then we shall have to infer utilities such that the utility of A minus that of B stands in some determinate proportion to that of B minus that of C. But these utilities will represent the agent's preferences only if he obeys the requirement of continuity, by being indifferent between B and any gamble that links the payoffs A and C to a contingency whose probability is the inverse of this proportion.

The representation theorem of decision theory is irrefutable. Like any mathematical theorem, however, it is a formal structure that needs to be interpreted. And the interpretation of this particular theorem is highly problematic.

One might think that the utility and probability functions constructed in the theorem could be interpreted as representing how much value the agent places on the various possible outcomes and how much credence he places in the

<sup>3</sup> That is, we infer that  $\text{Pr}(p)/\text{Pr}(\neg p) = U(B)-U(C)/U(A)-U(B)$ .

contingencies on which those outcomes depend. In that case, they would represent what we ordinarily call the agent's values and beliefs. Yet if the utility and probability functions constructed in the theorem are to represent the agent's values and beliefs, we shall have to define the term "preference" in such a way as to denote a phenomenon from which an agent's values and beliefs can be so constructed. How, then, shall we define the term?

A preference might simply be a behavior or behavioral disposition; that is, preferring A to B might consist in taking or being disposed to take A when choosing between A and B.<sup>4</sup> Alternatively, preferring A to B might be a kind of affect, such as liking or wanting A more than B. Or this preference might be a rudimentary value judgment to the effect that A is preferable, more desirable, or simply better than B. Will any of these phenomena yield utilities and probabilities that actually represent the agent's state of mind?

Inferences about the agent's state of mind are especially questionable when preference is defined behavioristically. From an agent's taking or being disposed to take A instead of B, we cannot necessarily infer that he values A more, since his behavior might also manifest weakness of will with respect to a contrary evaluation, or might be compatible with his inability or refusal to make any comparative evaluation at all; and yet the representation theorem instructs us to assign him a higher expected value for A, if his behavior or behavioral disposition qualifies as a preference.

The mentalistic conceptions of preference are not necessarily more conducive to inferences about the agent's values and beliefs. Even an agent's rudimentary value judgments, such as regarding A as preferable to B, may not provide a basis for such inferences. For if the agent prefers A to B, the theorem instructs us to infer, not just that he places a higher value on A—something that might indeed be entailed in his regarding A as preferable—but that he places a higher *expected* value on it, a value decomposable into values placed on its possible outcomes, and probabilities placed on the relevant contingencies. Yet an agent's regarding A as preferable to B may not in fact indicate that the higher value he thereby places on A is decomposable into values and probabilities that he places on its constituents.<sup>5</sup>

This problem becomes clearest when the agent has a preference for or

<sup>4</sup> Actually, taking or being disposed to take A instead of B is more likely to constitute the disjunctive state of preferring A *or* being indifferent, since our ordinary notion of someone's being indifferent between A and B still allows room for his picking one at random so as to avoid the fate of Buridan's ass. Decision theorists who favor a behaviorist interpretation of preference are therefore inclined to state the theory in terms of the disjunctive state of preference-or-indifference. I shall henceforth ignore this complication.

<sup>5</sup> The possibility that a person's preferences might represent judgments of value but not judgments of expected value is discussed by John Broome in "Utility," *Economics and Philosophy* 7 (1991) 1–12.

against taking risks.<sup>6</sup> Consider an agent who prefers A to B, and B to C, but is indifferent between B and the gamble [A if p, C if not-p]. As we have seen, the representation theorem will instruct us to infer that the agent regards his potential gain for getting A instead of B, and his potential loss for getting C instead of B, as canceling each other out once they have been discounted by the probabilities of p and not-p, respectively. But if B is not itself a chancy option and the agent is averse to risk, then he wouldn't have accepted the chancy alternative unless his expected value for the gain was more than would be required to cancel that for the loss, since his aversion to risk leads him to avoid even fair gambles. Conversely, if the agent enjoys risky ventures, then his accepting the gamble may be consistent with his assigning the gain an expected value too small to cancel that of the loss, since the value of risk itself makes even unfair gambles worth his while. In either case, the inference mandated by the theorem—that the agent's expected values for gain and loss are equally balanced—will not result in an accurate representation of his attitudes.<sup>7</sup>

Thus, the representation theorem does not show how an agent's preferences indicate what values and beliefs he actually holds. At most, it shows how an agent's preferences indicate values and beliefs that he appears to hold, in the sense that he behaves *as if* he holds them. Even a risk-averse agent behaves as if he holds values and beliefs corresponding to the utilities and probabilities that the representation theorem constructs for him: specifically, he behaves *as* he would *if* he held those attitudes *and* he wasn't risk-averse.<sup>8</sup> These "as if" attitudes can still be attributed to the agent, provided that they are conceived as having no reality beyond the patterns of preferences from which we construct them. That is, we can think of these values and beliefs as attitudes that are emergent in the agent's preferences—as complicated ways of preferring.

By sticking with the technical vocabulary of subjective utility and probability, we can remind ourselves that the decision-theoretic constructs are not values and beliefs in the ordinary sense. And having interpreted subjective utility and probability as ways of preferring, we can construct them from preferences of any kind—behavioral dispositions, feelings, or value judgments, all of which can display the patterns in question.

<sup>6</sup> See Bengt Hansson, 'Risk Aversion as a Problem of Conjoint Measurement,' in *Decision, Probability, and Utility; Selected Readings*, ed. Peter Gärdenfors and Nils-Eric Sahlin (Cambridge: Cambridge Univ. Press, 1988), 136–58.

<sup>7</sup> Here I am considering a case in which the agent's aversion to risk does not lead him to violate the axioms. Some economists will insist that the theory can represent such cases of risk-aversity, in the form of a "concavity" in the agent's utility function. Intuitively, however, we think there is a difference between how an agent feels about various outcomes and how he feels about their depending on chancy contingencies; whereas the theory represents the latter, if at all, only in terms of the former.

<sup>8</sup> Or weak-willed, or . . .

A second challenge for the interpreter of decision theory is to find within it some norm of rational preference. Decision theory is generally assumed to involve, not just a method for representing the preferences that you actually have, but a norm prescribing which preferences or patterns of preference you ought to have. The interpretive problem is to say what the relevant norm of rational preference is and where it appears in the formal theory.

The relevant norm is often assumed to be the one that tells you to prefer the option with the higher expected utility. To be sure, the prescription to maximize your expected utility may well be the norm implicit in the theory; but in the context of formal decision theory, this norm has a force somewhat different from that which it is ordinarily thought to have.<sup>9</sup>

As ordinarily understood, the prescription to maximize your expected utility presupposes that there is some measure of expected utility that applies to you and that your preferences are therefore obliged to maximize. But in the context of decision theory, the utility and probability functions that apply to you are constructed out of your preferences, and so your expected utility is not an independent measure that your preferences can be obliged to maximize; rather, your expected utility is whatever your preferences *do* maximize, if they obey the axioms. Hence, the injunction to maximize your expected utility can at most mean that you should have preferences that can be represented as maximizing some measure (or measures) of expected utility, which will then apply to you by virtue of being maximized by your preferences.

The only cases in which this injunction yields any criticism of your preferences are those in which you violate the axioms. In such cases, the theory implies that your preferences do not maximize your expected utility. But this criticism doesn't mean that there is some antecedent measure of your expected utility that your preferences fail to maximize: according to the fundamental theorem, your violating the axioms entails that no measure of expected utility can be assigned to you. The only criticism that can be directed at your preferences when they violate the axioms is precisely that there is no measure of expected utility that can be assigned to you, because there is none that they maximize. And no matter how you bring your preferences into conformity with the axioms, you will thereby silence this criticism, since you will maximize some measure (or measures) of expected utility, which will consequently turn out to be yours.

In the context of traditional decision theory, then, the injunction "Maximize your expected utility" means no more than "Obey the axioms, and you

<sup>9</sup> Several of the points made in this and the following sections are summarized by John Broome in 'Should a Rational Agent Maximize Expected Utility?', in *The Limits of Rationality*, ed. Karen Schweers Cook and Margaret Levi (Chicago: Univ. of Chicago Press, 1990), 134.

will have maximized any measure of expected utility that might be yours.” Put more simply, the injunction says, “Obey the axioms, and expected utility will take care of itself.”

The problem with the latter injunctions is that they lack the intuitive appeal of the prescription to maximize expected utility, as that prescription is ordinarily understood. The prescription to maximize expected utility has intuitive appeal if interpreted as presupposing that you already place different values on different outcomes, and that you therefore have reason to prefer those things which are most likely to promote whatever you already value most. But in the context of decision theory, the values that you place on outcomes are “as if” values constructed out of your preferences; and so those values aren’t antecedently available to generate reasons for having preferences. To be sure, the theory guarantees that if your preferences conform to the axioms, then you will systematically prefer options that promote what you value, in some sense of the word; but you will systematically prefer options that promote what you value, in this sense, only because you will turn out, by definition, to value whatever your preferred options systematically promote. And the question is why you ought to prefer things that promote what you value in this *post facto* sense.

Exponents of decision theory claim that they needn’t derive normative force from the notion of maximizing utility, because the axioms have normative force of their own. Thus, Savage writes:<sup>10</sup>

[W]hen it is explicitly brought to my attention that I have shown a preference for f as compared with g, for g as compared with h, and for h as compared with f, I feel uncomfortable in much the same way that I do when it is brought to my attention that some of my beliefs are logically contradictory. Whenever I examine such a triple of preferences on my own part, I find that it is not at all difficult to reverse one of them. In fact, I find on contemplating the three alleged preferences side by side that at least one among them is not a preference at all, at any rate not any more.

Here Savage is saying that the transitivity requirement has a normative force for him in its own right, as a requirement that he feels obliged to obey in his preferences, just as he feels obliged to obey the requirements of consistency in his beliefs.<sup>11</sup> Savage is therefore content with the suggestion that the only norm implicit in the theory is an injunction to obey the axioms.

But why exactly does Savage feel “uncomfortable” when he finds himself with intransitive preferences? Are intransitive preferences really like inconsistent beliefs?

What makes a triad of intransitive preferences seem inconsistent is that they

<sup>10</sup> L. J. Savage, *The Foundations of Statistics* (New York: John Wiley and Sons, 1954), 19–21.

<sup>11</sup> That Savage defines preference as a disposition to choose can be confirmed on p. 17.

tend to conflict, in the sense that acting on any two of them would entail frustrating the third. That is, if someone with Savage's intransitive preferences is offered a choice between H and G, his preference between these options commits him to taking G; and if he is then offered a choice between keeping G and getting F instead, his preference between them commits him to taking F; but his rejecting H in favor of G, and G in favor of F, would add up, in effect, to his rejecting H in favor of F, which would be contrary to his own preference between these two options.

Some theorists prefer to demonstrate the conflict among intransitive preferences by showing that the agent who holds them can be turned into a "money pump." Granted H, the agent should be willing to trade it plus a small sum of money in exchange for G, which he prefers to H; whereupon he should be willing to trade G plus a small sum for F, which he prefers to G; whereupon he should be willing to trade F plus a small sum for H, which he prefers to F; whereupon he will be back where he started but substantially poorer. What this argument shows is that the agent's intransitive preferences commit him to giving up something that he prefers—namely, having more money rather than less—without gaining any preferred option in compensation. Similar arguments can be constructed to show that preferences violating other axioms commit the agent to contravening his own preferences in the same way, by accepting the frustration of preferences that were previously satisfied without gaining the satisfaction of any preferences that were previously frustrated.<sup>12</sup>

Note, however, that all of these arguments depend on the assumption that an agent's preferences can be satisfied or frustrated by transactions other than the binary choice over which they are defined.<sup>13</sup> The preference for H over F, for example, is defined in the context of a choice between these two options. Yet in arguing that intransitive preferences are mutually conflicting, we assumed that the agent committed to rejecting H for G and G for F was thereby committed to a course of action that would frustrate his preference for H over F, since it would in effect (as we put it) entail his rejecting the preferred alternative. But rejecting H for F *in effect*, by choosing first between H and G and then between G and F, is not strictly the same as rejecting H for F in a choice between H and F themselves—the choice over which the preference for H over F is defined. Similarly, sacrificing wealth for poverty in the course of buying and selling F, G, and H is not strictly the same as sacrificing wealth for poverty when these two financial conditions are the only alternatives on offer.

Thus, the axioms of decision theory express conditions of consistency for

<sup>12</sup> No such argument can be constructed for the continuity axiom. And arguments constructed for the other axioms may depend on the agent's willingness to engage in an infinite series of transactions.

<sup>13</sup> This claim is equivalent to the thesis of Frederic Schick's 'Dutch Bookies and Money Pumps,' *The Journal of Philosophy* 83 (1986) 112–19.

preferences only if each preference is to be regarded as transcending its context—that is, as satisfied or frustrated not only in the context of the choice over which it is defined but also in the context of other choices that somehow add up to that choice in effect. Otherwise, violations of the axioms cannot be characterized as committing the agent to accepting losses in preference satisfaction.

Unfortunately, preferences needn't be context transcendent. Indeed, we can explicitly restrict each preference to its context by incorporating a description of that context into our specification of the options involved. And the effect of this redescription will be to short-circuit the normative force of the axioms, by making them impossible to violate.<sup>14</sup>

The apparent intransitivity of Savage's preferences, for example, can be eliminated if he claims to be holding preferences—not for F over G, G over H, and H over F—but rather for F-rather-than-G over G-rather-than-F, for G-rather-than-H over H-rather-than-G, and for H-rather-than-F over F-rather-than-H. Under these contextualized descriptions, no alternative is repeated across any two of the agent's preferences: each of his preferences ranges over its own unique pair of options. Hence the fundamental conditions for transitivity or intransitivity are lacking from these preferences, and any opportunity to violate the transitivity axiom with these preferences has vanished.

The contextualization strategy works by spreading out the agent's inconsistent preferences, so to speak, over an expanded range of options. It refracts each option through the prism of other options, so that F is split into F-not-G and F-not-H, G is split into G-not-F and G-not-H, and so on. Each of his initially inconsistent preferences can then be insulated from possible conflict with the others by being reassigned to its own, distinct pair of options. Because the agent's preferences over F, G, and H violated the axioms to begin with, his preferences over the expanded range of options may be uncoordinated, in the sense that preferences involving F-not-G will bear no systematic relation to preferences involving F-not-H (or to preferences involving F simpliciter, if any such preferences remain). Though uncoordinated in this sense, however, the agent's preferences will no longer violate the axioms, and they will no longer be inconsistent, strictly speaking.

One might think that redescribing the options doesn't save the agent from conflicts of the sort we have already considered.<sup>15</sup> For if he already has H but prefers G-not-H to H-not-G, then he should be willing to trade H plus a sum

<sup>14</sup> For a recent discussion of this issue, see John Broome, 'Can a Humean be Moderate?', in *Value, Welfare, and Morality*, ed. R. G. Frey and Chris Morris (Cambridge: Cambridge Univ. Press, 1992), 51–73.

<sup>15</sup> This response is discussed by Broome in 'Can a Humean be Moderate?' Broome gives a slightly different rejoinder (which I describe in n. 17, below).

preferences only if each preference is to be regarded as transcending its context—that is, as satisfied or frustrated not only in the context of the choice over which it is defined but also in the context of other choices that somehow add up to that choice in effect. Otherwise, violations of the axioms cannot be characterized as committing the agent to accepting losses in preference satisfaction.

Unfortunately, preferences needn't be context transcendent. Indeed, we can explicitly restrict each preference to its context by incorporating a description of that context into our specification of the options involved. And the effect of this redescription will be to short-circuit the normative force of the axioms, by making them impossible to violate.<sup>14</sup>

The apparent intransitivity of Savage's preferences, for example, can be eliminated if he claims to be holding preferences—not for F over G, G over H, and H over F—but rather for F-rather-than-G over G-rather-than-F, for G-rather-than-H over H-rather-than-G, and for H-rather-than-F over F-rather-than-H. Under these contextualized descriptions, no alternative is repeated across any two of the agent's preferences: each of his preferences ranges over its own unique pair of options. Hence the fundamental conditions for transitivity or intransitivity are lacking from these preferences, and any opportunity to violate the transitivity axiom with these preferences has vanished.

The contextualization strategy works by spreading out the agent's inconsistent preferences, so to speak, over an expanded range of options. It refracts each option through the prism of other options, so that F is split into F-not-G and F-not-H, G is split into G-not-F and G-not-H, and so on. Each of his initially inconsistent preferences can then be insulated from possible conflict with the others by being reassigned to its own, distinct pair of options. Because the agent's preferences over F, G, and H violated the axioms to begin with, his preferences over the expanded range of options may be uncoordinated, in the sense that preferences involving F-not-G will bear no systematic relation to preferences involving F-not-H (or to preferences involving F simpliciter, if any such preferences remain). Though uncoordinated in this sense, however, the agent's preferences will no longer violate the axioms, and they will no longer be inconsistent, strictly speaking.

One might think that redescribing the options doesn't save the agent from conflicts of the sort we have already considered.<sup>15</sup> For if he already has H but prefers G-not-H to H-not-G, then he should be willing to trade H plus a sum

<sup>14</sup> For a recent discussion of this issue, see John Broome, 'Can a Humean be Moderate?', in *Value, Welfare, and Morality*, ed. R. G. Frey and Chris Morris (Cambridge: Cambridge Univ. Press, 1992), 51–73.

<sup>15</sup> This response is discussed by Broome in 'Can a Humean be Moderate?' Broome gives a slightly different rejoinder (which I describe in n. 17, below).

of money in exchange for G; whereupon, if he prefers F-not-G to G-not-F, he should be willing to trade G plus a sum of money for F; whereupon, if he prefers H-not-F to F-not-H, he should be willing to exchange F plus a sum of money for H—whereupon he'll be back where he started but substantially poorer.<sup>16</sup> Hence, the agent's preferences still seem to conflict with his preference for more money rather than less.

The problem with this version of the money-pump argument is that it depends on the agent to be less than thorough in implementing the contextualization strategy. For it assumes that the agent can still be led around a circle that returns him, with less money in his pocket, to an earlier state of play. But an agent who is sufficiently thorough in contextualizing his preferences can avoid ever having to return to the same state of play. Having exchanged H-not-G for G-not-H, he can describe the next pair of options—not as G-not-F and F-not-G—but as (G-not-H)-not-F and F-not-(G-not-H); moreover, if he takes the latter option, he can describe the succeeding pair as [F-not-(G-not-H)]-not-H and H-not-[F-not-(G-not-H)]; and so on, indefinitely. The agent can thereby avoid ever revisiting the same state, and so he can avoid the discomfiture of having shelled out money to do so. Contextualized descriptions can thus eliminate the conditions necessary for being described as a money pump, just as they can eliminate the conditions necessary for violating the axioms.<sup>17</sup>

A sufficiently thorough implementation of the contextualization strategy can similarly undermine any attempt to criticize that strategy as leading to losses in preference satisfaction. This criticism always depends on a judgment that some sequence of transactions fails to satisfy any previously unsatisfied preferences while frustrating some preference that wasn't previously frustrated. But the contextualization strategy ensures that every transaction in a sequence will satisfy a preference that *couldn't* previously have been satisfied, precisely because its satisfaction requires the prior occurrence of all the preceding transactions.

The axioms of decision theory cannot rule out this evasive maneuver. For

<sup>16</sup> Strictly speaking, the agent is not back where he started: he started with H-not-G and ended up with H-not-F, and he prefers the latter of these outcomes to the former. From his perspective, then, he has been compensated for the outlay of money by the satisfaction of an additional preference. Nonetheless, if he is once again offered G in exchange for H, then his preference for G-not-H over H-not-G commits him to accept, and he will indeed return to an earlier state of play with less money in his pocket.

<sup>17</sup> John Broome offers a different objection to this version of the money-pump argument. Being offered G a second time in exchange for H (as described in the preceding note) will have forced the agent to choose between H-not-G and G-not-H, thereby depriving him of the preferred outcome that he had managed to buy—namely, H-not-F. The fact that someone can then sell H-not-F back to him doesn't show that he can be exploited. All it shows, as Broome puts it, is that someone can steal his shirt and then get him to buy it back. (See 'Can a Humean be Moderate?')

although the axioms stipulate that the agent's response to an outcome F as it appears in one choice must be coordinated with his response to F as it appears in other choices, the axioms do not require that these choices be described, to begin with, as involving one and the same outcome F rather than as involving different outcomes, consisting of F-in-different-contexts. Nevertheless, the maneuver of eliminating inconsistencies by contextualizing the options strikes us as violating the spirit, if not the letter, of the theory. It strikes us, in a word, as cheating.

An account of the theory's normative force must be able to explain why contextualization is illegitimate. The threat posed by this strategy is not merely that it will prevent the axioms from requiring any adjustments in our preferences, so long as we're willing to adjust our descriptions of them. A more fundamental threat is that the axioms won't count as requirements on preference, to begin with, if preferences are confined to their contexts, whether by description or otherwise.

Even if we don't describe our preferences contextually, our sense that they are subject to the axioms as requirement of consistency depends on the assumption that they transcend the choices over which they are defined. We cannot say what's inconsistent about preferring F to G, G to H, and H to F without assuming that each of these preferences can be satisfied or fulfilled outside of the associated choice, in a sequence of the other two choices. Our sense that these *uncontextualized* preferences are inconsistent thus depends on the sense that they are such as *shouldn't* be described contextually. Unless we can explain why contextualization seems wrong in this sense, we cannot defend our conception of the axioms as requiring anything.

As John Broome has noted, contextual descriptions can be appropriate in particular cases, because they highlight evaluatively significant features of the choice at hand. Broome's example of this possibility is an agent who feels better about staying home when the alternative is sightseeing than he does about staying home when the alternative is mountaineering, because staying at home would be cowardly in the latter instance but not in the former.<sup>18</sup> This agent can justifiably think of himself as choosing between sightseeing-rather-than-staying-home and staying-home-rather-than-sightseeing, or between mountaineering-rather-than-staying-home and staying-home-rather-than-mountaineering.

Contextualization is problematic only in other cases, Broome suggests, in which it entails drawing distinctions that make no significant difference. For many instances of F, G, and H, the alternative F-rather-than-G won't differ from F-rather-than-H in any respect that one might rationally care about. Con-

<sup>18</sup> *Weighing Goods* (Oxford: Blackwell, 1991), ch. 5. See also 'Can a Humean be Moderate?', 58.

What we're saying, I think, is that contextualizing one's preferences satisfies the axioms while somehow defeating the point of doing so. That is, there must be some ulterior point or purpose to having preferences that are transitive, monotonic, and so on; and it must be a point or purpose that could somehow be defeated when transitivity, monotonicity, and the rest are achieved by contextualization rather than by straightforward means. But what is the point or purpose that obedience to the axioms is meant to serve?

Well, a purpose that we know to be served by obedience to the axioms is representability. The proof of the representation theorem demonstrates that obeying the axioms enables us to represent our preferences in terms of utility and probability functions. Of course, the same proof applies to any preferences that obey the axioms, including those whose obedience has been achieved by contextualization rather than straightforwardly. Hence, representability alone cannot be the point that's defeated by contextualization.

There is this difference, however. If an intransitivity in one's preferences among F, G, and H is eliminated by the straightforward reversal of a preference, then the resulting preferences can be represented by assignments of utility to the three options F, G, and H.<sup>19</sup> But if the intransitivity is eliminated by means of contextualization, the resulting preferences must be represented by assignments of utility to the six distinct options that contextualization has produced: F-not-G, G-not-F, G-not-H, H-not-G, H-not-F, and F-not-H. In the former case, one's preferences are represented by pairs of utilities drawn from the same three values; in the latter case, one's preferences are represented by pairs of utilities that are entirely disjoint, six different values in all.

Although contextualization succeeds in making preferences representable, then, an important virtue is lacking from the resulting representation. The representation made possible by contextualization doesn't reduce three preferences to alternative pairings of the same three values; rather, it elaborates them into unrelated pairings of six different values. Contextualization makes one's preferences representable, but only in more complex terms, each of which carries less descriptive power.

Thus, a purpose that obedience to the axioms generally serves, but contextualization defeats, is the purpose of representing one's preferences in concise and powerful terms. So let's reflect, for a moment, on the descriptive power of the representation whose possibility is asserted by the representation theorem.

An agent's options can include not only every possible outcome but also

<sup>19</sup> Of course, if F, G, or H represents chancy options, then the utilities assigned to it will be expected utilities. I shall ignore this complication.

textualizing the descriptions of these alternatives will entail drawing senseless distinctions, according to Broome, thereby violating what he calls principles of rational indifference.

I agree that contextualized descriptions can draw sensible distinctions in some cases and not in others. But I don't think that we can explain what's wrong with contextualization by saying that it sometimes entails drawing senseless distinctions among the options, since we aren't currently considering it as a way of drawing such distinctions, in the first place. We are considering contextualization as a way of formalizing a conception of preference—namely, the conception that each preference is specific to the choice over which it is defined. We are therefore considering descriptions such as “staying-home-rather-than-sightseeing” and “staying-home-rather-than-mountaineering,” not to indicate supposed differences between the options so described, but rather to express the notion that staying home can be an object of preference only in the context of binary choices involving specific alternatives. And when contextualized descriptions are used to express this conception of preference, they cannot be rejected on the grounds that they indicate no significant difference between the options, since they don't purport to indicate such differences, anyway.

What's more, rejecting contextualized descriptions on grounds of their misrepresenting the options won't rule out the conception of preferences that they were in fact meant to express. For even if the agent agrees to describe his options simply as staying home, sightseeing, and mountaineering, he can still conceive of his preferences among them as capable of being satisfied or frustrated only in the choices over which they are defined; and he will then have no reason to see the axioms as requirements of consistency for these preferences. In that case, we shall still have to tell him why his preferences shouldn't be conceived as context-bound, whether or not he expresses this conception in how he describes the options.

The question therefore remains why we shouldn't contextualize preferences in general, irrespective of how we describe particular cases. The normative force of the axioms has yet to be explained in a way that answers this question. I hope to offer such an explanation.

Let me develop my explanation by returning to our intuitive dissatisfaction with contextualization as a means of obeying the axioms. We initially expressed this dissatisfaction by saying that contextualizing one's preferences satisfies the letter of the axioms while violating their spirit. Surely, the “spirit” of the axioms is what we're trying to understand when we try to understand their normative force. The way to understand the normative force of the axioms may therefore be to figure out what exactly is violated by the strategy of contextualization. What are we talking about when we say that contextualizing one's preferences violates the spirit of the axioms?

every possible permutation of outcomes and contingencies, combined to form gambles. Since the agent can have preferences over every possible pairing of these multifarious options, his preferences can be exponentially multifarious.<sup>20</sup> Yet the representation theorem guarantees that so long as his preferences obey the axioms, they will be fully characterizable in concise and systematic terms—specifically, in terms of a single number for each basic outcome, a single number for each basic contingency, and a simple formula for computing the actuarial products of these numbers. This characterization of the agent's preferences will be concise because it will attach numbers only to basic outcomes and contingencies rather than to permutations of their permutations, which is what pairs of options are. And the characterization will be systematic because it will relate every preference to these numbers by means of a single mathematical formula.

Suppose that the world can be in any one of ten states, and that each state can result in the agent's receiving any one of ten different payoffs. These outcomes and contingencies can be combined to form  $10^{10}$  gambles, which can be paired in  $(10^{10})!$  possible binary choices. Yet, even if the agent has a preference over each of these choices, his preferences can be represented by ten utilities, ten probabilities, and a single formula for computing expected utility—provided, of course, that his preferences obey the axioms. Obedience to the axioms thus guarantees truly impressive economies in the agent's self-description.

What the representation theorem tells us, then, is that preferences obedient to the axioms will be synoptically describable. And to my way of thinking, this is the ultimate basis of the axioms' normative force.

After all, to guarantee that something will be synoptically describable is to guarantee that it will make a certain kind of sense. When we have managed to comprehend manifold items under a simple but informative description, a description that subsumes their multiplicity under some uniformity, we have then made sense of those items—that is, rendered them intelligible—in the most fundamental way. Representable preferences will make sense in this way, by being subsumable under the uniformity of maximizing expected utility, as calculated from a common set of values. What I claim is that making sense is the point of having representable preferences, and hence the ultimate point of conforming one's preferences to the axioms of decision theory.

Indeed, I think that for preferences to make sense, by being synoptically describable, just is for them to be formally rational. I therefore interpret the

<sup>20</sup> Some versions of the theory actually require an agent to have preferences over all of these choices; others do not. My point is simply that the agent *can* have preferences over all of these choices and still represent them in the same concise and powerful terms spelled out by the representation theorem.

representation theorem as asserting, and the proof of that theorem as proving, that obedience to the axioms yields formally rational preferences.

This brief statement of my view contains what may strike the reader as an equivocation. I say that formally rational preferences are the preferences that make sense; and the reader may think that what I say is true enough, but only if “making sense” is understood as a normative expression synonymous with “being rational.” When I say that rational preferences are the ones that make sense, however, I am not propounding the tautology that rational preferences are the ones that are rational. Rather, I’m saying that formally rational preferences are the ones that are accessible to a particular mental act, an act of synoptic characterization or comprehension. How can I assume that this purely psychological conception of making sense can be substituted for the normative conception?

Well, I’m not exactly assuming this; I’ve argued for it elsewhere,<sup>21</sup> and I’m arguing for it again here. My argument is partly an inference to the best explanation: for as I shall show in a moment, the identification of rationality with intelligibility helps to explain our dissatisfaction with the strategy of contextualization.

But my argument is also partly an appeal to brute intuition. I am asking the reader to reflect on the felt authority of the requirements expressed in the axioms of decision theory, and to consider the following hypothesis. When you feel obliged to make your preferences transitive (for example), isn’t it because you feel obliged to align them into some coherent posture toward the options? And isn’t the point of such a posture simply that it will lend your preferences an intelligible order, a unifying thread, a common orientation—in short, a *rationale*? What it is for preferences to have a rationale—some rationale or other, whether it be good or bad—is simply for them to follow some organizing principle under which they can be comprehended. The felt obligation to have transitive preferences, I am claiming, is an obligation to have preferences that cohere around a rationale in this manner. It’s thus an obligation to have preferences that make sense.

In the past I have argued that practical rationality consists in intelligibility of a more robust kind. I have argued, for example, that a rational thing to do is a thing that would make sense, or be intelligible, in that one would be able to explain doing it. And I have assumed that explaining what one does entails citing its causes. I have therefore argued, for example, that desires and beliefs are reasons for acting insofar as they are causes by reference to which an action might be explained.

Yet to represent one’s preferences in terms of utilities and probabilities is not to explain them causally. Remember, the utilities and probabilities that help

<sup>21</sup> *Practical Reflection* (Princeton: Princeton Univ. Press, 1989).

to represent consistent preferences are emergent in, or supervenient on, the preferences that they help to represent: they are complex ways of preferring. And ways of preferring cannot cause the very preferences in which they emerge, or on which they supervene.

The utilities and probabilities that represent consistent preferences make them susceptible, not to causal explanation, but merely to perspicuous summarization. I do not find it helpful to describe this rudimentary vehicle of understanding as an explanation of any sort. To subsume something under a synoptic characterization is not really to explain it. But some philosophers have analyzed explanation as a kind of synoptic description, and their analyses can help me to clarify what I have in mind.

One such analysis is an account of scientific explanation offered by Michael Friedman in the 1970s.<sup>22</sup> Friedman's paradigm of scientific explanation is the derivation of one natural law from a more general law or theory—for instance, the derivation of the Boyle-Charles law of gases from the kinetic theory of molecular behavior. Friedman asks:

How does this make us understand the behavior of gases? I submit that if this were all the kinetic theory did we would have added nothing to our understanding. We would have simply replaced one brute fact with another. But this is not all the kinetic theory does—it also permits us to derive other phenomena involving the behavior of gases, such as the fact that they obey Graham's law of diffusion and (within certain limits) that they have the specific-heat capacities that they do have, from the same laws of mechanics. The kinetic theory effects a significant *unification* in what we have to accept. Where we once had three independent brute facts—that gases approximately obey the Boyle-Charles law, that they obey Graham's law, and that they have the specific-heat capacities they do have—we now have only one—that molecules obey the laws of mechanics.<sup>23</sup>

Friedman thus arrives at the following account of scientific explanation:

[S]cience increases our understanding of the world by reducing the total number of independent phenomena that we have to accept as ultimate or given. A world with fewer independent phenomena is, other things equal, more comprehensible than one with more.<sup>24</sup>

Now, I suspect that Friedman's account neglects one aspect of his example that is partly responsible for its being truly explanatory. The derivation of the Boyle-Charles law doesn't just unify one fact with two other facts under the cover of a single, more comprehensive fact; it unifies one law with two other laws under the cover of a single, more comprehensive law. Friedman seems to obscure

<sup>22</sup> 'Explanation and Scientific Understanding,' *The Journal of Philosophy* 71 (1974) 15–19.

<sup>23</sup> *Ibid.*, 14–15.

<sup>24</sup> *Ibid.*, 15. For a critique of Friedman's theory, see Wesley C. Salmon, 'Four Decades of Scientific Explanation,' *Minnesota Studies in the Philosophy of Science* 13 (1989) 94–101. The debate here is primarily about whether Friedman's notion of "the total number of independent phenomena" can be satisfactorily formalized.

this distinction with the ambiguous word “phenomenon,” which can refer to either accidental or lawlike regularities. But I doubt whether unifying purely accidental regularities under more comprehensive but equally accidental regularities would look like scientific explanation, even though it would indeed yield a world that was more comprehensible by virtue of having “fewer independent phenomena.” Scientific explanation has to provide something more.<sup>25</sup>

Still, Friedman has identified one element of scientific explanation, and it’s the element that interests me at present. Whatever else scientific explanations may provide, Friedman is right that they provide greater comprehension—a synoptic grasp of several phenomena that were previously grasped independently. And this sort of comprehension is what I am claiming that we attain when we can apply the methods of the representation theorem to a congeries of preferences. If someone’s preferences obey the axioms of decision theory, we can grasp them as falling into the coherent pattern of promoting constant utilities in light of constant probabilities. Thus comprehended, the preferences make more sense.

Another source for the relevant concept of comprehension is Louis Mink’s account of narrative explanation in history. Dissatisfied with the suggestion that historical narratives render events intelligible by revealing their causes,<sup>26</sup> Mink characterized narrative understanding as comprehension in the literal sense of a “grasping together”—“a characteristic kind of understanding which consists in thinking together in a single act . . . the complicated relationships of parts which can be experienced only *seriatim*.”<sup>27</sup> When history is presented in a coherent narrative, Mink argued, “actions and events, although represented as occurring in the order of time, can be surveyed as it were in a single glance as bound together in an order of significance, a representation of a *totum simul*.”<sup>28</sup>

Mink presented this account of historical understanding as a variation on the views of W. B. Gallie, which he summarized as follows:<sup>29</sup>

<sup>25</sup> The distinction between brute facts and laws is more carefully observed in a passage that Friedman quotes from William Kneale: “[T]he explanation of laws by showing that they follow from other laws is a simplification of what we have to accept because it reduces the number of untransparent necessitations we need to assume,” *Probability and Induction* (New York: Oxford Univ. Press, 1949), 91–92.

<sup>26</sup> See Mink’s critique of Morton White’s *Foundations of Historical Knowledge*, in ‘Philosophical Analysis and Historical Understanding,’ in *Historical Understanding*, ed. Brian Fay, Eugene O. Golob, and Richard T. Vann (Ithaca: Cornell Univ. Press, 1987), 118–46. For another review of philosophical work in this area, see W. H. Dray, ‘On the Nature and Role of Narrative in Historiography,’ *History and Theory* 10 (1971) 153–71.

<sup>27</sup> ‘History and Fiction as Modes of Comprehension,’ in *Historical Understanding*, 50.

<sup>28</sup> *Ibid.*, 56.

<sup>29</sup> *Ibid.*, 46. The view being summarized here is set forth by W. B. Gallie in *Philosophy and the Historical Understanding* (New York: Schocken Books, 1968).

In following a story, as in being a spectator at a [cricket] match, there must be a quickly established sense of a promised although unpredictable outcome: the county team will win, lose, or draw, the separated lovers will be reunited or will not. Surprises and contingencies are the stuff of stories, as of games, yet by virtue of the promised yet open outcome we are enabled to follow a series of events across their contingent relations and to understand them as leading to an as yet unrevealed conclusion without however necessitating that conclusion.

Mink did not share Gallie's concern with unpredictability and its role in drawing us along through a story; indeed, he was not interested in how we follow a story when reading or hearing it for the first time, since historians often tell us stories whose outcomes we already know. Rather, Mink was interested in how the characterization of events in terms of their relations to an outcome enables us to comprehend them as a completed whole after the story is finished.

Consider, for example, the story of *Treasure Island*, whose very title already hints at the "promised although unpredictable outcome" in light of which the story's various episodes are to be comprehended. Every major event in the story has some intrinsic description of its own, but it also has some description in relation to the outcome in question. Each major event can be regarded as either motivating or furthering or hindering or somehow bearing on the pursuit of Flint's treasure. And within the story, other promised outcomes serve a similar organizing role. As soon as the word "mutiny" is uttered in the confrontation between the Captain and the Squire, subsequent events can be comprehended as revelations of, responses to, actions upon, or deviations from the sailors' plan to revolt (which can itself be comprehended as an obstacle to recovering the treasure). The mutiny and the recovery of the treasure are thus common points of reference towards which we can orient our conception of the other events in the story; and having thus aligned our conception of the events, we can grasp them together rather than merely review them in succession.<sup>30</sup>

Although we thereby gain comprehension, which might be called a kind of understanding, this mode of understanding doesn't necessarily rest on an explanation of the events understood. Of course, the narrators of *Treasure Island* offer explanations of many events, but these explanations are self-contained digressions from the narrative and do not contribute to the sort of comprehension that interests Gallie or Mink. Again, many of the events that are comprehensible by virtue of their relation to the mutiny, or to the recovery of the treasure, are related to these outcomes as individually necessary or jointly sufficient conditions for them, and so they provide a partial explanation of why the mutiny occurred or why the treasure was recovered.

<sup>30</sup> These retrospective characterizations of events are what Arthur C. Danto calls "narrative sentences," *Narration and Knowledge* (New York: Columbia Univ. Press, 1985), ch. 15.

But equally many events may be comprehensible by virtue of being related to these outcomes as hindrances, inhibitions, or obstacles; and the comprehensibility of the story does not depend on its making clear why the favorable conditions won out over the unfavorable. In short, how comprehensible the story is does not depend on how well it explains *why* the treasure was found. Rather, it depends on how well the events in the story can be grasped together as bearing on this outcome in some way or other, favorably or unfavorably.<sup>31</sup>

The orientation of events toward a foreshadowed outcome is only one of many ways in which narrative form renders events comprehensible. Philosophers of history, literary theorists, and cognitive scientists offer many other examples of, and criteria for, the intelligibility of stories. For my purposes, however, this one illustration of narrative intelligibility will do, since it serves as a convenient analog for the way in which preferences make sense when they obey the decision-theoretic axioms.

The completed analogy is this. Preferences that obey the axioms can be represented as jointly oriented towards the outcome of maximal expected utility, and so they are intelligible in the same way as disparate events that point toward the foreshadowed outcome of a narrative. Consistent preferences make sense because they hang together, like the episodes in a coherent story.<sup>32</sup>

I can summarize these digressions into the philosophy of science as follows. Whereas the principle of intelligibility in a story is called the plot, and the principle of intelligibility in natural phenomena is called a law, the principle of intelligibility in an agent's preferences, I am claiming, is called a rationale. And having a rationale, I claim, is the condition of formal rationality in preferences.

This interpretation of the norm embodied in decision theory enables me to explain why contextualizing your preferences seems like cheating. For I interpret the axioms of decision theory as directing you to coordinate your preferences in such a way that they will make sense; and when so interpreted, the axioms have a point that is indeed defeated by the strategy of contextualization.

<sup>31</sup> Conversely, the complete explanation of an outcome may convey more than an understanding of its explanandum, since it may also convey comprehension of the events mentioned in its explanans. An historical explanation of why the Civil War occurred, for example, may help us not only to understand the outbreak of the Civil War but also to grasp together many otherwise disparate conditions and events, by unifying them under the concept "causes of the Civil War."

<sup>32</sup> Alasdair MacIntyre also combines the concepts of rational action, intelligibility, and narrative. See, e.g., 'The Intelligibility of Action,' in *Rationality, Relativism and the Human Sciences*, ed. J. Margolis, M. Krausz, and R. M. Burian (Dordrecht: Martinus Nijhoff, 1986), 63–80. Other than the fact that we both combine these three concepts, however, I can find very little in common between us.

The point of having transitive preferences among three options, according to my interpretation, is to have preferences that can be easily comprehended as pairwise comparisons of the same three values. But if your preferences must be redescribed in terms of six different options, none of which is repeated, then they can no longer be unified in the same fashion. To be sure, utilities will still be assignable in such a way that the option with the higher utility is always preferred; but the utilities thus assigned won't serve as unifying threads among these preferences, aligning them into a coherent posture toward the options, since each utility will help to represent only one of the preferences, and each preference will be represented by different utilities. These utilities may also unify the preferences in question with yet further preferences over the expanded range of options that contextualization has generated; but the unification thus achieved will not provide a common rationale for the three preferences with which you began—the preferences whose intransitivity drove you to contextualize. It will therefore fail to resolve the problem that the intransitivity initially posed, and so it will defeat the point of removing that intransitivity.

The point, as I have suggested, is to comprehend the preferences you have as forming a coherent posture toward the choices you face. And this point requires not only that you coordinate your responses to the same outcome or contingency as it appears in different choices, but also that you conceive of your choices, in the first place, as containing the same repeatable outcomes and contingencies. Insofar as you redescribe each of your choices as unique, as sharing no components with other choices, you ensure that your responses to them will resist synopsis; and so you ensure that your responses to these choices will elude any synoptic grasp. You therefore violate the spirit of the axioms when you redescribe your existing options as having nothing in common with one another, even if you simultaneously invent a larger range of choices for them to have something in common with. Multiplying your options and preferences in this manner yields less intelligibility, not more, and so it violates the underlying norm of decision theory, which is the injunction to have preferences that make sense.

Note that this critique is aimed, not at contextual descriptions of particular options, but rather at the conception of preferences that's expressed by contextualization as a general strategy. It explains why there is a rational pressure against confining preferences to their own contexts. The reason is that making sense of your preferences requires you to see them as manifesting some constant, underlying posture toward the options, and hence as bearing upon one another as expressions of the same rationale. To regard your preferences for F over G and for G over H as having nothing to do with your preference between F and H would run directly contrary to the goal of having preferences that hang together so as to make sense.

Of course, merely making sense can render preferences rational only in a very weak understanding of the word. All sorts of bizarre, perverse, and otherwise unsavory sets of preferences are organized around some principle of intelligibility or other. Surely I don't intend to claim that all such preferences are rational?

My reply is that I am trying to identify the norm of rationality that's embodied in decision theory, and decision theory can claim to embody only one such norm, and a weak one at that. As many have noted, all sorts of bizarre, perverse, and otherwise unsavory sets of preferences can satisfy the axioms of decision theory, too, if they are consistently bizarre, consistently perverse, or consistently unsavory.<sup>33</sup> All that decision theory can claim to formalize is rational consistency—a virtue possessed by many agents whom we would still like to criticize. In looking for the norm that's formalized in decision theory, then, we should expect to find a norm that's very weak when compared with the other norms that we wish to apply.

These other norms, whatever they are, can be described in one of two ways. On the one hand, we might think that norms other than that of rational consistency can still be norms of rationality, expressing substantive demands that rationality makes over and above merely formal demands of the sort that are embodied in decision theory. We might think that substantive rationality requires us, for example, to prefer pleasure over pain, or not to prefer present pleasure over future pleasure, and so on.

On the other hand, we might think that consistency is the only requirement of rationality, and that all other requirements are expressive of other virtues. Consistent preferences may still be criticized on many grounds, we might think—e.g., as insensitive, short-sighted, masochistic—but not on the grounds of being irrational.

My interpretation of decision theory lends support to the former view. For if the rational consistency of our preferences consists in their having an intelligible structure, as I have claimed, then a coordinate, substantive mode of rationality can be discerned in their having an intelligible content. Some sets of preferences that make formal sense, in that they can be perspicuously summarized, may still be at odds with what we know about human nature in general, for example, or about ourselves in particular; and so they may still fail to make sense, substantively speaking. Synoptically describable preferences may still be inexplicable or inscrutable in their content, because we cannot understand or explain why we, or anyone, would have preferences with that content, no matter how internally coherent they may be. As we have seen, the mode of understanding that's provided by coherent narratives, according to

<sup>33</sup> See, e.g., Sen, 'Rationality and Uncertainty,' in *Recent Developments in the Foundations of Utility and Risk Theory*, ed. L. Daboni et al. (Dordrecht: D. Reidel, 1986), 4.

Mink, or by overarching generalizations, according to Friedman, is not all of the understanding we might want of the phenomena that they summarize. Similarly, summarizing our preferences in terms of utilities and probabilities may still leave us quite baffled by those preferences in many respects.

If so, our preferences will lack a virtue that's clearly of a piece with synoptic describability, and hence with rational consistency, as I conceive it. Indeed, this virtue is related to rational consistency in precisely the way that one would expect substantive and formal rationality to be related, since it is a substantive form of intelligibility distinct from, but coordinate with, the purely formal intelligibility found in preferences that can be synoptically described.

Note, however, that under my interpretation the relation between formal and substantive rationality is not as it is ordinarily imagined to be. Ordinarily, rational consistency is imagined to be a necessary (but not sufficient) condition for substantive rationality in one's preferences. All rational preferences are assumed to be at least rationally consistent; substantively rational preferences are assumed to be rationally consistent and more. A theory of substantive rationality is therefore expected to tell us which among the rationally consistent sets of preferences one ought to have.

Yet if formal and substantive rationality are related as formal and substantive intelligibility, then they have the potential to conflict. The preferences that make the most sense substantively, in light of human or individual natures, may not be the ones that are most perspicuously summarized when considered alone. Conversely, the preferences that can be summarized most perspicuously may be quite inexplicable or inscrutable as expressions of our personality, or of any human personality at all.

In order to understand this point, keep in mind that I interpret formal rationality as a matter of degree. Rational consistency, conceived as intelligibility, is not an all-or-nothing affair, and so it is unlike consistency as understood in other contexts.

Consider a set of preferences that depart only slightly from the requirements of maximizing expected utility. Suppose that you prefer two million dollars to one million dollars and yet prefer one million dollars to any gamble with payoffs of two million dollars and zero dollars. In that case, there are no utilities and probabilities whose actuarial products you are maximizing. Any assignment of utilities to these monetary sums will yield a determinate proportion between your potential gain, in getting two million dollars instead of one million, and your potential loss in getting no dollars; and there will therefore be a determinate probability of winning at which your discounted gain from the gamble is greater than your discounted loss. But there is no probability of winning at which you prefer to take the gamble, and so there is no assignment of values for which your preferences maximize expected utility.

Because your preferences in this example maximize no measure of expected utility, they do not add up to a coherent posture toward the outcomes as being simply better or worse. There are assignments of utility and probability that will represent your response to any choice in which your financial security is not at stake—that is, any choice that doesn't set a certainty of having enough money against a risk of having less. But the same utilities and probabilities, by themselves, will not represent your responses to the remaining choices, in which you are invited to place financial sufficiency at risk for further gain. Indeed, no assignment of utilities and probabilities can unify these two subsets of your preferences.

The discontinuity in your preferences makes them less unifiable, less synoptically describable, and consequently less intelligible in the sense that I have defined. Formally speaking, then, these preferences are less rational, not only according to traditional decision theory but also according to my interpretation of it.<sup>34</sup>

Although discontinuous preferences lack the highest degree of formal rationality, they still possess such rationality to some degree. For although they do not add up to a coherent posture toward the outcomes as being simply better or worse, they still add up to a somewhat subtler posture toward the outcomes, as being better or worse but also, in some cases, good enough.<sup>35</sup>

In my example you behave as if you value one million dollars less than two million and yet value one million as being enough, in the sense that you refuse to put it at risk for the sake of a chance to gain more. One thing that we sometimes mean in saying that we have enough is precisely that we would rather

<sup>34</sup> This loss of intelligibility is reflected, by the way, in the diminished degree to which your preferences can transcend their contexts. Suppose that you are offered, first, a choice between \$1 million and a chance of getting \$2.5 million if the toss of a coin comes up heads; and, second, a choice between \$1 million and a chance of getting \$2.5 million if the same toss comes up tails. In either choice considered alone, your threshold of sufficiency dictates preferring the certainty of \$1 million to the mere chance of getting more; but when both choices are considered together, they offer the prospect of getting \$2.5 million for sure, which you prefer to \$1 million. Acting on the former preferences will lead you to frustrate the latter, and vice versa.

The only way to avoid this conflict is to contextualize your preferences. What your underlying values should lead you to say is that, although you prefer \$1 million to a chance of \$2.5 million when choosing only between these options, you prefer the chance of \$2.5 million if you'll be getting another choice, in which you'll have the opportunity to raise that chance to the level of certainty. The discontinuity in your other preferences thus forces you, in this case, to frame two different preferences addressed to two different contexts in which the choice between these outcomes might arise.

This inability to generalize your preferences goes hand in hand with your inability to represent them as maximally unified. It therefore goes hand in hand with their falling short of maximal rationality in the formal dimension.

<sup>35</sup> Although I use the language of satisficing to describe your strategy in this example, it is not quite the same as the strategies of satisficing defined by Simon (e.g., 'A Behavioral Model of Rational Choice,' *Quarterly Journal of Economics* 69 (1955) 99–118) or Slote, *Beyond Optimizing; a Study of Rational Choice* (Cambridge, Mass.: Harvard Univ. Press, 1989).

pocket what we have than risk losing it, even if that risk is paired with a chance of further gain. A refusal to trade one million dollars in the hand for two million in the bush can thus be understood as expressing the attitude that one million dollars is sufficient.

Your behavior in my example can be summarized, then, by a value function for money and a threshold designating some amount of money as sufficient. When such a threshold is added to a value function, the result is still a coherent scheme for unifying disparate preferences.<sup>36</sup> And when represented by this scheme, the preferences still make sense. Hence your preferences in this example are rational in the same sense as preferences obedient to the axioms: they can be perspicuously unified by a general framework of values, which provide their rationale. They're just slightly less unifiable—and hence slightly less rational, formally speaking—than preferences that obey the axiom of continuity.

Must we avoid even this slight loss of formal rationality? I think not. For I think that this loss in formal rationality yields a considerable gain in rationality on the substantive dimension. Evaluating everything as continuously better or worse than everything else makes less sense, for creatures like us, than evaluating things not only as better or worse but also, at times, as sufficient or insufficient. Hence continuous preferences, though easier to formulate in themselves, are harder to understand as preferences of ours.

Fully defending this claim about what makes sense for human beings is not the business of the present paper. I think that there are many different aspects of human nature that make discontinuous preferences especially intelligible; here I shall mention just one. Human beings are subject not just to desires but also to needs; and while we often have desires to satisfy our needs in some ways rather than others (say, by gaining two million dollars rather than one million dollars), the satisfaction of those desires is of no importance to us when the relevant needs are at risk.<sup>37</sup> Thus, a creature whose preferences are intelligible in terms of an undifferentiated continuum of values may make less sense as a

<sup>36</sup> Provided, of course, that your other preferences are in keeping with the threshold of sufficiency. If you refuse to trade a certainty of \$1 million for any chance of \$2 million but willingly trade it for a chance of \$3 million, then you aren't consistently treating \$1 million as sufficient.

<sup>37</sup> Another relevant aspect of human nature, I think, is that we experience our lives from different perspectives, including not only a succession of momentary present-tense perspectives but also a perspective of tenseless reflection on our lives as a whole. I have argued elsewhere that discontinuities between these perspectives yield discontinuities in the kinds of value to which we are subject. What is good for us at a particular moment is not necessarily the same, I think, as what is good for us in life; and what is good for us in life is not simply a function of what is good for us at the various moments during which we're alive. (See my 'Well-Being and Time,' Chap. 3, above.) This view yields a scheme of values that may best be expressed in discontinuous preferences.

person than a creature whose preferences are intelligible in terms of values intersected by various thresholds of sufficiency.

I therefore think that fully continuous preferences—in which any loss will be risked for the chance of a large enough gain—can be so coherent in themselves as to be incoherent with our understanding of the people who have them. And if the formal rationality of preferences is a matter of their internal coherence, I don't see how it can be required of them to a degree that undermines their coherence more broadly construed. The virtue that we have found in obeying the decision-theoretic axioms is the virtue of being formally intelligible, which is of a piece with the virtue of being intelligible as a person, in general, and as one sort of person, in particular. How can we owe the former virtue an allegiance so strong that it requires us to forsake the latter?

My interpretation of decision theory thus suggests that preferences can have too much formal rationality. Substantively rational preferences aren't maximally consistent and more, according to my view: they may be *less* than maximally consistent, so as to make more sense in substantive respects.

Actually, a corresponding point may well be true of consistency in beliefs as well. That is, rational beliefs may not be consistent and more; substantive rationality in theoretical matters may sometimes favor sacrificing the consistency of our beliefs, for heuristic or other epistemic purposes. Sometimes consistency really is the hobgoblin of little minds.

Courtesy inhibits me from using this aphorism to sum up my assessment of decision theory as an account of rational preference. I'll sum up my assessment like this. Preferences that obey the axioms of decision theory may indeed possess the ultimate degree of formal consistency, but the ultimate degree of formal consistency is sometimes too much.

Of course, there are differences between consistency in belief and consistency in preference; but these differences only militate against demanding the latter to the same degree as the former. The primary difference, I think, is that belief constitutively aims at the truth, and the body of true beliefs must be fully consistent; whereas preference has no constitutive aim for which consistency is required.<sup>38</sup> Inconsistency in our beliefs entails that some of them are false; and so it is an unmistakable sign of failure, even if it isn't necessarily irrational in some epistemic circumstances. But the reason why obeying the axioms of decision theory is rational, if it is rational, is not that it's a necessary condition for attaining some goal that's essential to preference. Insofar as obeying the axioms of decision theory is rational, it's rational because it makes our preferences formally coherent, thus ensuring that they are intelligible. And sacrificing some degree of formal intelligibility may enable our preferences to

<sup>38</sup> On this contrast, see my 'The Guise of the Good,' Chap. 5, above.

make more sense substantively and may consequently better serve the spirit of practical rationality.

Insisting on decision-theoretic consistency in our preferences may amount to insisting on preferences that make perfect sense in themselves but no sense at all *for us*. Once we understand the point of obeying the axioms of decision theory, we can see that the same point, appreciated more broadly, may be better served by violating them instead.