CHAPTER TWO: DECISION THEORY AND INSTRUMENTAL REASONING

2.0 Introduction

Any constructivist project in ethics stands in need of some account of practical reasoning, of the way or ways in which reason bears upon action, and how action can be or fail to be rational.[1] Instrumental reasoning is a promising place to begin, and in decision theory, there is an impressive and impressively developed body of theory which is often thought to amount to a rigorous, formal and systematic treatment of instrumental rationality. If decision theory is really a systematization of ordinary instrumental reasoning, we can help ourselves to its methods and results; if not, we face further questions about the relation between the two and especially about which, if either, trumps

[1] I shall speak interchangeably of what is rational and of what is reasonable, and correspondingly, of what is irrational and of what is unreasonable. I do not intend (as, e.g., Rawls does [1999, 315-317]) to mark any distinction between rationality and reasonableness.

the other in case of conflict.

What I shall claim is that decision theory does not map well onto ordinary instrumental reasoning, and moreover, that this is not due to defects in ordinary instrumental reasoning, but rather due to the failure of decision theory to capture some of its essential features. The power of decision theory for dealing with a variety of restricted contexts and specialized problems is considerable, but it does not amount to a general theory of instrumental reason. That is perhaps unsatisfying – one would like to have a rigorous and general theory – but I do not believe one has been worked out. It really is ordinary instrumental reasoning, disciplined but not displaced by theoretical results, with which we have to work.

I shall begin by circumscribing somewhat the field and issues I wish to address (I do not intend a comprehensive survey) and by settling some terminological points to facilitate discussion. Then, I shall outline the features of decision theory that will concern me and proceed to a more detailed examination. My principal concerns will be with maximization, the standing of the axiomatic conditions on preference, the force of decision-theoretic considerations for those whose preferences do not satisfy the axiomatic conditions, and the normative distinction between means and ends.

2.1 Preliminaries

Decision theory, so far as it will concern me, addresses itself to questions as to how it is rational to act, given a set of preferences, constraints and beliefs (or

expectations).[2] About the preferences, we shall have more to say as we proceed. The constraints are constituted by what the agent has to work with in acting or bringing his plans to fruition. The beliefs and expectations are those of the agent with respect to how the world is and what consequences will or are likely to follow upon different courses of action. These beliefs, except in the special cases in which, e.g., they are probabilistically incoherent, are generally taken as given by decision theorists and not, in any special way, within their competence. Whatever there is to be said about them with respect to whether the beliefs and expectations are correct or mistaken or as to whether they are arrived at in (generally) reliable ways, will involve other kinds of investigation (probably, many other kinds, since there is no unified field or discipline that addresses the correctness of what people take into account in making decisions).

Central to decision theory is the development and statement of conditions for the coherence of sets of preferences. No substantive conditions upon the content of preferences are assumed or presupposed, but sets of preferences, or sets of preferences in combination with sets of beliefs and expectations, can fail to be coherent. Conditions upon the coherence of preference sets are expressed by decision theorists in the form of

[2] Other applications, e.g., to modeling in economics and other social sciences, are beyond my scope.

Decision theory is sometimes contrasted with the closely related field of game theory by saying that game theory is addressed to issues of strategic interaction, to how it is rational to act in interacting with other rational agents where (at least part of) what has to be taken into account is the fact that one is dealing with agents who can themselves take into account the fact that they are dealing with rational agents. The distinction can also be drawn in other ways, e.g., by describing game theory as the part of decision theory dealing with strategic interaction or viewing decision theory as the part of game theory that is confined to "games against nature," i.e., in which actions of and interactions with other rational agents are not relevant. The issues I wish to address do not require venturing into game theory, so I will set it aside. (They are still, however, relevant to game theory, since game theory embodies the same underlying conception of rationality.)

axioms or postulates said to be definitive of the coherence of preference sets. Then, a rational choice is defined as one that is based upon – that is, either determined by or permitted by – a coherent preference set (again, in conjunction with beliefs and expectations, but I shall not keep repeating this).[3]

Different ways of axiomatizing the conditions on coherent preference have been offered. In some cases, these amount to different ways to reach the same destination. The different axiom systems turn out to impose equivalent conditions. In others, genuine alternatives to key axioms, such as Independence,[4] are proposed, and a preference set that qualifies as coherent under one axiomatization may not under another. As representative, I shall confine my discussion to standard expected utility theory.[5] There are two reasons. First, it is both the most familiar and the most widely relied upon; if anything deserves to be considered canonical in the field, standard expected utility theory is it. Second, the issues with which I am concerned emerge clearly in connection with it. To the extent that these issues are not raised by other axiomatizations, I see no need, as part of my current project, to address them.[6]

---

[3] A decision theorist might or might not allow that there is some other sense of rational choice that is not captured by whatever is the correct set of conditions on the coherence of preference sets. If he does, then he would treat the definition of rational choice in terms of the coherence of preference sets as defining some more restricted notion, rationality relative to preferences, perhaps.

[4] It is not important to define this here. I only note that Independence is critical for the possibility of defining a cardinal utility function (to be explained somewhat further later).

[5] Strictly, I shall be discussing both ordinal utility theory, which does not require the construction of an interval scale, and standard expected utility theory, which does. This is because ordinal utility theory is presupposed by expected utility theory (which is my major concern). For the present, I shall try to steer clear of the issues that arise when uncertainty (where information about probabilities is unknown or not available) as distinct from risk (where information about probabilities is available or assumed) affects the picture. Choice under uncertainty will occupy us at some length later.

[6] So far as the same issues *do* arise with other axiomatizations, what I say here will apply to those

2.11 Matters of Preference

Some discussion of the key notion of preference is in order. What is a preference, and what kind of role does it play?

Broadly speaking, there are two ways preferences have been conceived by decision theorists. First – an approach which has been most popular among economists – preferences have been taken to be *revealed* in choice and action. Someone who selects one option over another is said to prefer the option selected. Choice of one option over another is taken to be *criterial* for the existence of a preference for the option selected over the other. Second, preferences have been conceived as mental states of some kind that underlie and explain choice behavior. Importantly, the second conception allows for some slippage between preference and choice-behavior: the fact that something is chosen is not sufficient (though it may be good evidence) to show that it is preferred.

I think the revealed preference interpretation is inadequate. One reason is that it fails to track what we ordinarily mean when we speak of preferences. For example, if preference is what is revealed in behavior, no one can ever be indifferent between or among a set of options.[7] But a sartorial unsophisticate such as myself may simply open the drawer and take the first shirt that comes to hand. I may prefer taking the first that comes to hand over some other selection procedure, but not the shirt I take to others I could

as well.

[7] Modifying a revealed preference account to say that choice reveals either that one prefers the option selected or is indifferent between them (equivalent to the notion of weak preference explained below) is not attractive, for then no evidence from choice behavior could show that one is not indifferent between all options.

equally well have taken. Between the shirts, I am indifferent.

Additionally, it seems that a person can have a preference that is never revealed in behavior because the occasion for it never arises. I may have a preference as to which car (beyond my present means) I would purchase if I won the Publishers' Clearing House Sweepstakes, but if I don't win, I will never get to select between them. If I don't actually select one, though, a consistent revealed preference theorist would have to say I have no preference between them.[8]

That ordinary usage is not captured by a revealed preference perspective is not decisive, however. There might be theoretical advantages to using the term in a specialized way. So, a more important and more fundamental objection is that adopting the revealed preference view makes it impossible to identify any sets of preferences as failing to meet the coherence conditions specified in decision-theoretic axiom systems. As David Friedman remarks (speaking of the use of assumptions about rationality in economics):

> In order to get very far with economics, one must assume not only that people have objectives but that their objectives are reasonably simple. Without that assumption, economics becomes an empty theory; any behavior, however peculiar, can be explained by assuming that the

[8] The revealed preference theorist might instead hold that only choice behavior proves preference but deny that the absence of choice behavior proves the absence of preference. However, this can hardly be comfortable for him. Once preference is admitted to have some reality distinct from choice behavior and presumably at least partially available to introspective access, it will be hard to explain why choice behavior is always proof of a corresponding preference and why introspective reports to the contrary are always to be discounted.

> behavior itself was the objective. (Why did I stand on my head on the table while holding a burning $1,000 bill between my toes? I *wanted* to stand on my head on the table while holding a burning $1,000 bill between my toes.)[9]

The point is that any behavior can be represented as being in accord with preferences, provided that preferences are sufficiently weird, and so can any collection or sequence of behaviors. More generally, the coherence conditions of decision theory amount to imposing certain kinds of consistency requirements upon preference sets. The problem, if the requirements are to have any bite, is that no two (or three, etc.) behaviors that actually occur, considered apart from some description of them as, e.g., following a rule or aiming at an objective, can possibly be inconsistent with each other.[10] (If they were, they would not all occur.) If we adopt the revealed preference perspective, there are no obvious theoretical gains, and there is a significant theoretical loss: we lose the ability to distinguish between coherent and incoherent preference sets, and therefore, when rational choice is defined in terms of coherent preference sets, between rational and irrational choice.[11] So, I shall assume that preferences are mental states of some sort that underlie

[9] Friedman 1996, 4. Friedman is concerned with what assumptions about preferences and rationality are needed in order to make use of them for explanatory and predictive purposes. My concern is normative, but an analogous point applies.

[10] For that matter, there is no inconsistency between doing *A* because it promotes *C* and doing *B* because it prevents *C* unless there is also some assumption to the effect that relevant preferences are the same at the times of the respective performances of *A* and *B*.

[11] I suspect this may have been an *attraction* for some. "Why," Nozick (1997, 133) asks, "does Mises [who provided an unusually explicit statement of a revealed-preference perspective, in his 1963, 19-21, 94-96, 102-104] think it is so important to argue that preferences cannot be irrational? Perhaps because he doesn't want anyone interfering with choices on the grounds that they arise from irrationally structured preferences." If that is the reason, the move has little to recommend it, for, in the first place, we might wonder why, if the bar is set so low that everything qualifies, it is either important or desirable to protect

choice behavior and that we can (at least in some cases) know what a person's preferences are without awaiting their revelation in behavior.[12]

An important further point about preferences deserves our attention. When one speaks of ends or objectives (and similarly for goals, desires and wants), their content can typically be construed in terms of single-place predicates. There is some envisioned action or state of affairs that is the content of the objective; if the action occurs or the state of affairs comes about, the objective is realized. By contrast, preferences are explicitly comparative and must be construed in terms of two-place predicates: *this* is preferred to *that*.

In many ways, this injects a salutary dose of realism. John Broome makes the point nicely:

> Imagine you meet a thirsty person. She wants water, she wants Coca-Cola, she wants beer, and she has a great many other wants too. Suppose you know all her wants in great detail. You know she wants half a pint of water; she wants a pint of water; she wants a litre of beer; she does not want Coke and beer together; and so on. Given all that, what should you give her to drink? A pint of water? A half-pint of Coke? Coke and water together? Just from knowing everything she wants, you cannot tell.
>
> To know what to give her, you need to know her *comparative* wants. You

such "rational" choices from interference, and, in the second, any attempt to interfere would *also* qualify as rational.

[12] I do not intend to enter into any intricacies of the epistemology of preferences. I assume that agents have some (fallible) introspective access to the content of their preferences and that there are also various indirect ways of supporting or undermining claims about their preferences.

> need to know what she wants more than what. You need to know her preferences, that is. Her preferences put all her options in an order: a pint of beer above a pint of water, a half-pint of Coke and half-pint of water above a half-pint of beer, and so on. If you know all her preferences, and if we grant [that she should be given what she prefers], you know what to give her; you should put her as high up her preference order as you can. But knowing just her wants is not enough. (1999a, 9-10)[13]

The real choices that people make are rarely matters of just realizing an objective or not, with no other considerations brought to bear. Of course, we *do* take action to realize fairly definite objectives, but other considerations, almost inevitably comparative, are part of the background against which this or that objective is settled upon.

2.2 Outline of Utility Theory

With this background, I proceed to outlining utility theory.[14] So far as possible, I shall avoid technicalities, but it is useful to introduce some notation to capture the way in which preference is comparative. We can begin with the notion of *weak preference* ( ).

[13] "Similarly," Broome adds, "if you only know what is good for a person, even if you know everything that is good for her, that is useless. If you know everything that is good generally, that is useless too. You need to know what is better than what." (1999a, 10)

[14] Largely, I am following Shaun Hargreaves Heap's treatment in his article, "Rationality," in Heap, *et al*. 1992, which in turn follows the approach of von Neumann and Morgenstern. In this approach, probabilities are assumed to be objective and given to the agent. In the alternative approach worked out by Savage and others (Savage 1973), probabilities are understood as subjective degrees of belief and attributed to the agent on the basis of choice behavior. For some useful discussion, see Found 2001.

One thing is weakly preferred to another when it is at least as good as (alternatively, no worse than) the other in terms of a preference ranking. "*A* *B*" should be read as "*A* is weakly preferred to *B*."

In terms of weak preference, we can define both *strict preference*[15] and *indifference*. *A* is strictly preferred ( )to *B* when it is definitely better in terms of a preference ranking – that is, when *A* is weakly preferred to *B* and *B* is not weakly preferred to *A*. Or, *A* *B* when *A* *B* and it is not true that *B* *A*. *A* and *B* are indifferent (~) in terms of a preference ranking when it is true both that *A* is weakly preferred to *B* and that *B* is weakly preferred to *A*. Or, *A* ~ *B* just in case *A* *B* and *B* *A*.

The resemblance of this notation to that for comparing numerical or algebraic expressions (>, ≥, =, etc.) is, of course, not accidental. The idea of utility theory is to represent the relations between preferences by some kind of numerical index, so that certain important properties of the numerical relations will apply also to the relations between preferences.[16] Such an index can be used to represent a person's utility from the options available to her, and the property that it is important for such an index to have is that, from among her options (leaving aside ties), the largest number is assigned to the option she most prefers. Then, if she selects what she most prefers, since that is identical

[15] When I mean to be talking about a case in which one item is strictly preferred to another, I shall generally just say that it is preferred.

[16] That there exists some orderly way of assigning index numbers to the elements of a set of preferences (which presupposes that the relations between the preferences themselves meet certain conditions, to be discussed further below) is what is meant by saying that a utility function can be specified (for that agent, with those preferences). In essence, a utility function amounts to a mapping of the elements of a preference set onto a number line, and the relations supposed to matter on the number line may be either cardinal or ordinal. Different utility functions – i.e., different mappings – may preserve the same set of relations between the elements of a preference set; thus, one correct mapping will be some transformation of other correct mappings.

to selecting the option associated with the largest index number, she can be said to be maximizing utility (or, to anticipate coming elaborations, can be said to be maximizing expected utility).[17]

2.21 Ordinal Utility Theory

Plainly, for this to be workable, certain conditions on the relations between preferences will have to be satisfied. The numerical relations work because the numbers they relate have certain properties. Similar requirements apply when transposed to relations between preferences.

For present purposes, the most important are Transitivity and Completeness. Assume an agent has a set of preferences over some set of elements. What Transitivity requires is that, for any three elements in the set, *A*, *B* and *C*, if *A* is preferred to *B* (*A* *B*) and *B* is preferred to *C* (*B* *C*), then *A* must be preferred to *C* (*A* *C*).[18] If this condition were not satisfied, then she could regard *A* as (preferentially) better than *B*, but *C* as

[17] In order to avoid misleading suggestions, it is useful to remark briefly on the notion of utility employed here. The term 'utility' has historically been used in a variety of ways, and it is important not to confuse how it is employed in decision theory with others. What must be avoided is the idea that the utility maximized when the agent selects her best option is some separate object of her pursuit (perhaps a warm glow of satisfaction) distinct from other elements of her preference set, such as finding a job that matches her skills or a restaurant that serves good Thai cuisine. Naturally, it may be that some intra-psychic state, such as a warm glow of satisfaction, is among the elements of her preference set, but, if so, it will itself have to be assigned a utility index and will affect the utility indices that can be assigned to other elements. The warm glow will not be identical with her utility but will instead feed into assessments of the utility of her various options. To put it slightly differently, if the warm glow is an element in her preference set, it will have to be ranked *vis-à-vis* other elements as preferred, dispreferred or indifferent to them, and it may be that getting something else will have greater utility than the warm glow. Utility should not be understood as a separate object of pursuit but instead as a representation of preferences. To say that a person is maximizing utility is just to say that she is doing what she most prefers. See Broome 1999b.

[18] The other relations, weak preference and indifference, must also sustain transitivity relations, but the strict preference relation is the most important.

indifferent to or better than *A*.

What Completeness requires is that any element in the set can be ranked *vis-à-vis* any other. Take any two elements, *A* and *B*; then, it must be that the agent prefers *A* to *B* (*A* *B*), that she prefers *B* to *A* (*B* *A*) or that she is indifferent between them (*A* ~ *B*). If her rankings are not complete in this sense, then it would be possible that there is some pair of elements which are not comparable in terms of her preferences: neither is preferred to the other nor are they ranked equally. I think this is a real possibility, but for the present what is important is that if there is some pair of elements between which no preferential relation can be established, then, so far as the relation between those elements is relevant, it will not be well-defined what it is that best satisfies the agent's preferences (just as when, in numerical comparisons, we say that some term, such as *i*, the square root of negative one, has no place on the real number line, we cannot compare it to any real number and say, for example, that *i* is greater than, less than or equal to two).

When a set of preferences meets these conditions (and certain others[19]), then the elements over which the preferences range can be ordinally ranked – that is, they can be ranked from (preferentially) best to worst (including ties). We can then define an ordinal utility function for the preference set. This just means that we assign a number to each

[19] Strictly, two more conditions, Reflexivity and Continuity, are needed. Reflexivity requires that any element in a preference set be at least as good as itself. Continuity requires that, for any pair of goods in a bundle of goods, one of the two can be made marginally worse and the other marginally better in such a way that the resulting bundle would be judged to be, without change in preferences, indifferent to the first bundle. Satisfying Continuity would disallow any strictly lexical orderings of preferences. That is, in a preference set satisfying Continuity, it cannot be the case that there is no quantity of a good, *B*, that would compensate for some small reduction in another good, *A*. Continuity does not, however, rule out preference-orderings that are practically equivalent to lexical orderings. (There are actually two different Continuity axioms, the one just explained, which is needed for ordinal utility theory, and the other for its extension to expected utility theory. I briefly discuss the other Continuity axiom later.)

element in such a way that, for any two elements, *A* and *B*, (1) if *A* $\succ$ *B*, then the number assigned to *A* is greater than the number assigned to *B*, and (2) if *A* ~ *B*, then the number assigned to *A* is equal to the number assigned to *B*. So, if Jennifer prefers beer to Coke, is indifferent between Coke and Pepsi, and prefers either Coke or Pepsi to water, we could assign six to beer, three to Coke, three to Pepsi and two to water. The particular numbers assigned do not matter, and in particular should not be taken to imply that Jennifer likes beer twice as well as Pepsi or Coke one-and-a-half times as well as water. Any other set of numbers that preserved the same ordinal relations – e.g., 903, 902, 902 and 107 – would serve equally well.

The number assigned to each element can be called its utility index or can be said to represent its (ordinal) utility. Assume now that the elements in an agent's preference set are outcomes of action that are known with certainty,[20] and that the preference set is complete and transitive. Then we can correlate each outcome with a corresponding action that is sufficient to bring it about. Assume also that these outcomes are all that matter – in particular, that no preferences with respect to the actions as distinct from the outcomes affect what the agent would, all things considered, prefer.[21] If these conditions hold, we

[20] Of course, a person may have preferences over elements that are not outcomes of action (at least for that person) at all, such as whether a scientific theory is true or whether a past event happened in a certain way. If such preferences have no action-guiding import – if they do not, for example, shape the course of an investigation – then they can be set aside as not bearing on the rationality of action.

[21] This can often be secured by 'loading' preferences with respect to actions into the descriptions of the associated outcomes. For example, it may mislead to say that I am comparing a mowed to an unmowed lawn. The right description might be that I am comparing having a mowed lawn, together with being hot and sweaty, to having an unmowed lawn, together with being cool and comfortable.

Such loading of preferences with respect to actions into outcome-descriptions is not always possible nor is it always clear, where it is not possible, that the agent whose preferences resist such redescription is being less than rational – see Hampton 1998, Chapter 8 and Verbeek 1999 – but I shall, for the present, set aside such problems.

can give content to the earlier claim that a rational choice is based upon a coherent preference ordering and can say unambiguously what action is best in terms of the agent's preferences. We assign numbers to each outcome in a way that preserves the right relationships, and then the action correlated with the outcome with the largest utility index (or one of them, in case of ties) is the one that would best satisfy her preferences. Her best choice is the one that maximizes utility – or, in other words, is one that selects an outcome with the largest associated utility index.

2.22 Expected Utility Theory

Though ordinal utility theory may be useful for dealing with preferences over certain outcomes, it has significant limitations in that, in most of our choices and deliberations, we do not know with certainty what the outcomes of our actions will be. Thus, even if an agent has a complete and transitive preference ordering over outcomes, that is not generally sufficient to make it clear what action will best satisfy his preferences (or what action to take given that he cannot be sure the actual outcome will best satisfy his preferences).

What needs to be done to extend utility theory so that it has application in a world of risk? I will begin with an assumption about risk and then follow up with two terminological points. The assumption is that the risk an agent faces in selecting from among his options can be characterized in terms of probabilities which he knows or takes as given. If there is some outcome which the agent would prefer to all others, then he

knows for each of his options that it has some definite probability, a point-probability as it is called, of leading to the outcome he prefers and also has some definite probability or probabilities of leading to some member of a set of relatively dispreferred alternative outcomes.[22]

With regard to terminology, I have spoken frequently of the *elements* of a preference set as what preferences range over. The term was selected deliberately to avoid any pre-judgment as to what preferences could range over. In ordinal utility theory, it is most convenient to assume that the relevant preferences range over certain outcomes, but there is no necessity for this.[23] An initial step to incorporating risk would be to replace talk of certain outcomes with talk of *prospects*, which may be defined as outcomes combined with a measure of their probability.[24] More precisely, any prospect, *A*, is a gamble in which some outcome, *B*, is received with some probability, *p*, and the alternative to *B*, not-*B*, is received with the complementary probability, 1 - *p*. A complication to be borne in mind is that the "outcomes," *B* and not-*B*, may themselves be prospects, embodying further gambles over outcomes and so on. So construed, outcomes that are certain are a subset of prospects, namely those in which some outcome, that does

[22] This does not seem to be the only way in which an agent can be ignorant of what the future holds (see note 5), but one problem at a time!

[23] Indeed, in an earlier note (20), I mentioned the possibility that one might have preferences about which theories are true, etc.

[24] To avoid cumbersome locutions, I shall generally speak of *certain* and *risky* prospects, where a certain prospect is one in which an outcome is assigned a one-hundred-percent probability (this could also be called a non-compound prospect), and a risky prospect is one in which an outcome is paired with some probability less than one hundred percent. This terminology – as contrasted with the more natural 'uncertain prospects' – is adopted with a view to avoiding confusion over the distinction mentioned earlier (note 5) between uncertainty and risk.

not itself involve any further gamble, has a probability of unity.

Second, we can also generalize the notion of utility and replace it with *expected utility*, which can be understood as a representation of an agent's preferences over prospects. We can stipulate that, whatever the utility of an outcome received with certainty is, that is also the *expected* utility of the corresponding prospect – that is, one in which the relevant outcome has a hundred-percent probability of occurrence. Thus, in parallel to the above treatment of the relation between prospects and outcomes, the utilities of certain prospects are special cases of the expected utilities of prospects. This, however, does not get us very far. In particular, it does not license any inferences that the prospect in which an outcome has a fifty-percent probability has an expected utility equal to half that assigned to the prospect in which the same outcome has a hundred-percent probability. The reason is that the selection of a particular number to represent the utility of a given outcome in ordinal utility theory is quite arbitrary, so long as certain relations are preserved with the index numbers assigned to other outcomes. Thus, two equally good ordinal utility functions representing the same set of coherent preferences might exhibit very different proportional relations among its elements. If we cannot treat Jennifer getting a beer as satisfying her preferences twice as well as her getting a Coke, we also cannot treat her having a fifty percent chance of getting a beer as being just as good as getting a Coke. (That is, we cannot without further information about her preferences and their interrelations.)

If we are to deal with this within the framework of utility theory, some way needs

to be found of further regimenting the numerical representation of preferences and their relations to one another beyond the ordinal relations already allowed for. Somehow, we must find a way of cardinally scaling the numerical relations between representations of preferences. To get a better grasp on what is needed, let us look a bit more closely at Jennifer's problem. Assume that only her preferences with respect to what to drink are relevant and that getting a Coke, a beer or remaining thirsty are the only options. She would rather have a beer than a Coke and would rather have either than remain as she is (namely, thirsty), so getting a beer for certain has to be assigned a greater expected utility than getting a Coke for certain, and the certainty of getting a Coke will itself be assigned a greater expected utility than will be assigned to the certainty of remaining thirsty. A risky prospect of getting either a beer or else remaining thirsty will have some intermediate value, but it is not yet clear how that intermediate value will relate either to the certainty of getting a Coke or to any of the infinitely many possible risky prospects in which she either gets a Coke or remains thirsty.[25]

Now, it is initially plausible that answers to questions of the type just suggested can be reached on the basis of the agent's preferences, that preferences have *degrees of strength* with respect to one another rather than just being ordinal. Jennifer may say that she would *very much* rather have a beer than a Coke. It is also plausible that these degrees to which one thing is preferred to another extend to the ordering of risky prospects. If

[25] I am simplifying by considering only cases in which improvements over the *status quo* (without worsenings) are being considered. Thus, a risky prospect of an improvement over the *status quo* should also count as an improvement. If worsenings from the *status quo* were under consideration, then a risky prospect of a worsening (without any prospect of an improvement) should also count as a worsening, but a lesser one than the certain prospect of a worsening.

Jennifer much prefers having a beer to a Coke, then it is likely that she would also prefer, say, a ninety-nine percent chance of getting a beer to the certainty of getting a Coke. What is needed, then, is to establish a cardinal scale along which preferences can be represented that (a) preserves the ordinal relations between the certain or non-compound prospects in the preference set, and (b) also allows the comparison and ranking of any prospect, whether certain or not, with respect to any other.[26]

It turns out, if certain further conditions are imposed upon a preference set, that this can be done. The intuitive idea can be presented fairly readily. We arbitrarily assign numbers to a worst and to a best outcome for an agent (with the larger number being assigned to the best outcome, of course), which are, respectively, worse than or better than any of the actual prospects we wish to compare. Call the best *Bliss* and the worst, *Torture*. It should be understood that whatever Torture is, it is so bad that the agent, if given a choice between Torture and anything else, would select whatever is the alternative to Torture. Similarly, Bliss is so good that, when compared to anything else, it would be selected over that alternative. If we assign, say, the number zero to Torture and one hundred to Bliss, it is plausible that all the agent's other preferences, whether for certain or risky prospects, can be arrayed along a number line, preserving their ordinal relations to one another, somewhere between zero and one hundred. But where should each one – Jennifer's preference for getting a Coke, for example – be placed? Well, since Coke is

[26] Cardinal measures and comparisons do not presuppose a non-arbitrary zero-point (such as zero mass for cardinal comparisons of mass). They can be constructed for domains in which there either is not or is not assumed to be such a zero-point (as temperature scales were constructed before it was recognized that there is an absolute zero).

equivalent neither to Bliss nor to Torture, we have (using obvious abbreviations) the following relation:

$$B \quad C \quad T$$

Since numbers have been assigned to Bliss and Torture, can we assign a number to Jennifer's getting a Coke as some function of her preferences between Bliss and Torture? The first step is to recognize that there does seem to be a way of combining her preferences with respect to Bliss and Torture so as to get an intermediate value. Specifically, we can construct a risky prospect, a gamble, consisting of a probability mix of the two, where she gets Bliss with some non-zero probability and Torture with the complementary probability. That is, we say that the gamble gives her Bliss with probability, p, and Torture with probability, $1 - p$ (where $p \neq 0$). We can symbolize this gamble as:

$$[B, p; T, 1 - p]$$

This gamble has to be preferred to Torture and dispreferred to Bliss, because having some chance of Bliss has got to be better than the certainty of Torture and some chance of Torture must be worse than the certainty of Bliss. This is put to use in two ways. The first is to calibrate the scale between Torture and Bliss. Assume that Jennifer is considering two different probability mixes between Bliss and Torture, $[B, p; T, 1 - p]$ and $[B, p^*; T, 1 - p^*]$, where $p^* > p$ – that is, in which the second gamble gives her a greater probability of getting Bliss and a smaller probability of getting Torture than the first. It seems reasonable to suppose that she will prefer the second gamble to the first:

$$[B, p^*; T, 1 - p^*] \quad [B, p; T, 1 - p]$$

If this holds true for all values of p and p^*[27], then, for any value of p in $[B, p; T, 1 - p]$, we can assign to that gamble the real number along the zero-to-one-hundred scale that is equal to p x 100.[28] Then, every point along the scale will correspond to a value for p that itself appears in one and only one particular gamble of the form, $[B, p; T, 1 - p]$. Moreover, if we use these points along the scale to provide expected utility index numbers for the corresponding gambles, then all of the infinitely many possible gambles between Torture and Bliss will stand in the right ordinal relations to one another.[29] Every gamble that gives Jennifer a greater probability of Bliss (and a smaller probability of Torture) than some other will be assigned a larger index number than that other.

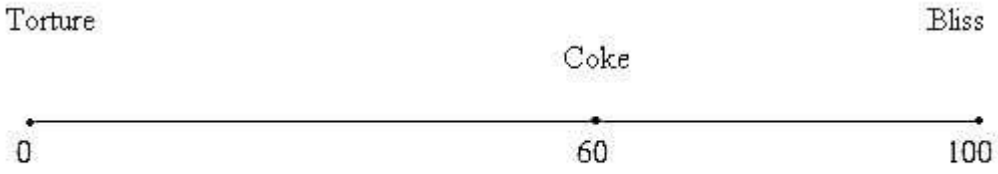
Once we have calibrated the scale in this way, the second use is to find out where along it to represent a preference for some particular prospect other than a gamble between the end-points, such as Jennifer's preference for getting a Coke. At this point, the idea is that, since we know a given probability mix between Torture and Bliss is, just like getting a Coke, preferred to Torture and dispreferred to Bliss, we can consult Jennifer and ask her whether she would prefer to have this gamble, for some specific value of p, or the certain prospect of getting a Coke. If she would rather have the Coke than accept the gamble, then the p-value is too small to represent her preference for getting a Coke. If she

[27] Again, where $p^* > p$.

[28] It was not accidental that I selected a zero-to-one-hundred scale. Those numbers were arbitrary in that others could have been used to anchor the end-points, but using zero for Torture and one hundred for Bliss simplifies the exposition.

[29] This is not the only possible way to secure these properties, just one that is simple and intuitive.

would prefer the gamble to the Coke, then the *p*-value is too large. In principle, if we present her with a large number of gambles between Torture and Bliss, we should be able to find one with respect to which she is indifferent between accepting that gamble and getting the Coke.[30] Then, the number on the Torture-Bliss scale that corresponds to that gamble can be assigned to represent her expected utility in getting a Coke. Then, we may get a graphic representation of the result that looks like this:

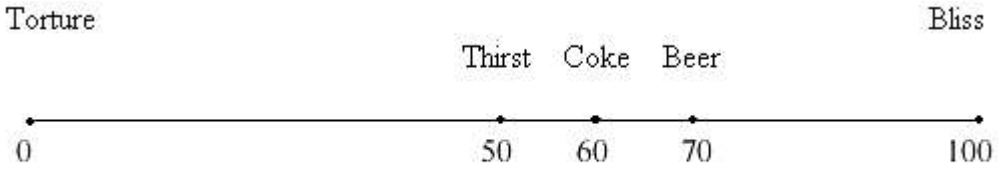


What this would mean is that, for Jennifer, C ~ [*B*, .6; *T*, .4] – that is, that she is indifferent between getting a Coke for sure and a gamble in which she has a 60 percent chance of Bliss and a 40 percent chance of Torture.

We repeat the process for Jennifer's other preferences, such as those for getting a beer or remaining thirsty, determining at just what gambles between Torture and Bliss she would be indifferent between those gambles and the respective certain prospects in which she remains thirsty or gets a beer, to find expected utility indices for those prospects as

[30] The "in principle" clause here is important since it may well be that Jennifer is getting more and more thirsty as we ask her questions about gambles between Torture and Bliss with differing *p*-values (rather than giving her a Coke!). As her thirst approximates Torture, she may be willing to accept gambles with lower *p*-values as being indifferent to getting a Coke. To be precise, we would have to say instead that, in a given situation, with given levels of thirst, etc., there is some gamble between Torture and Bliss, such that, *if* she were offered it, she would be indifferent between accepting it and the certain prospect of a Coke.

well. Suppose that, when we have done so, the slightly enriched graphic representation looks like this:



If we assume that Jennifer is thirsty and if all has gone well, we can not only rank her thirst, her getting a Coke and her getting a beer ordinally against one another, we can also determine, for example, that she would be indifferent between getting a Coke and a 50 percent chance of getting a beer or that she would prefer a 60 percent chance of getting a beer to the certainty of getting a Coke.

Generalizing a bit (and assuming the scale is worked out in the necessary detail), we can represent on the scale every certain prospect she faces. Further, we can represent all her preferences with respect to risky prospects as functions of the corresponding certain prospects on the same scale, and we can rank each prospect, whether certain or not, with respect to every other and assign to each an expected utility index. The index numbers assigned will have the property that any prospect strictly preferred to another has a larger expected utility index than the one to which it is preferred. Thus, Jennifer's choices, if she is actually selecting what she most prefers, maximize her expected utility. Further, by virtue of the possibility of constructing an expected utility scale that allows

ratio comparisons, that is, that allows us to say to what comparative *degree* her preferences are satisfied by different prospects, we are able to extend the reach of the idea that a rational choice is one that is determined or permitted by a coherent preference set to cover cases in which outcomes are not known with certainty.

Now, there are at least two reasons for being suspicious of what I have been doing in the last several paragraphs. The first is that I may have drastically over-simplified the kinds of options over which preferences range. It is not realistic to speak, say, of Jennifer's preferences with respect to beer as though it makes no difference what quality, temperature and quantity of beer is on offer, and so on. Nor is it enough to further specify in isolation the characteristics of the beer, for her preferences with respect to beer are also affected by how thirsty she is and what else she cares about in the present circumstances. This problem, however, is, in principle, easily remedied. What has to be realized is just that her preferences for beer, Coke and so on need a more fine-grained description if we are to be sure that *these* are the preferences which apply to her current situation. She need not have *general* preferences for beer over Coke over thirst. It suffices if she has preferences for beer (with the available characteristics) over Coke (with the available characteristics) over her actual degree of thirst in the current circumstances, when these highly specific preferences can be assigned expected utility indices.

The second concern is much more serious. I have alluded to further conditions on the coherence of preference sets, but have actually said very little about these further conditions. Instead, I have engaged in a good bit of hand-waving in the form of assertions

about what is plausible, what it is reasonable to suppose and the like. But *is* what I claimed plausible or reasonable to suppose? In particular, is it plausible or reasonable to suppose that they embody or exemplify *requirements* for the coherence of preference sets and therefore, since decision theorists define rational choice in terms of the coherence of the preference set from which it proceeds,[31] *requirements* for rational choice? To consider this question is a way of raising the issue whether and why the axiomatic conditions upon the coherence of preference sets are normative for choice. And that can hardly be approached without saying what the further conditions are.

A natural way to proceed at this point would be to spell out the further axiomatic conditions and then undertake to assess them, considering arguments in their favor as well as alleged counter-examples, but that is a well-trodden path and I shall take a different tack. I shall indeed *briefly* say something further, with only a minimum of commentary, about what the axiomatic conditions are.[32] But then I shall consider their normative standing from a different angle.

2.221 Additional Axiomatic Conditions for Expected Utility Theory

Four conditions have already been introduced in the discussion of ordinal utility theory: Reflexivity, Transitivity, Completeness and Continuity.[33] Six additional conditions

[31] See note 3 and accompanying text.

[32] I shall try to keep technicality also to a minimum, but some is almost unavoidable.

[33] Transitivity and Continuity, as introduced, were defined in terms of strict preference relations, but analogous conditions on weak preference and indifference must also hold. Similar analogous conditions are needed for the remaining conditions. (Technically, it is more elegant to state the conditions in terms of weak preference and derive the requirements for strict preference and indifference as needed. However, stating the conditions in terms of strict preference is more intuitive, since “strict preference” is

can be identified.[34]

The first of these can be called *Extension to Prospects*. What it says is that the first four conditions – Reflexivity, Transitivity, Completeness and Continuity – must also be satisfied by sets of preferences ranging over prospects. This seems just as reasonable as the initial conditions themselves since, when those were introduced, they were specified as applying to the ordering of elements of sets of preferences. To say that they apply to prospects is just to recognize prospects as elements of preference sets.

The second condition can be called *Preference Increasing with Probability*. What it requires is that for any pair of prospects, A and B, if A is preferred to B, then, for any pair of gambles over A and B, the gamble in which A is assigned a higher probability (and B a lower probability) is to be preferred to the other. That is, if $A \succ B$, then $[A, p; B, 1 - p] \succ [A, p^*; B, 1 - p^*]$ if and only if $p > p^*$. The idea is just that if some prospect is preferred to an alternative, then a larger chance of getting the prospect one prefers, rather than its alternative, has to be preferred to a smaller chance.

The third condition can be called *Closure*. What Closure requires is that, for any two prospects that are elements of a preference set, A and B, then, for any value of p, the gamble between them, $[A, p; B, 1 - p]$, is also an element of that preference set. That is, the elements of a preference set include any prospects that can be constructed from a gamble over other elements of the preference set.[35]

---

just what is ordinarily meant by "preference.")

[34] This is somewhat arbitrary. As noted earlier, there are different ways of axiomatizing decision theory, and what exactly the conditions are, and therefore what the precise number of necessary conditions is, need not be the same in different axiomatizations.

[35] Closure is included for expository convenience, but is not strictly needed since it can be derived

The fourth condition is sometimes called Continuity – somewhat confusingly, since there is already a Continuity axiom. I shall call it *Probabilistic Continuity*. Probabilistic Continuity says that for any three prospects, *A*, *B* and *C*, if *A* is preferred to *B* and *B* to *C*, then there is some probability-value, *p*, such that the gamble between *A* and *C*, [*A*, *p*; *C*, 1 - *p*], is indifferent to *B*.

The fifth condition can be called *Strong Independence*.[36] It "means that, in any prospect, any component object or prospect can be replaced by an object or prospect indifferent to it, and there will be indifference between the resulting prospects and the original one."[37] In other words, what it requires is that an agent who is faced with a gamble between two prospects, *A* and *B*, and for whom *B* is indifferent to some third prospect, *C*, should be indifferent between the gamble, [*A*, *p*; *B*, 1 - *p*], and the gamble, [*A*, *p*; *C*, 1 - *p*]. Since prospects involve gambles over outcomes that may themselves involve gambles over further outcomes, this means that more or less complicated gambles, between which the agent is indifferent, may be interchanged for one another in a course of deliberation and that refusal to accept such interchange of indifferent prospects marks a preference set as incoherent.

Finally, it is required that probabilities be combined in the normal way. Suppose that there are three prospects, *A*, *B* and *C*, and a compound gamble between them, [[*A*, *p*; *B*, 1 - *p*], *p** ; *C*, 1- *p**]. Then the probability of getting *A* is equal to *p* x *p**, the

---

from the other axioms.

[36] There are other, weaker, Independence axioms that figure in different axiomatizations. Where weaker versions of Independence are used, other axioms have to be strengthened to compensate.

[37] Heap, *et al*. 1992, 10.

probability of getting *B* is equal to (1 - *p*) x *p**, and the probability of getting *C* is equal to 1 - *p**. Accordingly, if there is some expected utility index assigned to each of *A*, *B* and *C* – represented respectively as *u*(*A*), *u*(*B*) and *u*(*C*) – then the expected utility of the gamble is equal to ((*u*(*A*) x *p*) x *p**) + ((*u*(*B*) x (1 - *p*)) x *p**) + (*u*(*C*) x (1 - *p**)).

2.3 Decision Theory: Some Limitations

Much can be said about the various axiomatic conditions, taken one by one. Some, such as Transitivity, are almost universally accepted, while others, such as the Independence requirement, have attracted considerable suspicion.[38] What I shall try to do, however, is to press a somewhat different question, why the axiomatic conditions on preference sets should be taken to be normative for choice.

An entry point for considering that question is to remember that the source of the mathematics used to define cardinal utility functions is in measurement theory. It had been shown that, in any domain of elements with certain properties, properties which can be specified by a set of axioms, cardinal measures and therefore cardinal comparisons could be defined. What von Neumann and Morgenstern did, in working out the mathematics of expected utility theory, was to transpose that set of conditions to the domain of preference. Thus, they were able to show how a set of preferences could be understood to meet the axiomatic conditions and so, how it was possible to develop a cardinal measure in terms of which preferences could be ranked and compared with one

[38] Still, it remains true that some form of Independence is accepted by most decision theorists.

another.[39]

Strictly, what had been shown was that meeting the conditions is sufficient for it to be possible to define a cardinal measure. So far as I know, the additional claim that meeting those conditions is also necessary for a cardinal measure has not been proven at all, but, for the sake of argument, let us suppose that to be true as well.[40] If it is, then of course it has to apply to sets of preferences, so the axiomatic conditions on preference sets are necessary to define a cardinal expected utility measure.[41] But why exactly are the conditions necessary to establish a cardinal measure also normative for choice? Or, to put matters the other way around, why suppose that action in accordance with a preference set for which no cardinal measure can be defined is *rationally* defective?

So far, this is only a question intended to raise a doubt, but it can be fleshed out a bit. Consider the following argument[42]:

---

[39] Dawes 1988, 150-151.

[40] If it is *not* true, then the case for saying that the (assumed) axiomatic conditions on preference sets are normative for choice is weakened, for there might be some other set of conditions that the same preference set satisfies that is also sufficient to define a cardinal measure. If we call the standard set of conditions, A_1, and an alternative, A_2, then it might be that some action (outcome, etc.) best satisfies preference in terms of A_1 but not in terms of A_2, or *vice versa*. It is also possible that, though satisfaction of the conditions of either A_1 or A_2 is sufficient to cardinally order a preference set, a given preference set may satisfy one without satisfying the other. Then, the question would arise as to which, if either, is normative for choice.

[41] Though not proven, this is made plausible by the fact that no one seems to have any idea how one might go about defining a cardinal measure when one or more of the axiomatic conditions is unsatisfied.

[42] This argument assumes that rational choice must range over cases in which risky prospects are relevant. Since some measure of ignorance as to what the future holds almost always infects the options among which we choose, this amounts to little more than saying that rational choice must apply to the options we face. (We *could* have a conception of rational choice which applies only to few or none of the choices with which we are actually faced, but it would be hard to explain why we should be much interested in it.)

(1) Unless the axiomatic conditions on preference sets are satisfied, we cannot have a cardinal measure of preference satisfaction.

(2) Unless there is a cardinal measure of preference satisfaction, there cannot be a determinate answer as to what best satisfies preference.[43]

(3) Therefore, there can be no rational choice (with respect to a preference set[44]) unless the axiomatic conditions are satisfied.

(4) And therefore, if rational choice is possible, the axiomatic conditions must be satisfied.

Plainly, those conclusions do not follow from the first two premises alone (which, for the present, I am granting). The first two premises could be true and the conclusions false if there could be a rational choice that does not determinately best satisfy preferences. The additional premise needed to derive the conclusion, (3), and the equivalent (4), would be something like:

[43] I stipulate that "what best satisfies preference" means "what satisfies preference at least as well as any alternative." In other words, ties are permitted at the top of a preference ordering. In the case of a tie for what best satisfies preference, any of the tied items may be chosen and would count as best satisfying preference. Additionally, for there to be a *determinate* answer as to which of a set of alternatives best satisfies preference, it must be true that an alternative that best satisfies preference is at least weakly preferred to any other. (There could be indeterminacy at the top of a preference ranking if there were a set of two or more options which could not be ranked with respect to the other[s], but each member of the set was ranked more highly than any option that was not a member of the set. This qualification is added to foreclose the interpretation that members of such a set of mutually unranked options could be said to satisfy preference at least as well as any alternative.)

[44] I shall not continue to repeat this qualification.

2a. Unless there is a determinate answer as to what best satisfies preference, there can be no rational choice.

That premise, however, admits at least three interpretations. The weakest is:

2a'. Unless there is a determinate answer as to which of a set of options best satisfies preference, there can be no rational choice among those options.

So understood, the premise may be unexceptionable. It can be taken to assert just that if the members of a set of options cannot be ranked as preferentially better than, worse than or equal to one another, there can be no decision between them on the basis of preference. But this is too weak to support either (3) or (4) in two respects. First, it does not imply that there is anything *irrational* about selecting such an option.[45] Second, it does not imply that members of the set cannot be ranked in terms of preferences against other options that are not a part of the set. There might be some option that is definitely better than any member of the set; if so, then it would be irrational to select a member of the set rather than that option. Or there might be some option definitely worse than any member of the

[45] It is relevant here that there are two different senses of 'rational.' It may, on one hand, mean *rationally required*. In that sense, there can be no preference-based rational choice among options that are not preferentially ranked. (For that matter, there can be no rational choice in *that* sense among options that are preferentially ranked, but ranked as indifferent to one another.) Or, 'rational' may mean *rationally permitted*. If rational choice is rationally permitted choice, then the selection of one from among a set of options that are not preferentially ranked may be rational. A case that selection of one of a set of preferentially unranked options is irrational would involve the claim that so choosing would be *rationally forbidden* or, alternatively, not rationally permitted.

set, in which case it would be irrational to select it rather than a member of the set. There could still be rationally required choice, even if not *among* the members of the set. In other words, (2a') is consistent with the existence of a partial ordering of the elements of a preference set.

So, consider a stronger interpretation:

2a". Unless there is a determinate answer as to which of a set of options best satisfies preference, there can be no rational choice among any options.

This is sufficiently strong to support (3) and (4), but, as it stands, it is implausible. For suppose there is a pair of options, *B* and *C*, which cannot be ranked preferentially against one another. From (2a'), and therefore from the stronger (2a"), it follows that there can be no preference-based reason for selecting one over the other. But suppose that the actual decision problem facing an agent is between *A* and *D*, that *A* is strictly preferred to *D*, and that *B* and *C* do not figure in his deliberations at all. Surely, then he could, and rationally should, select *A* over *D*. The fact that there are other options which he is unable to rank with respect to each other – options that, as it happens, he does not have to choose between – should make no difference. Once again, the possibility of a partial ordering of the elements of a preference set undermines the conclusions.

What is needed to support (3) and (4), then, is not precisely a stronger

interpretation – (2a") is already sufficiently strong – but something that makes explicit a claim that (2a") only presupposes:

> 2a'". Unless there is always a determinate answer as to which of any set of options best satisfies preference, there is never a determinate answer, and so, there can be no rational choice among any options.

What (2a'") rules out is the possibility that a set of preferences may be partially ordered. In effect, it asserts that a partial ordering is no ordering at all. I want to draw attention to two closely connected features, one almost explicit and the other implicit in (2a'"). The first is that rational choice depends upon there being an ordering of options, combined with the further claim that only a complete ordering is genuinely an ordering. Obviously, this is closely connected with the Completeness axiom.

But why should this be accepted? This leads to the second feature. The implicit answer to this question depends on the assumption that rational choice requires or is to be identified with *maximizing*, with selecting what is best in terms of all of one's preferences taken together. And the possibility of doing that depends upon one's preferences being such that every element in one's preference set can be unambiguously ranked against every other.[46] When the elements of a preference set include prospects, that is only possible for practical purposes if a cardinal measure of preference satisfaction can be

[46] Unambiguous ranking carries with it, of course, the requirement of Transitivity. If the ranking is not transitive, then, for at least some cases, the same option will be ranked as both better and worse than some other, depending upon the order in which it is compared with other options.

defined. Perhaps an infinite mind could (just) ordinally rank all prospects, including those defined as gambles over other prospects.[47] But no finite mind could follow suit. Suppose some agent prefers *A* to *B*, *B* to *C*, and *C* to *D*. How, without a cardinal measure, is she to rank the gambles, [*A*, *p*; *C*, 1 - *p*] and [*B*, *p**; *D*, 1 - *p**], where $0 < p < 1$ and $0 < p^* < 1$? A preference for *A* over *B* over *C* over *D* would not imply any particular preferential relation between the specified pair of gambles. The only remotely tractable procedure for a finite mind is to treat preferences over gambles as a function of preferences over outcomes comparable on a cardinal scale.

The earlier argument, then, can be reformulated in this way:

(1) Rational choice presupposes maximizing.

(2) Maximizing presupposes the possibility of a cardinal measure of preference satisfaction.

(3) The possibility of a cardinal measure of preference satisfaction presupposes that the axiomatic conditions upon the coherence of preference sets are satisfied.

(4) Therefore, rational choice presupposes that the axiomatic conditions upon the coherence of preference sets are satisfied.

[47] Strictly, a genuinely complete and transitive ordinal ranking of all prospects could always be represented by some cardinal function or other, but it would not need to be the case that any particular ranking of some subset of prospects was derived from the cardinal function.

That argument is certainly valid, but since I am willing to grant the second and third premises, everything important in it depends upon the truth of the first. What I shall do in much of the remainder of the chapter is to focus in various ways upon its truth and upon that of the closely related requirement that one's preferences completely and unambiguously order all of one's options. My general strategy will be to run the argument in reverse: to argue that it is implausible that our preferences completely order our options and therefore implausible that rationality requires *of us* that our actions be determined or permitted by a preference set that satisfies the axiomatic conditions. That strategy would have no prospect of success, however, if there were some other argument showing that satisfaction of the axiomatic conditions on preference sets was necessary for rational choice, so the first matter to attend to is whether any such necessity *is* established by other arguments.

2.31 Defending the Axiomatic Conditions: Three Approaches

Nobody supposes that the axiomatic conditions on preference sets amount to logically necessary truths about rational choice or, therefore, that it is contradictory to deny or reject one or more of them. Nor is it supposed that it would be contradictory to deny one while keeping the others. If it were, then the one denied could be derived from the others and so would not be needed as an independent axiom.[48] The actual defenses

[48] This may be slightly misleading. There are first, as has been noted, different ways of axiomatizing expected utility theory and some may be more or less compact in terms of the number of axioms relied upon than others. (For example, I included Closure among the axioms, though it can be derived from the others.) Second, there is theoretical interest in seeing what the most compact or minimal set of axioms necessary is. However, it remains true that, for whatever is the most compact set of axioms,

offered for the axioms fall into three classes, which are not often clearly distinguished from one another.[49] First, it may be claimed that the axioms are, individually, intuitively secure or compelling or that they can be derived from something which is. Second, a coherentist defense can be mounted to the effect that accepting the set of axiomatic conditions makes the best sense of our pre-theoretic practice and of our assumptions and convictions about rational choice. Third, and most interestingly, a pragmatic case can be developed that we will do better if our preferences and choices conform to the axioms.

2.311 The First Approach: Intuitive Security

A defense in terms of the intuitive security of the axioms runs into two sorts of empirical problems. One, which has been demonstrated by various psychological studies, is that situations can be constructed in which people systematically make choices that violate the axiomatic conditions. By now, there is a large literature on such violations.[50] Regularly and predictably, people choose in ways in which they would not if their preferences conformed to the axioms. It is an important fact that these violations are systematic. It is not just that mistakes are made about what is rationally preferable. That might be explained in any of several ways, including lack of time to work out what is best, insufficient familiarity with relevant procedures and the like. Then, however, one would

each of them (and each proper subset of the axioms) is logically independent of and therefore not derivable from the rest.

[49] I am idealizing in describing pure cases of each of the approaches. It may be doubted whether anyone advocates taking any one of these approaches, to the exclusion of the others, to defending all of the axioms. I shall later briefly address the possibility that the best defense of the axioms consists of some appropriate combination of the different approaches.

[50] See, e.g., Camerer 1995 and Kahneman and Tversky 1990.

expect a random distribution around the correct answer. But when the violations of the expected utility axioms are systematic – when test subjects tend to converge upon the *same* alternative to what would be required by the axioms – that suggests that something deeper, such as the common rejection of some axiomatic condition, is responsible for test-subjects' choices. Moreover, the systematic violations persist to some substantial degree even when the decision-theoretic arguments for an alternative choice are explained. The fact that the violations are systematic, even on the part of test-subjects who have been exposed to the arguments for an alternative, suggests that most people do not find the standard axiomatic conditions to be intuitively secure or compelling nor do they see them as flowing from something which is.[51]

The second empirical challenge is a more specialized version of the first. It can be called the challenge of the experts. Significant numbers of people who have as good a claim as anyone to expertise in decision theory find that, in certain decision problems, they are inclined to choose in ways that violate some axiom, usually some version of Independence. This inclination is often reflectively stable in people who have considered and understood all that can be said on behalf of conforming to the axioms.[52] Even if it might be said with some color of plausibility to be unsurprising if ordinary people, untrained in decision theory, have unreliable intuitions, it is far more difficult to make it

[51] An analogy: Suppose a simple arithmetical problem, such as 24 + 27, were posed to an elementary school class. It would not be terribly surprising, and would be less surprising the younger the students, if a large percentage got the answer wrong. But it would demand explanation in terms other than simple error if the majority of the class agreed upon a *particular* mistaken answer, such as 45. The need for such explanation would be still more patent if it was explained to the class why the answer was 51 and most of them *still* gave 45 as the answer.

[52] Two important examples can be found in Allais 1990/1979 and Ellsberg 1990/1961.

credible that experts are unreliable in the same way. An appeal to the intuitive security of the axioms seems to lead only to a contest of divergent intuitions.

2.312 The Second Approach: Coherentist Defenses

The second or coherentist defense seems to be in no better shape, and for essentially the same reasons. As E.F. McClennen puts it:

> What one hopes for … are starting points that command nearly unanimous acceptance, at least among thoughtful and knowledgeable researchers. Unfortunately, [the axiomatic conditions] do not appear to meet this test. [They] have been the subject of sustained, spirited and thoughtful questioning by a number of decision theorists. Thus, they appear to be unsuitable starting points. But this last consideration would also seem to work against any "coherentist" argument as well. The principles in question do not codify the choice behavior of competent or even expert decision makers. (1990, 4)

2.313 The Third Approach: Pragmatic Defenses

Pragmatic defenses are the most promising and have in one form or another often been offered in defense of various axiomatic conditions, but in the end they are also inadequate. Let me be clear what I am claiming here. It is not that *no* pragmatic defense of *any* axiom is *ever* adequate. Whether it is or not will depend on the details of the case.

Rather, it is that it cannot be the case that *all* of the standard axioms can be given a pragmatic defense that does not depend upon particular preferences or relations among preferences for the person to whom the defense is offered. The satisfaction of some conditions – importantly including some conditions upon the agent's rationality – must be assumed to be in place before-hand.[53]

Since a pragmatic defense claims that we do better if our preferences and choice behavior conform to the axioms, the question I shall press is: Better *how* or in terms of *what*? I shall consider this in two stages. In the first, to illustrate some of the important points, I examine a paradigm of pragmatic argument in defense of an axiom, the 'money pump' argument in favor of Transitivity. In the second, I consider more generally what pragmatic defenses of the axioms can be expected to do.

Suppose I have intransitive preferences over three kinds of fruit: I prefer apples to bananas, bananas to cantaloupes, and cantaloupes to apples. Suppose also that I have a cantaloupe. Since I would rather have a banana than a cantaloupe, you can induce me to pay you some small sum to exchange the cantaloupe for a banana. Once I have the banana, you can induce me to pay a small sum to exchange the banana for an apple. Once I have the apple, you can induce me to pay a small sum to exchange it for a cantaloupe. I'm back where I started, with the cantaloupe, except that I'm poorer. Even worse, if my preferences over fruit remain the same, you can repeat the cycle as many times as

[53] It is important also to realize that a pragmatic defense may turn out *not* to be a defense of the standard set of axioms. McClennen 1990 is, among other things, an extended pragmatic argument against the Independence axiom. (Or better, it is an extended pragmatic argument in favor of what McClennen calls 'resolute choice,' which, in certain circumstances, commits the resolute chooser to violations of Independence.)

necessary to take all the money I have. (If you know those are my stable preferences and also how much money I have, you might even be well-advised to give me the cantaloupe if I don't already have one!) This is a version of the well-known money-pump argument for Transitivity.

The point of the story can be generalized. If my preferences are intransitive, I can be manipulated by others so that I inevitably end up worse off. Additionally, even without deliberate manipulation by others, sequences of events are possible in which, choosing on the basis of my intransitive preferences, I inevitably end up worse off.[54]

How compelling is this argument? Though it has considerable appeal, I am convinced that it is flawed. Its appeal derives from widely shared assumptions rather than from the strength of the argument itself.[55]

Consider again the case in which I have intransitive preferences over fruit and repeatedly pay to exchange a less preferred for a more preferred fruit. Why must I think that the result of such a sequence of exchanges is that I end up worse off?[56] If there is an answer, it appears that it would have to rest on the fact that I am assumed to have *transitive* preferences with regard to something else – in this case, with regard to quantities of money.

---

[54] It is of course not essential to the structure of the argument that the losses made inevitable by my intransitive preferences be monetary.

[55] The argument I present, though developed independently, closely parallels the one in Hampton 1998, 244-247.

[56] A related point has been made by Loren Lomasky (in discussion). Why, he asks, could not the person with intransitive preferences argue that he is made better off by each exchange and therefore by all of them? After all, for any of the exchanges, had he not preferred making that exchange, he would not have done so. So, why can he not regard the money as well-spent (and the fruit well-exchanged), since he got something he preferred each time?

And this is not all, for I could have transitive preferences with regard to quantities of money and *still* think I am not made worse off by the series of exchanges. In particular, I would not think I had been made worse off if I transitively preferred less money to more. Then, I would regard the series of exchanges as improving my financial condition as well as, each time, replacing a less preferred with a more preferred fruit.[57]

But suppose I have more normal preferences with respect to money and transitively prefer more to less. Even so, this is not by itself enough to show that my intransitive preferences over fruits are in need of revision. I would have a set of preferences over quantities of money *and* a set of preferences over different fruits. The argument will only work if I must or think that I must regiment my preferences over fruit in terms of the preferences over money. But why must I do that? For all that has been said so far, the regimentation could go in the opposite direction – that is, if regimentation there must be, I could adjust my preferences over money so they didn't interfere with the fruit exchanges in which I wish to engage.

So, what is needed for the money-pump argument to work is that, in addition to some set of intransitive preferences ranging over some domain, (a) the agent also has a set of transitive preferences ranging over some other domain, (b) that acting on the intransitive preferences insures that the transitive preferences invoked will be frustrated, and (c) that he considers it *more* important to satisfy the transitive preferences than those in the intransitive set. If those conditions hold, he will, if he can, adjust the intransitive

[57] Perhaps, if these were my preferences, I could be manipulated into worsening my financial condition by being required to accept a gift of a small sum of money with each fruit exchange!

preferences to make them transitive as well.[58]

But, this cannot work as a *general* argument for imposing transitivity upon one's preferences. For any given domain over which one's preferences intransitively range, the argument can provide a reason for imposing transitivity there only if there is some *other* domain in which one's preferences are already transitive (and with respect to which the other conditions are met). If there is no other domain within one's total preference set having the required characteristics, then no money-pump argument will work.[59]

Though I think this consideration of the money pump argument illustrates some important points, it is worth thinking further and more generally about pragmatic defenses of the axioms. A pragmatic defense of an axiom holds that we will, in some way, do better if we conform to it than if we do not. But how will we do better? So long as the

---

[58] Strictly, still more conditions are needed. The agent would also need to be aware that he has an intransitive preference set and would need to think that the risk of being manipulated because of it was worth the trouble of trying to change it. (Suppose he had an intransitive preference cycle over a thousand elements. First, he would probably not be aware of it, and second, even if he were, might think it unlikely that anyone would be able to find and exploit the intransitivity.)

Also, if his judgment that it is more important to regiment the intransitive preferences by the transitive than *vice versa* is itself modeled as a preference – say, a preference in a domain ranging over options of preference-revision – then the preferences in *that* domain will also have to be transitive.

[59] It might be objected, without contesting the details of my line of argument, that my conclusion has little or no practical importance, given the preferences people actually have. It is, after all, extremely common for people to have transitive preferences for greater over lesser quantities of money and for them to take the fact that some course of action is sure to lose money, without any off-setting gains, as a decisive consideration against it. (Does the fact that I, with my intransitive preferences over kinds of fruit, am *pleased* to make each exchange count as an off-setting gain?) It might be said that as long as this (or some analogue) is true of the preferences people actually have, it hardly matters that money-pump arguments are not decisive for all possible preference profiles: It is enough if they are decisive for actual preference sets. They may still show that all of *us*, with the preferences we actually have, should regiment our preferences into transitive relations. Transitivity may be rationally, because pragmatically, binding upon us without being rationally binding for all logically possible agents.

As an objection, this is misconceived, for it concedes the point that the argument for Transitivity depends on the character of other preferences we happen to have. It just adds that we *do* happen to have the other preferences needed.

theorist urging the pragmatic defense adheres to the decision-theoretic orthodoxy that there are no substantive requirements upon preferences, and so holds also that decision-theoretically rational choice does not presuppose that we hold particular preferences, the only option for the pragmatic defender would appear to be to claim that, by conforming to the axiom in question, we will do better in terms of our preferences.

But this is deeply problematic. To see why, recall some of the results already reached, namely, that satisfaction of all the axiomatic conditions is necessary to define a cardinal measure of preference satisfaction, and that the possibility of a cardinal measure is intimately connected with the existence of a complete ordering over options. Specifically, if there is a cardinal measure, then there must be a complete ordering. A slightly weaker claim was defended earlier about the converse relation: *for a finite mind*, if there is a complete ordering, then there is a cardinal measure. So, for a finite mind, there is a cardinal measure if and only if there is a complete ordering.

The problem we were considering is whether the person addressed by a pragmatic defense does better in terms of his preferences to conform to the axiom in question. But, *ex hypothesi*, his preferences do not satisfy all the axiomatic conditions upon preference sets. There are two possibilities. The first can be disposed of quickly. If a cardinal measure is needed to show that he does better in terms of his preferences, the argument fails: it will not be the case that conformity to the axiom would better serve his preferences,[60] since no cardinal measure can be constructed unless all the axioms are

[60] Of course, this does not imply that he would do worse by conforming or equally well by not conforming. It just means that, so long as his preferences do not satisfy the axioms, there is no determinate answer.

satisfied.

The more interesting alternative is to deny that a cardinal measure is always needed to show what is better in terms of a set of preferences; at least sometimes, it is possible to show that one option is better than another in terms of a set of preferences without relying upon the existence of a cardinal measure. If this is assumed, then a pragmatic argument in defense of an axiom will succeed just in case conformity to the axiom is better in terms of the agent's preferences than non-conformity. That is, it will succeed when his preferences suffice to order the options of conformity to and non-conformity to the axiom and rank the first above the second.[61]

Suppose the pragmatic defense succeeds. This means the agent's preferences are such as to unambiguously order conformity over non-conformity to the axiom. There are two noteworthy implications of this fact. The first is that, since the agent addressed by the pragmatic defense does not already satisfy all the standard expected utility axioms and therefore does not completely order options in terms of his preferences, it is possible to have an ordering of preferences which is only partial but nonetheless sufficient for rational choice, at least within certain domains or over certain sets of options. This means, among other things, that the argument sketched earlier for the axioms from the premise that rational choice presupposes maximizing must be mistaken. If any pragmatic argument at

[61] There are still of course ways in which a pragmatic defense of some axiom might fail. One is that the agent's preferences may be so disordered that none of his options in fact *are* ordered by his preferences. His preferences, so to speak, point in all directions and therefore not in any. Another is that though his preferences order some options, they do not order the options of conformity and non-conformity to the axiom. Still another is that, though conformity and non-conformity may be ordered by his preferences, conformity does not actually get ranked above non-comformity.

all works, then rational choice does not presuppose maximizing. Or, by contraposition, if rational choice presupposes maximizing, then no pragmatic defense of any of the axioms works. One cannot coherently hold both that rational choice presupposes maximizing and that pragmatic arguments can be given in defense of any of the axioms.

The second is that, in order for any pragmatic defense to succeed, the agent's preferences must already meet certain conditions. If any of those conditions are themselves to be defended as normative for choice, their defense must be conducted on other grounds. It is not possible in principle that a pragmatic defense can be given for all the axioms of standard decision theory,[62] or for that matter for all the axioms of whatever alternative to standard decision theory a particular pragmatic defender favors. Every pragmatic defense works, if it does, by relying upon the fact that the preferences of the agent to whom it is addressed already meet certain conditions.[63]

2.314 Can the Approaches be Combined?

In summary, none of the common approaches – not an appeal to what is intuitively secure, not a coherentist defense and not a pragmatic defense – is capable of establishing

[62] Since every pragmatic argument rests on some assumptions about conditions met by the preferences of the agent to whom it is addressed, it might be that a pragmatic case can be made for conformity to *each* of the axioms without its being the case that a pragmatic argument can be made for *all* of them. The case for conformity to one would presuppose the satisfaction of certain conditions; the case for another would presuppose satisfaction of a different set of conditions, and so on.

[63] It might be thought that in principle a pragmatic defense either of an axiom or of the full set of axioms could be thoroughly dispositive. For suppose that there is some logically exhaustive way of characterizing sets of preferences, such that, say, all preference sets satisfy one of the sets of conditions, C_1, C_2 or C_3. Then, an argument for some axiom, A_1, might proceed to show that conformity to A_1 is pragmatically better relative to each of the sets of conditions the relevant preference set might satisfy. This apparent possibility, however, is an illusion, for it is surely logically possible that a preference set be such that it does not rank conformity and non-conformity to the axiom at all.

that all of the axioms of standard expected utility theory are genuinely *requirements* upon the coherence of preference sets or, therefore, upon rational choice.

A natural question, then, is whether the different approaches can be combined in some way, so that what cannot be secured by any of the approaches operating individually is secured by their judicious joint application. It might be, for example, that some subset of the axioms is intuitively secure and that arguments of other kinds can be given for the remainder. It might even be thought that something like this is, in fact, somewhat inchoately at work in leading to the conviction, on the part of many decision theorists, that the full set of expected utility axioms is normative for choice. I do not know whether an argument of that sort can be adequately fleshed-out. Perhaps it can be. Still, it has not *been* done, so far as I know, especially not with careful attention to and distinction between what is taken to be intuitively secure, what is to be defended on the basis of broader coherentist considerations and what, given any previously defended conditions upon preference sets, is to be given a pragmatic defense. If it can be done, I would like to see the argument.

2.32 The Standing of the Axioms

In light of the foregoing, what can be said of the normative standing of the standard expected utility axioms? In their favor are both their mathematical elegance and tractability, plus the not inconsiderable attraction of the point that *if* one's preferences satisfied the axioms, it is very difficult to see how it could be denied that an action with a

larger expected utility index is better in terms of one's preferences than any action with a smaller expected utility index.

A further attraction for some is perhaps better described as a motivation than as a reason for accepting the expected utility axioms. It derives from the thought that there must exist some procedure which amounts to an algorithm for resolving any decision problem with which one might be faced.[64] Applying the algorithm may be difficult in practice for any number of reasons, but, in principle, there is always a correct answer to be found, and it always *can* be found by correct application of the algorithm. For reasons that are not clear to me, some find the alternative that there are no algorithms in a given domain, but that we can sometimes identify and avoid mistakes or can sometimes find correct or better answers to questions posed within the domain, to be unacceptable. They are willing to tolerate any amount of practical difficulty in coming up with the correct answers, so long as they do not have to admit any theoretical indeterminacy in what the correct answers are nor that they lack methods for finding the correct answers.

What can be urged against the axioms is just the fact that, at the current stage of discussion, there is no adequate defense of the entire set of expected utility axioms. For an agent whose preferences do not satisfy the axioms, it cannot be simply a foregone conclusion that his choices, in light of his preferences, are less than rational. Of course, they may be less than rational, but pointing only to non-conformity to an axiom is not

[64] I take the existence of an algorithm for any decision problem to imply that there is some decision procedure which is sufficient to pick the best option, or one tied for best, out of any set of options that may be available in the given decision problem. A procedure that sometimes failed to do so, or sometimes failed to rank any option as best or tied for best, would not, in this sense, be an algorithm.

sufficient to establish the fact.

I think the most we can do at this point is to treat the claim that rational choice must proceed from a preference set that satisfies the expected utility axioms as an *hypothesis*. If the hypothesis is that it is both necessary and sufficient for a choice to be rational that it be based upon a preference set that satisfies the axioms,[65] then the hypothesis could in principle be tested in either of two ways. On one hand, we could try to find some case in which a choice based on such a preference set fails in some way to be rational. On the other, we could try to show that there are rational choices that are not based on such a preference set. For either kind of test, apparent failure will tend to count in favor of the hypothesis and apparent success against it.

For my purposes, I shall set the first kind of test aside[66] and concentrate entirely upon the second. Specifically, I shall argue that it is virtually certain that our preferences do not in fact satisfy all of the axiomatic conditions, but that this does not (nor does anything else) show that we do not or cannot make rational choices. Part of this I take to be obvious and uncontroversial: we can and at least sometimes do make rational choices, choices that are better than their alternatives in terms of our preferences and objectives.

---

[65] A slight qualification is needed. It might be objected that the orthodox decision theorist would surely admit that a choice can be rational if based on a set of preferences ranging only over certain outcomes (and where only certain outcomes are relevant), when the preference set satisfied only the axioms of ordinal utility theory rather than the full set of expected utility axioms. This is true, but easily side-stepped. For the kind of case described, satisfaction of the expected utility axioms is sufficient but not necessary for what an orthodox decision theorist will recognize as a rational choice. For most cases, however, risky outcomes are relevant, and it is only those cases that concern me here, so the hypothesis can be expressed as the claim that, where risky outcomes are relevant, it is both necessary and sufficient for a choice to be rational that it be based upon a preference set that satisfies the expected utility axioms.

[66] It is difficult or impossible to find uncontroversial examples. Any proposal is likely to be met with the claim that the choice in question, though it may be counter-intuitive, is nevertheless rational.

For that, I intend to offer no further argument than has already been presented. The other part, that we do not satisfy all of the decision-theoretic axiomatic conditions, requires more elaborate support. I shall begin with further consideration of the requirement that the elements of a preference set be completely ordered.

2.321 Completeness and the Inscription Thesis

One of the axioms of expected utility theory is that the elements of a preference set must be completely ordered.[67] It must be possible to rank any element with respect to any other as preferred, dispreferred or indifferent to that other element. This is necessary in order to define a cardinal measure of preference-satisfaction and, consequently, for maximization in terms of the preference set to be well-defined.

If we are going to maintain that the Completeness condition is satisfied or may be satisfied for our actual sets of preferences, then a closer look at what is involved in having a preference is needed. For instance, we might suppose that an agent has a preference between two elements of his preference set, *A* and *B*, just when he has considered *A* and *B* together and either ranked one above the other or else ranked them as indifferent to one another. But if so, then it is clear that agents such as ourselves do not have complete orderings over our preference sets. There are innumerable pairs of items which are elements of our preference sets – that is, both of which enter into some preferential

[67] In what follows, I assume (as mentioned in note 46) that Completeness is not satisfied unless Transitivity is as well. Technically, however, the two requirements are independent. A preference set might be completely but not transitively, or transitively but not completely, ordered. Since I take it that there is little doubt that a coherent preference set must be transitive, I have chosen, except where it might make some difference to the argument, to speak only of Completeness.

relation or other – but the members of which have not been compared to each other. I may prefer chicken over fish for dinner and Jones over Smith in the municipal election, but may never have considered whether I would prefer Jones's victory to chicken for dinner.

Generalizing a bit, if, in order to satisfy Completeness, I must explicitly compare each element in my preference set to each other, then, for three elements, *A*, *B* and *C*, I need to perform three comparisons: *A* to *B*, *A* to *C* and *B* to *C*. For four elements, six comparisons are needed, for five elements, ten comparisons, and so on. Evidently, unless the number of elements is small, this is going to quickly get out of hand. For example, for 50 elements, 1225 comparisons would be needed.[68]

The point is even more obvious when we consider risky prospects. For between any two prospects in which, for the sake of argument, some outcome is assigned a probability of one hundred percent, such as having chicken for dinner (*C*) or Jones's electoral victory (*J*), there are infinitely many gambles, corresponding to the infinitely many possible values of p (with $0 < p < 1$) in $[C, p; J, 1 - p]$. Before, for some finite number of elements, the task of comparing them, if the number of elements is large, was (merely) forbiddingly difficult. Here, for Completeness and Closure both to be satisfied, each of the infinitely many gambles between chicken for dinner and Jones's victory must also be preferentially ranked with respect to every other element of the preference set (including every other gamble over any other elements of the preference set!). So, if

[68] In general, for n elements, the number of comparisons needed is equal to the sum of the integers from zero to $n - 1$.

preferential ranking presupposes explicit comparison, the task is, for finite minds (over finite periods of time), strictly impossible.[69]

Taken together, these facts imply the following disjunction: Either we do not (and *cannot*) satisfy the Completeness condition, or else, preferential ranking does not presuppose explicit comparison. Accordingly, if we assume that it is possible for us to satisfy the Completeness condition, we must also assume that preferential ranking of the elements of a preference set is possible without explicit comparisons. It can be true that an agent has some preference ordering over prospects that she has never considered together or, for that matter, has never considered at all.[70]

Let us see what this implies. If we assume that some agent's preferences satisfy the Completeness condition (together with the other axioms), the most obvious and most important implication for my purposes is something that I shall call *the inscription thesis*. The inscription thesis holds that the preferences or preferential relations involved or expressed in explicit comparisons have an underlying structure, not necessarily attended to but which is nonetheless present within those preferential relations, and which is sufficient to determine the remaining preferential relations between all the elements of the preference set.[71] The preferences involved in explicit comparisons have, inscribed within

[69] I assume that there is some minimum, non-zero, time required to perform an explicit comparison.

[70] She may never have considered either of a pair of prospects at all because the presence of each as elements in her preference set is guaranteed by the Closure axiom. She may rank *A* over *B* and *C* over *D*, by virtue of having explicitly compared them, but never have considered, for specific values of p and p^*, either $[A, p; B, 1 - p]$ or $[C, p^*; D, 1 - p^*]$. Nonetheless, for Completeness to be satisfied, it must be true that she has a preferential ranking between those two gambles.

[71] By "sufficient to determine," I mean that the underlying structure has features such that there is a unique answer as to what the further preferential relations are.

them, so to speak, all the preferential relations among all the elements of the agent's preference set. The preferential relations of the options explicitly considered embody already strengths or degrees or weights that can be compared to one another. Some limited number of explicit comparisons has been performed and preferential rankings between the items compared have been established, but somehow there can (truly) be ascribed to the agent a complete ordering over all the elements of her preference set, including those that have never been explicitly compared.

The inscription thesis appears to me very doubtful, and in what follows, I shall try to cast further doubt upon it. But before doing that, I will address two lines of defense that might be offered.

2.3211 Two Defenses of the Inscription Thesis

A defense of the inscription thesis might be grounded in the claim that if the agent were presented with a choice between any pair of the elements in her preference set, including among those elements all of the gambles that can be defined over other elements, then she would make some choice or other.[72] In this form, the proposal faces crippling objections. One is that unless determinism is true (or true for the conditions under which she is supposed to make the choice), it may simply be false that she would make some definite choice. She would make some choice or other, but there may be nothing about her or her situation that settles which choice she would make. And, even if determinism does hold for the conditions under which she is supposed to be choosing, it

[72] Dawes (1988, 154-155) suggests something like this in defense of Completeness.

may be true that she would make some definite choice between the options presented to her, but it does not follow either that she prefers that option to the other or that she is at least indifferent between them, unless we assume, what we have already rejected, some version of a revealed preference theory. The fact that she makes a certain choice is not sufficient to show that she has at least a weak preference for what is chosen over what is not, for it may be that the counterfactual, 'if she were presented with a choice between A and B, she would select A,' is true, but true in virtue of some feature of her situation other than her preferences.

So, the suggestion has to be amended to say that if she were presented with any pair of elements in her preference set (whether for choice or not), she would have some preferential ranking between them. Now, it does not seem obvious to me that this is true, but even if it is true, that is still not sufficient to underwrite the inscription thesis. For there is still the possibility that, if presented with such a pair, she would then *form* a preferential ranking between them – perhaps some definite preferential ranking, not just some preferential ranking or other – but that the preference she then forms is not determined by her pre-existing set of preferences.[73]

But suppose we avoid this possibility as well and assert that if she were presented with any pair of elements in her preference set (whether for choice or not), she would have some preferential ranking between them that is determined by her preference set as it was before she was presented with the alternatives. *Perhaps* this is so, but it seems exactly as

[73] It is, if the counterfactual is true, presumably determined by *something*, but what determines the preferential ranking need not be her preference set, or her preference set alone.

doubtful as the inscription thesis itself, for the simple reason that it is equivalent to the inscription thesis: *This* counterfactual will be true just in case the inscription thesis is true, and false otherwise; hence, its truth cannot provide the inscription thesis with any independent support.

The other approach to defending the inscription thesis can be treated more briefly. It appeals to the techniques for constructing a cardinal utility function and points out that, for any elements of a preference set that can be located with a cardinal measure along a real-numbered scale, the preferential relations in which they stand to one another, including the preferential relations between all gambles defined over the elements of the preference set, can be derived. The idea is that only a few fixed points are needed rather than an infinite set of comparisons. The rest of the preferential relations are functions of the few fixed points. But as a defense of the inscription thesis, this is confused, because the argument is circular. A cardinal utility function can only be defined for a preference set if the expected utility axioms, including Completeness, are satisfied. But it was the apparent fact that Completeness might not be satisfied by actual preference sets that was the rationale for interpreting Completeness in terms of the inscription thesis. Of course, *if* Completeness and the other axioms are satisfied, the inscription thesis is true (for any finite mind), but that yields no assurance that Completeness and the other axioms *are* satisfied.

So far, I have examined the only two arguments I know for the truth of the inscription thesis and found both wanting. Still, it *might* be true. Or more precisely, it

might be that the inscription thesis is true of the preference sets of at least some of us. So, what I will do at this point is to turn to presenting a series of considerations against the truth of the inscription thesis. I do not know of any direct way of demonstrating that the inscription thesis is false – false, that is, with respect to the preference sets we have – but I think it can be thoroughly undermined: it can be shown that it is very unlikely to be true and thus, that it is not reasonable to believe that all of the weightings or degrees of strength needed for the truth of the inscription thesis are actually present in our preferences.

2.3212 Undermining the Inscription Thesis

I shall look at three kinds of considerations[74] which are aimed at showing that it is implausible to think that we *can* satisfy the axiomatic conditions. The first two have to do with *novelty*, with either new objects of preference or with previously unconsidered decision problems, while the third has to do with *uncertainty* (and its relation to probability). There is some overlap between these, and it will not be possible to keep them completely separate, but that fact has certain advantages for my thesis, for it implies

[74] There is another kind of argument, to which I have already alluded, for the incompleteness of most persons' preference sets. This consists of the considerable empirical evidence that people regularly and systematically violate the axioms of expected utility theory. (See, e.g., Dawes 1988 and Kahneman and Tversky 1990 – which represent only a small sampling from a very large literature.)

This kind of evidence, however, is widely known and has not prevented people from thinking that satisfying the axioms represented an appropriate ideal. My intentions are more radical – not to argue that we fall short of satisfying the axioms and therefore should try harder, exercise more care or the like, but instead that there are deep reasons for thinking that satisfying the axioms is not something that we can do and therefore is not, for beings like us, an appropriate ideal – which of course is not to say that there are no standards of rational choice appropriate to us and in light of which we *should* try harder, exercise more care and the like.

that the considerations cannot be answered in isolation. An adequate answer to one will have to address the others as well, so far as they overlap with it.

2.32121 Novel Objects of Preference

The first issue to consider is how to understand what happens to a preference set when some new element is introduced, for it is an important fact about preference sets that they have histories. The relatively simple preference sets of children become, as the children mature, more complex and come to include elements about which their younger selves would have had no preferences. This may happen in many ways, but what is important here is the extent to which this is a matter of novel experience introducing an agent to something which he did not previously rank preferentially at all. Whether it be a matter of new tastes or sensations, new activities, or new dimensions of concern or interest, they must somehow be integrated into and thereby alter the agent's pre-existing preference set. A great deal of this kind of change must occur in the course of a normal agent's life.

If the inscription thesis is true of an agent who must integrate some novel element into his preference set, and if it remains true after the new item is integrated, then the preferential relation between that item and everything already a part of his preference set – its relative weight or importance – must somehow be *inscribed* into his preference set by virtue of his coming to preferentially rank it. How is this inscription process supposed to work?

Consider the following schematic illustration. An agent prefers *A* to *B* and *B* to *C*. Such an ordinal ranking is of course not sufficient to insure that this is a complete ranking, even within this limited range. For that, we must suppose also that there is some definite gamble between *A* and *C* such that [*A*, *p*; *C*, 1 - *p*] is indifferent to *B*. But let us grant that. What happens when the agent considers some previously unranked option, *D*?

Suppose the agent considers the relation of *D* both to *A* and to *C* and is sure that *D* falls somewhere in the range (exclusive of the endpoints) between *A* and *C*. Thus, *A* is preferred to *D* which is preferred to *C*. For *D* to be fully integrated into a complete preference ordering, though, he must establish at what gamble between *A* and *C* he would be indifferent between the gamble and getting *D*. Moreover, to avoid introducing intransitivities into his preference set, he must get this *exactly* right. It is not good enough to conclude that he is indifferent between *D* and the gamble, [*A*, .6; *C*, .4], if it would be more accurate to say he is indifferent between *D* and the gamble, [*A*, .599; *C*, .401]. In particular, the gamble between *A* and *C* that he accepts as indifferent to *D* must stand in exactly the right relation to the corresponding gamble with respect to *B* – which he has not considered in establishing the ranking!

It does not much matter here how likely it is for the agent to assign exactly the right value to the new object of preference, provided only that it is less than certain.[75] Even if he will very likely get each one right, the fact that the preference sets of adults have

[75] It does not help to claim that there is no sense to 'getting it right' that is independent of the actual ranking assigned. Apart from other problems, such as its close kinship with a revealed preference view, there is one obvious way of getting the ranking wrong: it may turn out to be probabilistically incoherent with other rankings one would assign. If so, they cannot all be correct.

come to be as they are through the addition and ranking of *many* new elements insures that it is enormously unlikely that *all* of the new rankings are exactly correct.[76] And if any such ranking is not correct, there will be intransitivities and therefore failures to completely and unambiguously order his options.[77]

2.32122 Novel Decision Problems

Novelty is an issue for the completeness of preference sets in another way, connected with previously unconsidered decision problems. It is a familiar fact that in the normal course of events we are faced with decision problems that we have not faced nor even considered before. Some choice must be made from among a set of options, when the agent has never compared all of them to one another.[78]

To adapt an earlier example, suppose Caroline knows she would prefer chicken to fish for dinner and Jones to Smith in the municipal election. But she has never compared chicken for dinner to Jones's victory. Still less has she ever compared her actual options, a high probability of chicken for dinner to a slightly increased chance that Jones will win. But we can suppose that circumstances force the choice upon her. How is she to make a

[76] Suppose there is a ninety-nine percent chance that each new object of preference will be ranked correctly. Then there is only about a thirty-seven percent chance that all of one hundred new objects of preference will be ranked correctly. For larger numbers (or a smaller chance of getting each one right), the chances are of course even less.

[77] Strictly, the argument for this presupposes that at least four non-compound elements of a preference set are being ranked *vis-à-vis* one another.

[78] Here, I am restricting myself to options that are *currently*, in advance of the choice, part of the agent's preference set. By this I mean that each of the options can be defined in terms of elements of her preference set that are present because, at some point, there was an explicit comparison and ranking with respect to at least one other element.

decision between options she has never before compared?

If the inscription thesis is true of Caroline, the answer must somehow be there, present to be elicited, in her preferences. She must really, albeit not yet explicitly, prefer the high probability of the chicken dinner to the slightly increased probability of Jones's victory, or *vice versa* or else be indifferent between them.

Now, there are two sorts of cases where it seems that it may be true that the answers really are present to be found in her pre-existing preferences. First, Caroline may immediately know which she prefers as soon as she realizes that she is facing the choice. If she were compelled to choose between having her thumb smashed by a hammer and having chicken for dinner, we would be surprised if she hesitated to answer, even if she had never explicitly compared those options before. If the choice between the chicken dinner and Jones's victory were like that, it seems reasonable to say that the answer was already present in her preferences. Second, she may not immediately know how she ranks the options, but, in reflecting upon them, she comes to some definite conclusion, perhaps by taking into account others among her preferences and how they will be affected respectively by the chicken dinner and Jones's victory. She may have no sense that she is doing anything other than more richly articulating, and thereby bringing to the surface, preferences she already has. If this is what is going on, it may well be that she is finding an answer that was already present in the structure of her preferences.

If the novel decision problems Caroline faces are all of one of the foregoing two sorts – where she immediately or upon reflection realizes how her preferences bear upon

the decision problem – it might be reasonable for her to suppose, at least so far as this is the only relevant issue, that the inscription thesis is true of her. But there is another type of case in which it seems more doubtful that she is only discovering something already true of her preferences.

Suppose that when Caroline is faced with the choice between the relevant probabilities of chicken for dinner and of Jones's victory, she does not immediately know how she ranks them (so it is not a case of the first sort), but also that further reflection does not yield any answer (so it does not appear to be of the second sort, either). Since we are supposing this is a forced choice, she will of course *select* one or the other, but may still say that her selection is not a matter of *preferring* one to the other nor is it a matter of being indifferent between them. She is not confident that her selection has the best expectation of serving her preferences nor that it can be expected to do as well as the alternative.

Now, in a case like this, especially if Caroline has had ample time to elicit her preferences between the options, I think we should take her word for it: She has not succeeded in eliciting a preferential ranking between her options because it was not there to be found.

But it is of course possible to maintain, despite Caroline's actual non-success in finding it, that the answer is still somehow present in her preferences. It is not obvious that this helps for two reasons. First, it may raise a parallel question on a different level, namely, what to do in selecting between a pair of options when, under the circumstances,

a preferential ranking for them cannot be established (or, perhaps, discovered). Should she flip a coin, substitute an easier problem (e.g., act as she would if she thought her choice would completely determine whether she got chicken for dinner or whether Jones got elected), act on impulse or what? Surely her response might be that she cannot establish a preferential ranking for *those* options, either.

But there is a second and more interesting level of response. No one ever has unlimited time or cognitive capacity to investigate the content of her own preferences. Accordingly, since it is not being assumed that all preferences must be either conscious or instantly available to reflection, anyone may, in a given case, be incapable of determining how some option stands with respect to her preferences. Nonetheless, we typically assume that, under good conditions,[79] agents are reliable (not infallible) in determining what their preferences are. Suppose that Caroline is not especially rushed in coming to a conclusion and that there are no obvious interfering factors, but that she reports that she is confident that the answer is not there to be found – that is, that the answer is not already present in her preferences. She simply does not know how to rank the options. Surely, this is possible.

But there is a dilemma here for friends of the inscription thesis. On one hand, Caroline says that her preferences are not sufficient to rank her options, and there is no special reason to doubt her reliability. But if the inscription thesis is true of her, she must be mistaken. On the other hand, we have no *better* warrant for accepting her reliability in

[79] This is vague, and I shall not try to spell it out, but 'good conditions' are meant to rule out various circumstances, such as distractions or tiredness, that can be expected to interfere with or distort judgment.

the kind of case in which she reports, after reflection and under good conditions, that her preferences *do* order her options. Just as she could be mistaken in thinking that her pre-existing preferences do not order a pair of options, she could also be mistaken in thinking that her pre-existing preferences *do* order a pair of options. Just as she might have overlooked or failed to attend in the right way to some feature of her preferences that would succeed in ordering a pair of options, she might have overlooked or failed to attend in the right way to some feature of her situation, extraneous to her preferences, that determined the apparently successful ordering she reported. If her judgment in one case is doubtful, and therefore not sufficient to show that her preferences fail to completely order her options, it appears that her judgment in the other case is equally doubtful and therefore not sufficient to show that her preferences in *that* case completely ordered her options. What is gained on one hand is lost on the other.

It seems to me that issues such as Caroline faces here, though perhaps not common,[80] may affect many decisions. By this, I do not mean of course that many of us are faced with deciding between chicken and the victory of a favored candidate, but that almost any of us can be placed by circumstances in a position in which we do not know how to rank the options we face and in which further examination of our preferences leads no closer to an unambiguous ranking.[81] If this is both correct and correctly represents the actual extent of order in our preference sets, then our preferences do not completely order

[80] It is difficult to be sure just how common they are, since they may often be present when there is insufficient time, before a decision must be made, to identify them and reliably rule out alternative explanations.

[81] That this strikes me as plausible may only show that *my* preferences are not complete!

our options.[82]

2.32123 Uncertainty

There is a third reason for the claim that our preferences are incomplete. To this point, I have for the most part spoken as if the probabilities to be assigned to outcomes were unproblematically available. Further, it is important, if the axiomatic conditions are to be satisfied, that these probabilities be entirely definite point-probabilities,[83] so I have in effect assumed as well that point-probabilities are unproblematically available. To bring out what this involves, we need to look at the contrast case, at the alternative to the availability of point-probabilities.

The conventional way to do that is to draw a distinction between *risk* and *uncertainty*.[84] Conceptualizing a situation as one of risk involves a particular characterization of an agent's ignorance or knowledge of the future. A person making a decision under conditions of risk may not know what the outcome of his decision will be

[82] The plausibility of this conclusion is reinforced by the fact that, for someone in a position like Caroline's, a very natural response – if a decision need not be made immediately – is to turn from considering which of her options she *does* prefer to considering which she *should* prefer. She may of course find no answer to that question, either, but the fact that she raises it and hopes to find an answer implies that she does not think her existing preferences provide everything needed to make a decision.

[83] Consider three elements of a preference set, *A*, *B* and *C*, and suppose that *A* is preferred to *B* and that *B* is preferred to *C*. Assume further that there is some definite value, *p*, such that [*A*, *p*; *C*, 1 - *p*] is indifferent to *B*. Then, if $1 \geq p^* > 0$, [*A*, *p**; *C*, 1 - *p**] must be preferred to *C*, and if $1 \geq p^{**} > 0$, [*B*, *p***; *C*, 1 - *p***] must be preferred to *C*. But unless *p** and *p*** have entirely definite values, it will not always be possible to compare [*A*, *p**; *C*, 1 - *p**] and [*B*, *p***; *C*, 1 - *p***]. It will be consistent with the suppositions that the first would be preferred to the second, that the second would be preferred to the first or that they are indifferent to one another. But if that is the case, Completeness will not be satisfied.

[84] I have written elsewhere (1994) on problems of choice under conditions of uncertainty and reached no definite conclusion save that it is a hard problem and that the most plausible proposal for converting it into an easy problem, the one to be outlined here, is by no means rationally compelling.

but knows exactly the probabilities attaching to the different possible outcomes. For instance, a person considering playing Russian roulette knows (assuming the gun is working properly) that if he pulls the trigger he has a one-in-six chance of getting a bullet in the head and a five-in-six chance of not getting a bullet in the head. Certainty about what the future holds is just a limiting case of risk, one in which a single (non-compound) outcome has a probability of one hundred percent.

The polar opposite case of ignorance about the future can be illustrated in this way. Suppose the person considering playing Russian roulette does not know how many chambers the gun contains or how many chambers are loaded. There may be any number of chambers, *n*, and any number of them, from zero up to *n*, may be loaded. Then, he would have *no* definite probabilities assignable to the possible outcomes of getting or not getting a bullet in the head. He is completely uncertain, with respect to those alternatives, what the future will hold if he pulls the trigger.

Once both risk and uncertainty have been characterized, we can of course imagine any number of intermediate cases of partial uncertainty. For instance, the potential Russian roulette player may know that the gun contains either six or nine chambers and that either one or two chambers are loaded, so, assuming he doesn't want a bullet in his head, he has at worst a one-in-three chance of getting a bullet in the head and at best a one-in-nine chance. He would know that the probability of getting a bullet in the head, if he pulls the trigger, can be represented by some value in the set, {1/9, 1/6, 2/9, 1/3}. Further elaboration of the example could yield probabilities equivalent to some

unspecified value in the closed interval between one-ninth and one-third ($1/9 \leq p \leq 1/3$) – or, for that matter, within any other interval of probability values. Or – not so readily exemplified by ringing changes on possible arrangements for Russian roulette – there may only be ordinal probability information available to the agent: he may know that one outcome is more likely than (or about as likely as, much more likely than, etc.) another.[85]

There are at least three important points here. The first two are fairly obvious; the third requires a bit of elaboration. The first is that most ordinary reasoning about probabilities bearing upon decisions to be made in fact involves at least partial uncertainty rather than risk. Our probability-judgments do not usually assign well-defined point-probabilities; instead, they usually take the form of assigning approximate values or else are simply ordinal. Moreover, there is reason to think that this is not an *eliminable* feature of our reasoning about the unknown future, something that we could replace with point-probabilities if we were more careful or more assiduous in gathering evidence. Apart from other difficulties,[86] this is clear because one of the sources of partial or complete uncertainty in probability-judgments is the *known* possibility of *unknown ignorance*. There may be some possibility relevant to a prospective decision which is not recognized at all, and which is therefore not assigned any probability value.[87]

---

[85] Ordinal probability information is not reducible to some interval probability. '*A* is more likely than *B*' is not equivalent to the probability of *A* being equal to some value in the interval, $.5 < p \leq 1$ (and the probability of *B* being equal to some complementary value in the interval, $0 \leq p < .5$), because there may be some alternative or set of alternatives to *both A* and *B*, the probability of which is unknown or only partially known.

[86] One is the fact that greater care or additional effort in gathering evidence may themselves demand resources, especially in the form of time, that are not available at the time a decision must be made.

[87] Abstractly, a case of unknown ignorance can be described in this way: Suppose an agent considers the probability to be assigned to each of a pair of outcomes, *A* and *B*. Suppose also that he

The second point is that if genuine and irreducible uncertainty (whether complete or not), as distinct from risk, characterizes what an agent knows about the future, there is no well-defined sense to maximizing.[88] The way in which decision-making under risk is assimilated to maximizing is to replace, as the appropriate maximand, actual utility with expected utility, where the expected utility of an outcome is its actual utility discounted by its probability, and the expected utility of an option is the sum of the expected utilities of the various possible outcomes of selecting that option. But when point-probabilities to assign to the outcomes are not available, it is unclear by what the utilities of the various outcomes are to be weighted or discounted in order to determine which option has the greatest expected utility.

The third point requires a bit more background. There is a way of extending standard expected utility theory to cover cases of uncertainty.[89] The result of the extension, *subjective expected utility theory*, requires a strengthening of the axioms, especially Completeness, so that what is necessary to satisfy Completeness includes a

---

assigns fully definite point-probabilities to each. That may be a mistake, for there may be some other outcome, *C*, which he has not considered and to which he has assigned no probability. Since the unconsidered possibility, *C*, may make a difference to what he would decide (had he considered it), the probabilities assigned to *A* and *B* can at best be sufficient basis for judging (say) that *A* is more likely than *B*, but not for judging that the probabilities assigned to *A* and *B* give the correct factors to be employed in weighting the respective utilities of the two.

For an example, consider an agent estimating her potential liabilities under the terms of a contract. She may, having taken all reasonable steps, conclude that her maximum liability is a certain sum and that the advantages she expects to derive can be represented by some different sum. She may assign probabilities to the various events that condition these gains or liabilities and conclude that she should sign the contract. However, it may be that her actual potential liability, due to some inadequately understood provision of the contract, is much greater and that, had she been aware of it, she would have concluded instead that she should not sign the contract.

[88] This will be qualified somewhat below.

[89] Useful discussion may be found in Luce and Raiffa 1985, Chapter 13.

complete ordering over not only risky prospects but also over *uncertain prospects*, where an uncertain prospect is one that may include partial or complete uncertainty about the probability to be assigned to its constituents.[90] Then, the relevant maximand is *subjective expected utility* – that is, expected utility ranging over uncertain prospects and based on whatever probability information or beliefs the agent has (his subjective probabilities, as these are called).[91] Once we admit uncertain prospects as elements to be ordered in a preference set, then, if the preference set satisfies the subjective expected utility axioms, it can be shown that any outcome for which one has only limited probability information can be treated as indifferent to some gamble from a reference set that can be expressed in terms of point-probabilities.[92]

So if, for instance, an agent is faced with partial uncertainty in a choice in which he believes a given action will lead to either *A* or *B* as an outcome and believes that *A* is

[90] To smoothly extend standard expected utility theory to cover these cases, it should be assumed that certain and risky prospects are special cases of uncertain prospects. Certainty will be conceived as varying from zero, or complete uncertainty, to one, or complete certainty. It is an interesting question whether the variation in uncertainty should be conceived as admitting infinitely many values or degrees (whether continuously or not) or whether there is some finite set of degrees of uncertainty in the range. I think that we cannot limit ourselves to finitely many possible degrees of uncertainty. For suppose that there are at least two degrees of partial uncertainty. Suppose that one of them holds when all the probability information available is that one option, *A*, is more likely than another, *B*, and that the other holds when *A* is much more likely than *B*. Could we not then construct a further uncertain choice in which it is, say, completely uncertain whether the first or the second is true, and will not that further choice have to have some degree of uncertainty not to be identified with either of the other two? Now, with a third degree of uncertainty, we can repeat the argument, comparing it with either the first or the second, to get a fourth, and so on.

[91] There are of course further requirements attendant upon the inclusion of uncertain prospects and subjective probabilities. They must satisfy the axioms of expected utility theory (or appropriately strengthened versions of them), and, in particular, the subjective probabilities must also conform to the usual rules for combining probabilities.

[92] This conclusion about limited probability information extends to what might be termed the maximal case of limited probability information: the case in which one has *no* probability information – that is, to choice under conditions of complete uncertainty.

more likely than *B*, there will have to be, by the appropriately strengthened versions of Probabilistic Continuity and Strong Independence, some gamble between *A* and *B* such that the agent would be indifferent between facing it and the corresponding uncertain prospect. That is, there must be some value of *p* such that the agent would be indifferent between being offered the gamble, [*A*, *p*; *B*, 1 - *p*], and being offered the uncertain prospect between *A* and *B* described above.

Once we have gone this far, it is clear that choices under uncertainty, whether partial or complete, can be assimilated to maximizing. Even if, as claimed above, we cannot eliminate uncertainty about the future, we are able, by finding gambles that can be expressed in terms of point-probabilities that are indifferent to uncertain prospects, to say by what values the associated outcomes are to be weighted in decision-making. This is sufficient to show that if the agent's preferences and subjective probabilities satisfy certain conditions, then we can make sense of saying that there is some quantity maximized in rational choice.

This is not to say, however, that it is plausible that anyone actually does satisfy the conditions. For it is a significant fact that the conditions to be satisfied are more demanding than those of standard expected utility theory. Since it is vastly implausible that our preferences satisfy the conditions of standard expected utility theory, it is even less plausible that they satisfy the more demanding conditions of subjective expected utility theory. To put it differently, if standard expected utility theory is to be applicable to an agent's preferences, then those preferences must satisfy the inscription thesis:

somehow, the strengths or weightings of all risky prospects, including ones never explicitly considered but constructed from elements of his preference set, must be present in his preferences and their relations to one another. For subjective expected utility theory, what must be 'inscribed' in the agent's preferences must include not only all of those, but must include in addition relative strengths or weightings for all uncertain prospects that can be constructed from elements of his preference set. For any pair of non-compound risky prospects, *A* and *B*, there can be constructed, in addition to all of the infinitely many gambles possible between *A* and *B*, at least one uncertain prospect,[93] one in which the agent is completely uncertain whether she gets *A* or *B* (designate this as [*A* *B*][94]). This uncertain prospect will have to be assigned a utility index such that it is indifferent to some gamble between *A* and *B*. But [*A* *B*] will also enter, as an element, into the construction of further risky and uncertain prospects, for example, into gambles against risky prospects, such as [[*A* *B*], *p*; *C*, 1 - *p*], into compound uncertain prospects, such as [[*A* *B*] [*C* *D*]], and so on. Since the inscription thesis, adjusted to accommodate uncertain prospects, requires more than the version adapted to merely risky prospects, it must be less likely that it is true of an agent.[95]

---

[93] In fact, there will be at minimum several uncertain prospects, reflecting differing degrees of uncertainty, that are not equivalent to some (merely) risky prospect, e.g., in which *A* is more likely than *B*, in which *A* is much more likely than *B*, in which *A* is about as likely as *B*, and possibly more. In principle, if we take seriously the idea that uncertainty can vary among infinitely many values (see note 90), there will be infinitely many uncertain prospects that can be constructed using *A* and *B* as elements. So, the problem of integrating uncertain prospects into an agent's preference set is actually harder than outlined in the text.

[94] I adopt " " simply to represent some operation of concatenation between prospects, such that the outcome of the operation is uncertain (as distinct from risky).

[95] It is implausible, then, that an agent's preferences with respect to uncertain prospects satisfy the relevant version of the inscription thesis, but there is also a consideration that calls into question whether maximizing subjective expected utility is rationally required, even for the imaginary agent whose

2.3213 Incompleteness

I have been arguing that three kinds of considerations – related to novel elements in a preference set, to novel decision problems and to uncertainty – show that our actual preferences do not satisfy the inscription thesis.[96]

These considerations have two important features in common. The first is that they appeal to cognitive limitations rather than to rational defects. What I mean can be brought out as follows. It is not possible to sharply distinguish cognitive limitations from

preferences *do* satisfy the relevant conditions. In terms of subjective expected utility theory, we can understand what would have to be true of an agent's preferences in order to characterize his choices under uncertainty as maximizing some quantity, namely, subjective expected utility.

But the appeal of maximizing lies in the thought that the maximizer does better in some relevant sense than an otherwise similar agent who does not maximize. In the face of certain outcomes, the utility maximizer does better than the non-maximizer by achieving outcomes that he at least weakly prefers to all others, while the non-maximizer does not. In the face of risk, the expected utility maximizer does better on average than the non-maximizer. In both cases, we have some understanding of the sense in which the maximizer does better than a non-maximizer with the same preferences in the same situation. But, in the face of uncertainty, the subjective expected utility maximizer does better …. how? Only in terms of a metric constructed so as to have a consistent way of saying what it is to do better under uncertainty. What is not clear is that he does better in any other, intuitively acceptable, sense. He need not do better as things actually turn out, nor on average, nor at avoiding disasters, nor at achieving benefits nor in any other independently specifiable way. Moreover, it is not just that the maximizer of subjective expected utility *may* fail to do better in any independently specifiable way, but also that there cannot be a sound argument that he would do better in some other way, for, if there were, we would not be dealing with genuine uncertainty.

What this means, I think, is that here it is the requirements for mathematically representing a quantity that can be maximized that is driving the argument rather than the plausibility of maximizing as a rational requirement. Put differently, since, when uncertainty enters the picture, maximizing has no well-defined independent sense, there is no warrant for supposing that it is a rational requirement. We can *give* it a sense along the lines of subjective expected utility theory, but even for agents, if there are any, whose preferences satisfy the subjective expected utility axioms, we cannot provide any further argument that they will do better to maximize subjective expected utility rather than adopt some other course. They will have only the Pickwickian consolation that they will do better at maximizing subjective utility.

[96] An additional powerful reason for thinking that our preferences do not fully order our options derives from the prevalence of unconscious motivation and particularly from unconscious biases that may shape what options are considered. If some options are systematically prevented from coming into consideration, then their preferential ordering with respect to other options comes into question. Be that as it may, I do not need to rely upon any such arguments.

rational defects, since anything that may be called a rational defect – e.g., the disposition to endorse arguments that have the form of denying the antecedent (that is, P Q; ~P; ~Q) – can also be represented as a possibly severe case of cognitive limitation. Nonetheless, we do, in a rough and ready way, distinguish the two, attributing some mistakes to ignorance or inability and others to defects in reasoning. Roughly, we say that there are rational defects when we think the inference or action in question is one that the person (rationally) should, given his knowledge and abilities, have gotten right, but did not. That, of course, depends on what we take his knowledge and abilities to have been – that is, upon what his cognitive limitations are – so drawing the distinction presupposes that we already know something about which mistakes result from cognitive limitations and which from rational defects. Though we cannot make the distinction perfectly sharp, we still have excellent reason for drawing it, for, on one hand, it would be impossibly demanding to treat all mistakes as the product of rational defects, and, on the other, intolerably lax to excuse all mistakes as due to cognitive limitations. The fuzziness of the distinction does not matter here, so long as it is granted that the kinds of considerations I urge against the truth of the inscription thesis need not involve rational defects.[97]

To return, two of the considerations, the first and third, are in essence arguments from finitude and rely upon the implausibility of the claim that we can correctly and consistently solve problems the solutions of which require unlimited precision. The

[97] If it is objected that the considerations I have urged *do* rely upon the presence of rational defects, it is incumbent upon the objector to specify what those are, and importantly, to do so without begging the question – specifically, without relying upon the assumption that any failure to satisfy the inscription thesis is *ipso facto* sufficient evidence for the diagnosis of a rational defect.

second is more complex. It relies upon three points, first, upon the plausibility of the claim that anyone can be faced with decision problems such that, even under good conditions, she cannot determine which of her options best serves her preferences, second, that this is good evidence that her preferences do *not* completely order her options and therefore do not satisfy the inscription thesis, and third, that any argument casting doubt on the quality of the evidence and therefore on the conclusion would equally, though in a different way, cast doubt on the genuineness of the cases in which her preferences *appear* to order her options.

None of these three arguments assumes that we are rationally at fault for the incompleteness of our preferences.[98] We would be at fault only if there were some procedure available to us for preference acquisition and modification that would non-accidentally result in the completeness of our preference sets. Since we have no such procedure, the incompleteness of our preferences, and therefore their failure to satisfy the inscription thesis, must be ascribed to cognitive limitations.

There is a second important feature which follows from the first. Given that our preferences do not satisfy the inscription thesis, there is no obvious way of *coming* to satisfy it that is not subject to the same problems. If anything, the problems are compounded in the absence of the assumption that the preference set to be revised in order to satisfy the inscription thesis is already complete and transitive. When the problem is

[98] This is not, of course, to deny that failures of rationality may be operative in generating sets of preference that do not completely and unambiguously order options. However, irrationality was not relied upon in the arguments presented, so those arguments amount to a case that even if no rational defects are involved in preference formation or revision, we have strong reason to think that resulting sets of preferences will fail to satisfy the inscription thesis.

only, at a given stage, to integrate a single new object of preference into the elements of a preference set that is already supposed to be complete and transitive, its relations and weighting with respect to other elements of the set must, indeed, be gotten exactly right to avoid introducing intransitivities. However, provided that there is some way to detect that an initial assignment of relative weight to the new element is not correct, only *its* weight needs to be adjusted. But when the preference set is not supposed already to be complete and transitive, it is not evident either where to begin or where to stop. In particular, imposing transitivity upon some subset of the elements in a preference set may introduce intransitivities in other, over-lapping, subsets; rectifying those may introduce yet further intransitivities, and so on.[99] If intransitivities are present, they *may* be uncovered by some piecemeal examination, but there can be no assurance of finding them, short of a complete survey of all the preferential relations that obtain among the elements of a preference set, a survey that is beyond our capacities. Nor, short of a complete survey, can there be any assurance that a contemplated rectification of some intransitivity does not generate other intransitivities.

2.33 Maximizing and Satisficing

Our preferences do not satisfy the inscription thesis and therefore do not completely order our options. Hence, it is not true in general that we can maximize with

[99] Imposing completeness, if it can be done, does not appear subject to an analogous problem. That is, imposing a complete ordering among some subset of the elements of a preference set is not liable to introduce *in*completeness elsewhere. The rub, of course, is "if it can be done." It is not clear that there is any procedure available to a finite mind for imposing completeness upon a preference set, especially if what must be completely ordered includes risky or uncertain prospects.

respect to our preference sets since maximizing is not well-defined with respect to incompletely ordered preferences. Any of us can find ourselves in situations in which there is no answer as to which of our options best serves our preferences.

This can be over-stated or misunderstood, however. It does not imply that maximization is never appropriate, but rather that maximization is not always appropriate. For maximization to be a general requirement upon rational choice, it must be possible to apply it to any decision problem that can be constructed from the elements of a person's preference set. Regardless of the options with which the agent is faced, it must be possible (in principle) to identify one of them as being at least weakly preferred to all others. Denying that maximization is, generally, a rational requirement is consistent with maintaining that, for *some* sets of options for choice, one of those options may be weakly or strictly preferred to all others.[100] And, when this is the case, it may be entirely appropriate to hold that the agent should maximize with respect to her options in the given decision problem.[101] Thus, it is not true that denying that maximization is a requirement for

[100] There are two ways this may be so. First, the actual set of options may be fully ordered by the agent's preferences (though not all possible sets of options would be). Second, some subset within the actual set of options may not be preferentially ranked with respect to each other, but there may be some option strictly preferred to any member of the unranked subset. Two further possibilities, neither of which, arguably, should be assimilated to maximization, would obtain when either (a) there is some mutually unranked subset of options such that each member of the subset is strictly preferred to any option in the complementary subset of mutually ranked options, or (b) when there is some option which is weakly but not strictly preferred to every member of a subset of mutually unranked options and at least weakly preferred to any other option that is not a member of the subset of mutually unranked options.

[101] A complication is that an agent may have adopted some action-guiding principle as a result of a non-maximizing decision (when no maximizing decision was available) and that the principle dictates a non-maximizing choice in a new choice situation in which a maximizing choice is available – i.e., in which some option is at least weakly preferred to all others. On the assumptions that it can be reasonable to adopt such a principle and that it should, at least *ceteris paribus*, govern decisions to which it applies, then it may be that the rational thing to do is to select an option that would, but for the principle, be strictly dispreferred to some available alternative.

rational choice is liable to infect all ordinary decisions, e.g., about what to have for dinner.[102] Over limited domains, there may often be a maximizing choice and nothing I have said should be taken to imply otherwise.[103]

The point remains, however, that maximizing cannot be appropriate to all choices because it is not always well-defined what a maximizing choice would be. Further, the larger the scope of a choice – that is, the greater the extent to which it has effects which can be expected to be substantial and lasting – the more likely it is that maximizing will not be apt. Choice of a career or of a mate provide good examples, for in each case other decisions will in turn depend upon, will be altered or modified, will even be made possible or impossible, in consequence of the earlier decision.

Part of the point is that the further effects cannot be foreseen in detail and may therefore impinge in unforeseeable ways upon matters made relevant by one's other preferences. But so far that is only a problem of uncertainty. There is an additional dimension due to the fact that one of the features of long-term plans is that their execution makes a significant difference to what the person is doing over the term of the plan and that *the person herself* is altered in the process. She engages in different activities, spends time with different associates, and acquires different preferences as an indirect result of executing the plan. Importantly, some preferences relevant to the choice to adopt and

---

[102] Nor, more broadly, is it the case that denying that maximization is a rational requirement is equivalent to denying that there are any rational requirements or *desiderata* when maximization is not appropriate.

[103] There is a further question: How does a domain get limited? Domain-limitation may be the result of *non*-maximizing choices – for example, that only options for dinner are to be considered and, of those, only the members of some short list.

execute the plan may be preferences the person *does not have* when the plan is adopted. The uncertainty involved runs deep: not only is the agent uncertain what the future may bring, she is also uncertain how the unknown future will matter when the time comes. The larger the scope of a choice, the larger is the set of preferences that may be relevant, and the set of relevant preferences (assuming that set is well-defined – which it may not be[104]) probably no more than intersects with the *complete* set of the chooser's preferences at the time of choice.[105]

If, then, maximizing cannot be applied to all choices – and, ironically, is least likely to apply where we would most like some clear-cut decision procedure – what can we do instead? The most popular, and also I think the most plausible, answer (apart from a dogged insistence – or presupposition – that we *can* somehow manage to maximize) is that we should lower our sights and settle for *satisficing*.[106] The core idea is that the agent should seek and select an option that is *good enough*, rather than one that maximizes. Its most natural application is to cases in which an agent is still searching for an acceptable option. (It would normally make little sense to select a worse member of a set of options

[104] Set aside, for the moment, any concerns about how to determine in practice the membership of the set of relevant preferences. Then suppose that each of a pair of options would have different effects upon the preferences of the chooser such that, if one option is selected, the chooser will come to prefer *A* to *B*, whereas, if the other is selected, she will come to prefer *B* to *A*. Does it make sense to say that one of those preferences, to the exclusion of the other, belongs in the complete set? Surely, both are in some sense relevant and both have the same claim to be included, but if both *are* included, the preference set will not be consistent.

[105] Will the chooser have preferences about the ways in which her preferences are subject to modification in consequence of some far-reaching choice (which preferences can then feed back to provide additional criteria or desiderata for the choice)? Quite possibly, but there is no more reason to expect that these preferences will completely order her options than that her other preferences will do so.

[106] I was introduced to the term by Nozick (1981, 300), who cites Simon's 1957 *Models of Man*. The idea has been much discussed, both by Simon and others. See, e.g., Schmidtz 1995, Simon 1996/1969, and, without using the term, his 1990/1983.

known to be available simply because it is still good enough.[107]) Then a satisficer, rather than trying to determine what option is best in terms of all her preferences together, delimits some range within which a decision problem arises – such as what to have for dinner, whether to accept a job offer, whether to buy a house or keep searching – and then settles upon criteria such that, if they are satisfied, an option would, by her lights, count as *good enough*. The options are compared in light of the antecedently established criteria, and the first to qualify as good enough is selected.[108, 109]

Much can be done in the way of formal analysis of satisficing, but I shall leave that to others,[110] except for noting the interesting point that it appears that any rationale for satisficing must itself be a satisficing rationale. An argument cannot be mounted that satisficing is the best we can do (given uncertainty and incomplete preference orderings), for, apart from the fact that 'best' may have no determinate reference in the face of incompleteness, its success would be its failure. If there *were* a sound general argument

[107] In special cases, it might – for example, if there is neither a best member nor any tied for best in the set of available options. See Schmidtz 1995, 42-43. Also, see note 101.

[108] If the criteria turn out to appear *too* easy to satisfy, they may be revised upward, or if too difficult, then downward. In either case, what is "too difficult" or "too easy" is itself at least implicitly a function of a satisficing judgment – that the effort and resources devoted to the search is or is not good enough. See Nozick 1981, 300.

[109] There are indeterminacies, intransitivities and practical dilemmas to which a satisficer is prey. Her choice in favor of one option and against others may be shaped by the order in which questions are asked and considerations brought to bear rather than by the relative merits of the options. If we could, we would like to avoid such difficulties. *In principle*, the maximizer escapes them, but even at its best, the escape amounts to less than may appear. For a maximizer, choosing in the face of risk or uncertainty, the maximizing choice may be to select the best member of a limited set of options, consisting of, say, *A* and *B*. It may still be true that, had he considered a third option, *C*, he would have ranked it above both *A* and *B*. Being a maximizer does not protect an agent against the possibility that actual decisions may depend upon the order in which options are presented or upon other extraneous factors, rather than upon the relative merits of the options. More importantly, the promised escape from practical dilemmas is only an illusion in any case unless we can (always) be maximizers – which we cannot.

[110] See Schmidtz 1995, Chapter 2 and especially 55-57.

that satisficing, in our circumstances, is the best we can do, that would assimilate satisficing to maximizing. Satisficing would then be what maximizing under those conditions amounted to. Satisficing can only be a genuine alternative if its rationale is something other than that it is the best procedure for selection among options.[111] And if the rationale is not that it is good enough, or satisficing, what could it be?[112]

Much can also be done in the way of providing a rationale for satisficing by exhibiting problems with the maximizing model, but that I take to be sufficiently complete for present purposes. What I shall do in the next section is attend to a feature of satisficing that in its turn suggests what I think is the deepest problem with standard, maximizing decision theory.

2.34 Means and Ends

How does a satisficing agent guide her action? Within some domain of concern, she selects as an objective some state of affairs which she believes can be brought about or promoted through her action. She is guided by her judgment that the selected state of

[111] Schmidtz says that satisficing can only be of instrumental value "because to satisfice is to give up the possibility of a preferable outcome, and giving this up has to be explained in terms of the strategic reasons one has for giving it up." (1995, 45) Though he makes it clear that he thinks that the strategic reasons for (sometimes) satisficing are rooted in maximizing from a larger perspective, the conflict with my view is more apparent than real, since he admits (46) that there may often be no optimum from a global perspective: an agent may have to make a choice when nothing unequivocally favors one option over another.

[112] The various axiomatized methods for choice under uncertainty do not provide alternative, non-satisficing routes to the selection of satisficing because they are all ways of identifying some maximand which completely orders options. The satisficer does not have any general procedure for inducing a complete ordering over options. (It is an interesting question for further exploration whether the selection of one of those methods for choice under uncertainty might presuppose satisficing in that there is no proof that one of those methods is best.)

affairs is good enough, that it answers satisfactorily or well enough to her desires and preferences. In other words, she selects a goal and, then, barring alteration of the goal itself, guides subsequent action with respect to that domain by what she understands to be its suitability for the promotion of that goal rather than by its suitability for maximizing the satisfaction of her preferences in general.

Thus, there are two distinguishable stages in the deliberation by which a satisficer guides her action. First, there is goal-selection carried out in light of the agent's preferences, but it is not assumed to be necessary either that selection of the particular goal or even that the selection of some goal or other (then and there) is a maximizing choice.[113] The fact that the goal is selected as being good enough, as answering well enough to her preferences (which will not normally fully order her options), has the important implication that it need not be abandoned instantly should something better or apparently better come along. Since it was not selected for being the best, even proof that it is not the best will not necessarily lead to its abandonment.[114]

Second, once a goal has been selected, action within the relevant domain is guided by its relation to that goal rather than by maximization. To take a simple example, an

[113] Of course, it *may* be, but the satisficing agent has no general procedure for ordering her options so as to insure that a maximizing choice can be identified. See note 100.

[114] The deliberation relevant to abandoning a goal in favor of something else might be said to be *dissatisficing* in structure. It will be appropriate to abandon a goal when it turns out to be bad enough. For a satisficer, there will be a gap between barely finding something else to be better than a currently pursued goal and appropriately abandoning its pursuit.

There are interesting comparisons to be made with Joseph Raz's conception of authoritative reasons as pre-emptive: "the fact that an authority requires performance of an action is a reason for its performance which is not to be added to all other relevant reasons when assessing what to do, but should exclude and take the place of some of them." (Raz 1986, 46; emphasis in original omitted) Adopting a goal is analogous to recognizing an authoritative reason and pre-empts other reasons that would have been relevant had the goal not been adopted.

agent who has embarked upon an investment plan may have chosen to set aside a given percentage of her income every month. Having decided that, she does not reconsider what to do with that portion of her income, whenever an unanticipated opportunity for expenditure arises. She does not, in a typical case, ask whether she would really be better satisfied, all things considered, with new furniture.[115]

There is a clear sense in which the satisficer selects a goal and then guides her action by its relation to that goal. Whether or not the maximizer can be said in the same sense to have a goal or to guide his action in terms of a goal, the issue I wish to pursue is connected to whether he can distinguish his goals from the means appropriate to them – more precisely, whether he can distinguish the differing ways in which preferences with respect to outcomes and preferences with respect to the steps involved in bringing about those outcomes are relevant to his choices. If the distinction cannot be adequately drawn, then there will be a sense in which the maximizer *cannot* be said to guide his actions in terms of his goals.

Consider the following all-too-common problem. An agent has adopted a plan at a

[115] Contrast what a maximizing agent is supposed to do. In the first place, it becomes less clear what it means to settle upon a goal. The maximizer may, of course, in light of all his preferences taken together, undertake to bring about some desired state of affairs, but whether this amounts to settling upon or having a goal is open to question. The problem is that, intuitively, in settling upon and then in having a goal, there is an element of inertia: the goal governs subsequent action but is not itself readily subject to reconsideration. If it is also granted that there may be some motivational state that falls short of constituting the having of a goal, then there is at least a potential gap between being in some way motivated by an envisioned state of affairs and having it as a goal to bring about that state of affairs. We can expect the maximizer to be less attached to his goal than the satisficer, for he will be constantly ready to give it up should something better come along. Perhaps this degree of attachment is not sufficient for having a goal. On the other hand, it is of course true that even a non-maximizing agent is prepared at some point to reconsider, so readiness to reconsider alone cannot disqualify the maximizer as a goal-pursuer. The disqualification may be grounded in his being *too* ready to reconsider, but I do not know how to determine what degree of readiness is too great. Since none of the argument to come turns upon this being, by itself, an important difference, I shall not pursue it.

time, t_0, to bring about a preferred outcome at a later time, t_2. Execution of the plan requires performance of a particular action (the Step) at an intermediate time, t_1. I shall suppose that at t_1 there has been no change in relevant information available to the agent nor has there been any unforeseen change in the agent's preferences, but at (or just before) t_1, the agent strictly prefers not to take the necessary Step. In addition, we can suppose that the preference change with respect to the Step was itself foreseen when the plan was adopted.

Such situations are familiar. An example might be deciding upon a diet. There is an envisioned outcome, losing weight, ranked above other accessible future outcomes and a necessary step, such as refraining from between-meal snacks. In addition, at the time the plan is adopted, it is recognized that there will be temptations to snack between meals: when the Step must be taken, the agent will prefer snacking to sticking to the diet. On one hand, it appears that the agent's reasons for taking the Step are just the same as for initially adopting the plan – no unanticipated information or preference has entered the picture. If the plan was initially well-conceived – that is, if it was reasonable to adopt it – the agent ought to take the Step. On the other hand, now that the prospect of snacking is immediate, the agent does *not* prefer the outcome expected from refraining from the snack. He would, right then, rather snack than lose weight. Why must he be bound by his preferences of a few hours earlier? If it is rational for him to guide his actions by his preferences, why are the preferences at t_0 decisive, while those at t_1 are discounted – especially since it is the preferences at t_1 that are actually experienced at the time the

choice to snack or not must be made?[116]

Most of us – however difficult we find it to carry through in practice – suppose that the former argument is better: Having made a reasonable plan, and in the absence of relevant additional information not already taken into account in the formulation of that plan, it is reasonable for a person to take the necessary steps to implement the plan, even if those necessary steps are dispreferred at the time they must be taken.

However, according to standard decision theory, this misdescribes the situation. It is not that the first argument is invalid, but that it depends upon a false premise. In standard decision theory, the only reasons we have are based on preferences and expected consequences (subject to a budget constraint) at the time a choice is made. A decision cannot rationally depend – except insofar as this affects current preferences and expectations – upon a past event such as having adopted a plan. Thus, if the step needed to carry out the plan is such that one would prefer not to take it at the time of choice (i.e., when the step must be taken or not), then one has reason at that time not to take the step. But if this fact was really foreseen when the plan was adopted, the plan fails to be reasonable because its execution depends upon the taking of a step which it is not reasonable to take. One who accepts the rationality-defining postulates of standard decision theory should either not have formulated a plan aiming at that goal or else should have made provision that every step would be preferred to its alternatives at the time it would have to be taken. We can put this somewhat differently by saying that, for standard

[116] Does he still *have* the preference to take the Step, even if it is not motivationally salient? Perhaps we should allow that he may, but if so there is at least an apparent conflict among his preferences, and it is not clear which should govern his choice.

decision theory, reasonable plans are constrained by the requirement that they contain nothing but feasible steps, where feasible steps are all at least weakly preferred, when they must be taken, to their alternatives. If that requirement is not met, then the plan was not reasonable in the first place.[117]

This seems unsatisfactory. If standard decision theory is correct about situations of this sort, there may be an outcome which an agent would like to achieve and a plan that, if executed, would achieve that outcome, and it may be that if the plan were executed and the outcome achieved, the agent would be glad she had adopted the plan and taken all the necessary steps, but nonetheless, the agent cannot rationally adopt the plan because it incorporates infeasible steps. Her best available options are to either give up seeking that outcome or to undertake special arrangements to make sure that all the steps are feasible. Either option represents some cost, whether in the form of giving up the chance to obtain her most preferred outcome or in the form of making special provisions to avoid having to take infeasible steps.[118]

Why does this problem seem so difficult for the maximizer of standard decision theory? There are two points to note before trying to answer. First, the question is not (or not just) why *we* sometimes find it hard to carry through our plans. That seems adequately accounted for by imperfect rationality. Rather, the question is about why *ideally rational* agents would find themselves apparently having to settle for second-best. And second, it is specifically a problem for rational agents *as conceived by standard*

[117] See McClennen 1990, especially chapters 12 and 13.

[118] A further concern is that the feasibility-insuring provisions might themselves be so costly that, if they are necessary to achieve the outcome, then the outcome is not worth achieving.

decision theory. If, as I have been arguing, the conception of rationality embodied in standard decision theory is not normative for us, it may be possible to address the problems associated with taking steps to achieve a goal in ways not open to the maximizer.[119]

The reason the problem seems difficult, I think, is that standard decision theory has no satisfactory way of making a normative distinction between ends and means. If the distinction could be made, there would be conceptual room to hold that ends provide reasons for adjusting means but not *vice versa*. To see what the problem is, consider where or how such a normative distinction might be represented. There are two plausible candidates, that ends are to be characterized in terms of outcomes of actions or in terms of intrinsic preferences.

Suppose we identify ends with outcomes and hence identify means with steps that contribute to bringing about those outcomes. Then, the problem is simple:[120] Though we can say what steps contribute to what outcomes, all normativity vanishes because the fact that a step contributes to an outcome will not provide any reason for taking that step or for avoiding alternatives. Any step and any combination of steps will lead to some outcome or other. What is needed, at minimum, is some way of discriminating *among* outcomes, to identify one or some as ends, rather than others, and therefore to enable the identification

[119] Satisficers do not face the same problem, at least not in so acute a form, for they are not automatically subject to criticism for making non-maximizing choices and therefore not for taking counter-preferential steps. (I do not mean to suggest that being a satisficer is sufficient to deal with the problem in all its forms.)

[120] More difficult problems pertain to questions about individuating outcomes and setting an appropriate time-horizon for the identification of outcomes, but those will not concern me here.

of some options, rather than others, as means to those ends.[121] In short, in addition to the identification of outcomes and contributory steps, something exogenous is needed to represent the normative force of ends.

It might be thought that the exogenous factor can be readily supplied. Consider *wholly derived preferences*, or *derivative preferences* for short. At a given street corner, I prefer turning left over turning right because I prefer one grocery store to another. If not for my preference between stores, I would (then and there) have no preference for turning one way over the other. On pain of infinite regress, however, not all preferences can be wholly derived; there must be some which are non-derivative or *intrinsic preferences*.[122] The proposal, then, would be that ends are to be characterized in terms of intrinsic preferences. Ends will be intrinsically preferred to their alternatives, so, once ends are securely identified, we can turn to the consideration of contributory means – that is, we can show what derivative preferences an agent should have and act upon in light of his intrinsic preferences.

The problem with this is that no role is left for temptation. To return to the example of adhering to a diet, consider the (readily generalizable) case of George, whose

---

[121] And that is just the beginning, for many features of outcomes of action are not intuitively part of any end pursued in a given course of action. Typing rearranges small particles on my keyboard, but the arrangement or rearrangement is not what I aim at in typing. See also note 21.

[122] There may be single preferences, where *A* is preferred to *B*, which are wholly non-derivative in the sense that only the preference for *A* over *B* is relevant to any choice between the two. But it may be that a preference is not wholly derivative without being wholly non-derivative. The preference relation between the two may be part of some set of mutually supporting or interlocking preferences such that *A* would be preferred to *B* if nothing else were at stake, but that if something else were at stake, the preferential relations could be altered. Since no important part of my argument turns upon whether we are speaking about wholly or partially non-derivative preferences, I shall indifferently employ 'intrinsic preference' to cover both.

end or goal is to lose weight. So far as he has only derivative preferences between steps or means, the only explanation for his not taking a step that is better than available alternatives at contributing to his ends must be in terms of misinformation, ignorance, or inadvertence. He will certainly have no motivation to take a step that either leads away from or less effectively toward his ends. But then, whence comes the temptation to snack? Surely, yielding to temptation is not a matter of an accidental misstep on the way to his goals.

The answer must be that George's preferences with respect to the steps to be taken are not wholly derived. He is motivated to snack rather than stick to the diet because he has some intrinsic preference for snacking, then and there. If so, there are two possibilities, that the preference for snacking either can or else cannot be integrated into a consistent ordering with George's other intrinsic preferences. If it cannot, then there is no consistent set of intrinsic preferences to identify as the relevant end or ends, and therefore none in terms of which to regiment means.

Matters are no better, however, if we suppose that George's preference for snacking can be integrated into a consistent ordering. For then, at least *prima facie*, the act in question is not, strictly speaking, one of *yielding* to temptation; rather, it is an act licensed by its service to his ends. There is an intrinsic, not merely a derived, preference for what we are calling "yielding" and, since ends are to be identified in terms of intrinsic preferences, no genuine yielding after all.

Now, it might be supposed that room for the possibility of yielding to temptation

can be found in the thought that the snacking to which George is tempted is contrary to what he really or most prefers. Though his intrinsic preferences, including the preference for snacking, can be integrated into a consistent ordering, snacking then and there does not serve them.

This is very puzzling. Unless we wish to revert to a revealed preference theory, we should not complain that we cannot make sense of a situation in which, *ex hypothesi*, George can make a choice contrary to what he most prefers.[123] Nonetheless, there are difficulties, and though there are several possibilities, none seems adequate. To begin, what is the other element of George's preference for snacking: what is snacking preferred *to*? Presumably, it is preferred to not snacking. Also, however, sticking to the diet, which requires not snacking, is preferred to snacking. If that is not simply to amount to an inconsistent set of preferences and therefore to an inconsistent set of ends identified in terms of those preferences, there must be some sense in which the preference for sticking to the diet is what sets George's end while the preference for snacking does not. Since both the preference for sticking to the diet and the preference for snacking are intrinsic, we cannot distinguish between them on the basis of the presence or absence of an intrinsic preference. I have suggested we have to say that adhering to the diet is what George most prefers, but how are we to understand that? We do not mean that the diet-adherence

[123] Another possibility is that a revealed preference theory might be rejected on the grounds that it does not adequately accommodate indifference or incomplete preference orderings. However, in the case at hand, neither of these is supposed to be at stake. George is supposed to have a consistent preference ordering which is not served by snacking. One could object to the claim that *this* preference ordering (or one structurally like it) might not be revealed in choice behavior without being committed to the general claim that choice always reveals preference.

preference has greater introspectible intensity, for in those terms snacking may well be what George most prefers. Nor will it do to content ourselves, without further elaboration, with the formulation I gave earlier, that the act of snacking is contrary to what George most prefers, because, for the decision theorist, there are only formal limits on what preferences may enter into a utility function, and, subject to those constraints, any preference is to be considered on the same terms as any other. The set of his preferences is equally consistent if he snacks and alters the preference for adhering to the diet as if he refrains from snacking in order to adhere to the diet. What is needed is some further explication of the sense in which the preference for sticking to the diet is supposed to be of greater weight or importance than the preference for snacking.

That explication has not been, and, I submit, will not be, forthcoming. More precisely, it will not be forthcoming in terms that can be represented within standard decision theory, for the explanation being sought is one of the normative distinction between means and ends, not of the psychology or phenomenology of preference or desire. When we ask why adherence to the diet is more important than snacking, what we want to know is why George *should* abstain from snacking, and the answer to that lies in the fact that losing weight is George's goal or end. It is because losing weight is the goal that adherence to the diet is more important than snacking, not because adherence is more important that loss of weight is the goal. There will be no answer in terms of preferences alone.[124]

[124] Nor will there be an answer in terms of (just) preferences combined with beliefs and expectations. I do not mean, of course, that preferences (etc.) do not enter into the selection of goals, just that the role of goals or ends in the guidance of choice is not captured in those terms alone.

And that is the deepest problem with standard decision theory. Whether explicitly or not, the theory seeks to be reductive about ends, to account for them in terms of the satisfaction of preferences and the like.[125] But, as we have seen, the attempt to understand rational choice in terms of maximizing the satisfaction of preferences ultimately leaves the theory unable to express or represent the normative distinction between means and ends. Taking instrumental reasoning seriously requires that we go beyond decision theory.

2.4 Summary

Decision theory, understood as providing a normative account of rationality in action, given a set of beliefs, preferences and constraints, is often thought to be an adequate formalization of instrumental reasoning. As a model or representation of important features of instrumental reasoning, there is much to be said for it. However, if decision theory is to adequately account for or formalize (correct) instrumental reasoning, then its proposed axiomatic conditions must be normative for choice. That is, it must be that a choice is *rationally defective* unless it proceeds from a preference set that satisfies the axiomatic conditions.

Though some axiomatic conditions are largely uncontroversial, the same cannot be said for others. Accordingly, it is not easy to provide adequate support for the complete set of conditions. There seems to be no clear case that every agent who fails to satisfy the

[125] What I envision in the way of a non-reductive account of ends is along the lines of Bratman's planning theory of intention. Though he does not typically speak in these terms, roughly, an objective or goal is what intentional action is guided toward, and an intention is "a distinctive attitude, not to be conflated with or reduced to ordinary desires and beliefs" (Bratman 1999, 10) – nor, I would add, should it be conflated with or reduced to preferences.

complete set of conditions must be rationally mistaken. Indeed, the apparent fact that competent decision-makers, including experts in decision theory, make choices that would be disallowed by the axioms is one of the sources of doubt as to their normative standing.

For my purposes, the most important of the conditions is Completeness, the requirement that an agent's preferences completely order her options. Applied even to relatively small numbers of elements in a preference set, the number of comparisons required, if each must be made explicitly, quickly becomes too large to be plausibly managed. If extended to all the options that can be constructed from elements of her preference set under conditions of risk (to say nothing of uncertainty), a complete ordering would involve infinitely many pair-wise rankings. The agent cannot have explicitly performed all of these rankings, and so, if her preferences do completely order her options, it must be assumed that her preferences have an underlying structure which suffices to determine all the needed preferential relations. The claim that there is such an underlying structure to an agent's preferences, that a complete preferential ordering is inscribed in her preferences, is what I call *the inscription thesis*.

I focus upon Completeness and the related inscription thesis because, although we cannot satisfy the axiomatic conditions unless the inscription thesis is true of us, it can be shown to be either false or enormously unlikely that the inscription thesis is true of us. In neither case is it reasonable for us to believe it of ourselves. Further, if it is not true of us, there is little we can do to rectify matters. There are no obvious steps to take that would result in our coming to have a complete preference-ordering. The reasons that it is

implausible to think the inscription thesis true of us are also reasons to think that it is implausible that we can ever bring it about that it will, in the future, be true of us.[126]

Since a complete ordering of all of our options is not inscribed in our preferences, maximizing cannot be a general rational requirement. Rational choice may include maximizing when one option is weakly preferred to all others, but it is not defined by maximizing, for it is not always well-defined what it is to maximize.

The most important alternative to maximizing is satisficing, in which an agent selects an option because it is satisfactory or good enough in terms of her preferences, the rationale for which must ultimately itself be satisficing in nature. Its importance is, first, that it provides an alternative to maximizing that is within our capacities, and second, that it provides a natural way to model the selection of goals or ends in light of an agent's preferences, without implying that the having, adoption or pursuit of goals is reducible to or explicable entirely in terms of the agent's preferences.

Having a non-reductive account of ends or goals is, in turn, important in order to have a satisfactory account of ordinary instrumental reasoning, including such commonplaces as the fact that we can be tempted to act in ways in conflict with our objectives. Though there is much that can be learned from decision theory, it does not adequately represent instrumental reasoning.

---

[126] Relatedly, even if there were steps we could take to impose Completeness on our preferences, it is not clear that we would have any reason to do so, for the supposed reason would either depend upon a complete ordering of our preferences or not. If the former, then the argument for imposition is fatally compromised, while, if it is the latter, the incompleteness of our preferences leaves open the possibility that the reason will be undefeated, untied, but still not rationally decisive.