Must rational intentions maximize utility?[1]

Ralph Wedgwood

How is it rational for us to form, maintain, and revise our intentions, in cases where we are not certain of what the outcome of the available acts will be? The mental events in which we form or revise our intentions are choices or decisions. So, to answer this question, it is natural to look to the approach that has been developed by classical decision theory (also known as rational choice theory). According to this approach, rational choices should be defined as choices that maximize a certain kind of *probabilistic expectation* of a certain kind of *value*. In this essay, I shall assume that this approach to defining rational choice is broadly correct.

This assumption leaves it open exactly *what kind* of "value" rational choices should be defined as maximizing the relevant kind of expectation of. There are two main versions of this approach that have appealed to philosophers. On a neo-Aristotelian approach, rational choices should be defined as choices that maximize the relevant expectation of a certain kind of *objective goodness*. On a neo-Humean approach, rational choices should be defined as those that maximize the relevant expectation of *utility* – where "utility" is interpreted as a *measure of subjective preference*. In this essay, I shall argue against the neo-Humean approach, and in favour of the rival neo-Aristotelian approach.

Some arguments against the Humean approach offer an external critique of this approach. What authority, after all, do our preferences have, to dictate to us what we should do? If we can

[1] Earlier versions of this paper were presented to audiences at the Universities of Sydney, Adelaide, and Melbourne, and at the London School of Economics. I am grateful to those audiences for their helpful comments.

resist our preferences, why shouldn't we at least sometimes resist?[2] In this essay, however, I shall attempt a more internal critique. I shall try to show that even by the neo-Humean's own lights, the neo-Humean approach cannot be made to work: if preferences are what the neo-Humeans are committed to taking them to be, there is in fact no way of measuring these preferences by means of anything like a utility function.

In arguing for this point, I shall have to confront the famous "representation theorems" that loom large in the formal literature on decision theory. According to these theorems, so long as the rational agent's preferences meet certain conditions or "axioms", it can be *proved* that these preferences can be measured by means of a utility function. I shall argue that if preferences are what the neo-Humeans are committed to taking them to be, these preferences do not meet enough of these conditions or axioms to allow the neo-Humean to defend the claim that even the most ideally rational preferences can be measured by means of a utility function. There is, admittedly, an alternative interpretation of what "preferences" are that will make it plausible that – when interpreted in this way – these preferences do indeed meet the crucial axioms. But as I shall explain, this alternative interpretation is in fact a neo-Aristotelian conception of rational choice, not a neo-Humean conception. Rational choices – that is, rational ways of forming, maintaining, and revising our intentions – should not be defined as choices that maximize expected utility; they should instead be defined as choices that maximize a certain kind of expected goodness.

1. The concept of "utility"

Utility, as the term is being used here, is a *measure* of an agent's *preferences*. Some decision

[2] In effect, I offered such an external critique elsewhere (Wedgwood 2003).

theorists are willing to contemplate the idea that an agent's utility function, as Lara Buchak (2013, 18) puts it, corresponds to a "real mental state" of the agent, and that "preferences are merely a way to discover an agent's utility function…: while preferences are a good guide to the utility function in general, there is no constitutive link between the two." I shall not think of utility in this way. As I understand utility, an agent's utility function is precisely a measure of how strongly she prefers some prospects over others. If U(•) really is "your utility function" and U(*A*) is greater than U(*B*), then it follows that you do not prefer *A* over *B*. In general, if the difference between U(*A*) and U(*B*) is positive, then the greater this difference is, the more strongly you prefer *A* over *B*; if the difference is negative, then the greater the difference is, the more strongly you prefer *B* over *A*; if U(*A*) is equal to U(*B*), you are indifferent between *A* and *B*.

The notion of the *expected value* of a function V is defined in terms of some probability function P(•). To define this notion, we need a "partition" – that is, a set of jointly exhaustive and mutually exclusive states of affairs or propositions $S_1, \ldots S_n$. (To say that these propositions are "jointly exhaustive" is to say that this probability function P assigns probability 1 to the *disjunction* of all these propositions '$S_1 \vee \ldots \vee S_n$'; to say that these propositions are "mutually exclusive" is to say that P assigns probability 0 to the *conjunction* of any two of these propositions '$S_i \wedge S_j$', where $i \neq j$.) This partition also needs to have the property that for every proposition S_i in this partition, and for every possible prospect or object of preference *A*, there is a definite value that this function V assigns to *A* according to S_i, V(*A*, S_i). Then the simplest way of defining the *expected value* of a prospect *A* according to P is as the probability-weighted sum of *A*'s values according to these propositions – where each of these values is weighted by the probability of the relevant proposition:

$\sum_i P(S_i)\, V(A, S_i)$.[3]

If this value function V is a utility function, then this formula gives *A*'s *expected utility* according to this probability function P, $EU_P(A)$.

We are assuming here that some broadly probabilistic approach to rational choice is correct. To fix ideas, I shall suppose that it follows from this assumption that at every point in time, there is a unique probability function P that should guide the rational agent's thinking at that time; but I shall leave it completely open, for the purposes of this discussion, exactly what this probability function P is. According to the neo-Humean definition, rational choices are those choices that maximize expected utility according to the relevant probability function P: that is, a rational agent will choose an option *A* only if the agent believes that *A* is available, and for every alternative option *B* that the agent also believes to be available, the expected utility of *A* according to P, $EU_P(A)$, is at least as high as the corresponding expected utility of *B*, $EU_P(B)$.

In fact, however, the central claim of expected utility (EU) theorists is different – in effect, it is the claim that a rational agent's *preferences* maximize EU_P. More precisely, they claim that the utility function that measures every rational agent's preferences is *expectational*, in the sense that the utility of each prospect *A* is identical to its expected utility according to the relevant probability function P. This central claim about rational preferences implies the neo-

[3] Various refinements of this definition are possible. For example, causal decision theorists in the style of Lewis (1981) will insist that this partition $S_1, \ldots, S_n$ must be a partition of "causal states of nature" (or "causal dependency hypotheses"), which are utterly beyond the relevant agent's control; and evidential decision theorists such as Jeffrey (1983) will insist that the relevant probability should be the conditional probability of the relevant state S_i conditional on the assumption that the relevant prospect *A* is realized, $P(S_i|A)$. The subtle differences between these different versions of EU theory will not matter for my purpose here, since my concern is with all versions of decision theory that focus on the notion of utility at all.

Humean definition of rational choice given the assumption of a certain fundamental connection between rational *preference* and rational *choice*. Specifically, this fundamental connection is the following: a rational agent will choose a prospect *A* only if the agent believes *A* to be available, and there is no alternative prospect *B* that the agent also believes to be available, such that the agent prefers *B* over *A*. I shall assume that the plausibility of the neo-Humean definition of rational choice stands or falls with this central claim of the EU theorists. So the main focus of my discussion will be on this central claim – that is, on whether rational preferences maximize expected utility.

For this central claim to be true, it is not necessary that the preferences of any agent whatsoever can be measured by means of a utility function; all that is necessary is that any *fully rational* agent's preferences can be measured by means of such a utility function, and that any such agent's preferences maximize expected utility in the sense that I have defined.

Many theorists believe that it can be *proved* that any fully rational agent's preferences must maximize expected utility in this sense. What can unquestionably be proved is what have come to be known as "representation theorems." According to these theorems, if one's preferences meet certain conditions (or "axioms", as they are often called), then those preferences can be "represented" by a unique utility function and probability function.[4]

What does it mean to say that a probability function P(•) and a utility function U(•) "represent" one's preferences? If $EU_P(\bullet)$ is an expected utility function defined in terms of P(•) and U(•), it means that the following two conditions hold:

[4] Strictly, the probability function will be unique up to an arbitrary choice of a *unit*, and the utility function will be unique up to an arbitrary choice of a *unit* and a *zero point*. See for example the representation theorem that was proved by Savage (1954, chap. 3 and 5).

(i) For every prospect *A* in the relevant domain, $U(A) = EU_P(A)$.

(ii) For all prospects *A* and *B* in the relevant domain, one prefers *A* over *B* only if $U(A) > U(B)$; and one is indifferent between *A* and *B* only if $U(A) = U(B)$.

Clearly, the claim that the agent's preferences can be represented by a probability function P(•) and a utility function U(•) is equivalent to the claim that the agent's preferences maximize the corresponding expected utility function $EU_P(\bullet)$. If this probability function P is in fact the probability function that should be guiding the rational agent's thinking at the relevant time, and if the preferences of all rational agents must meet these axioms, then we would indeed have a proof of the EU theorists' central claim, that all rational preferences maximize expected utility according to the relevant probability function P.

To gauge the significance of these representation theorems, however, we need to appreciate that among decision theorists, there are two quite different interpretations of the idea of a utility function. José Luis Bermudez (2009: 47) refers to these interpretations as the "operational understanding" and the "substantive understanding" respectively, while Buchak (2013: 17) calls them "formalism" and "psychological realism" about utility. On the formalist or operational view, all the real psychological facts about preferences consist in the way in which these preferences *rank* or *order* the relevant prospects. On this view, the utility function is simply a way of modelling or redescribing this preference ranking: the representation theorem explains what it is for the utility function to be a way of modelling or redescribing this preference ranking. On the realist or substantive interpretation, on the other hand, the strength of one's preference is a real psychological fact about one's state of mind. This fact about one's state of mind may be revealed by these facts about how one's preferences rank all the relevant prospects, but it is not constituted by these facts; the representation theorem merely provides us with a reason for

accepting that the strength of a rational agent's preferences can be measured by a utility function.

On the formalist interpretation of utility, all of the axioms are crucial, since on this interpretation, to say that one's preferences can be measured by means of a utility function just is to say that one's preferences meet these axioms. On the substantive interpretation, on the other hand, it might not be necessary for one's preferences to satisfy all of these axioms, since on this interpretation the strength of one's preferences might in fact be measureable by a utility function even if we cannot prove that this is the case by means of the representation theorem.

For our purposes, a complete enumeration of these axioms will not be necessary. What is important, however, is to recognize that these axioms fall into three different groups. First, some of these axioms require that the rational agent's preferences must not be *trivial*: in particular, the agent must prefer some prospects over others. Since an agent whose preferences were trivial in this way would presumably have no need of making any choices, I shall not discuss these non-triviality axioms here.

Secondly, some axioms require that the domain of prospects that the rational agent has preferences towards must have a certain sort of *structure*, and that the rational agent must have a *complete* set of preferences over this domain (that is, for every two prospects *A* and *B* in this domain, the agent must either prefer *A* over *B* or prefer *B* over *A* or be indifferent between *A* and *B*). Following Joyce (1999, 82) I shall call these the "structure axioms".

Finally, some axioms require that the agent's preferences must meet various requirements of *coherence*; I shall discuss these axioms in Section 4. The most important of these coherence requirements are the following:

i. The rational agent's preferences must be *transitive*: whenever the agent prefers *A* over *B*, and also prefers *B* over *C*, she also prefers *A* over *C*.

ii. The rational agent's preferences must be *independent*: whenever the agent prefers *A* over *B*, and the truth or falsity of a proposition *p* is irrelevant to this preference, the agent also prefers the gamble that gives her [*A* if *p* is true and *C* if *p* is not true] over the similar gamble that gives her [*B* if *p* is true and *C* if *p* is not true].

iii. The rational agent's preferences must be *monotonic*: whenever the agent prefers *A* over *B*, and also prefers the gamble that gives her [*A* if *p* is true and *B* if *p* is not true] over the opposite gamble that gives her [*B* if *p* is true and otherwise *A*], then likewise, for every other pair of prospects *C* and *D* such that she prefers *C* over *D*, she must also prefer the gamble that gives her [*C* if *p* is true and otherwise *D*] over the opposite gamble that gives her [*D* if *p* is true and otherwise *C*].[5]

There is a crucial difference between these groups of axioms. The structure axioms are necessary for *proving* that the agent's preferences maximize expected utility, but the claim that these preferences maximize expected utility does not entail that these structure axioms are true. By contrast, the pure coherence axioms are not just necessary for proving that the agent's preferences maximize expected utility; these axioms are actually entailed by the claim that these

[5] To be precise, there are several subtly different representation theorems, and the proofs of these different theorems rely on slightly different axioms: Savage's axioms (listed in Joyce 1999, chap. 3) are slightly different from the axioms that are relied on by the proof of Jeffrey's (1983) theorem that was due to Ethan Bolker (listed in Joyce 1999, chap. 4). However, these subtle differences will not matter for my purposes. Each of the axioms that have been proposed is either: (i) an innocuous non-triviality axiom requiring that the agent prefers some prospects over others; (ii) some structure axiom requiring that the agent's preferences give a complete ranking of all the prospects in a certain domain; or (iii) a coherence axiom that is necessary for the truth of the representation theorem. Every version of the proof rests on some structure axiom requiring that this domain of prospects must be *infinite* – and that is the only feature of the structure axiom that concerns me here.

preferences maximize expected utility.

On any interpretation of utility, then – including both the "substantive" or "realist" interpretation and the "operational" or "formalist" interpretation – the claim that rational preferences must maximize expected utility implies that rational preferences must meet these pure coherence axioms. On the other hand, whether or not rational preferences must meet the structure axioms will depend on how the notion of utility is interpreted. On the operational or formalist interpretation, an agent's utility function is simply a way of modelling or redescribing the agent's preference ranking of the relevant prospects; and to say that one has a unique utility function just is to say that this preference ranking meets these axioms. So, on this operational or formalist interpretation, rational preferences would have to satisfy all the axioms – including the structure axioms – for it to be the case that rational preferences maximize expected utility. On the substantive or realist interpretation, by contrast, even if the structure axioms are not satisfied, it could still be the case that rational preferences maximize expected utility; we just would not be able to establish that this is the case by relying on the representation theorem.

I shall argue that rational preferences do not satisfy the structure axioms in Section 3. While this need not worry the proponent of the substantive or realist interpretation, it entails that, on the operational or formalist interpretation, the EU theorists will have to retreat from their central claim that for every agent and time, if the agent is rational at that time, there is a unique expected utility function that the preferences that the agent has at that time must maximize. In Sections 4 and 5, I shall argue that rational preferences – at least if they are mental states of the sort that the neo-Humean is committed to taking them to be – need not even satisfy the pure coherence axioms. First, however, in Section 2, I shall address a crucial preliminary question: What are preferences?

2. What are preferences?

There are certainly some silly interpretations that we could put on the word 'preference' that will ensure that "preferences" satisfy the relevant axioms. Suppose that there is some measurable quantity that prospects can have – for example, one such measurable quantity is the *amount of money* that the agent will gain from the realization of those prospects. Consider the following interpretation of "preferences": for you to "prefer" *A* over *B* just is for you to have a greater expectation of this quantity for *A* than for *B*. Since this interpretation of the word 'preference' just *defines* a preference as having a greater expectation of some quantity, this interpretation guarantees that "preferences" will satisfy the relevant axioms (at least so long as the degrees of belief that make it the case that you "have" a certain expectation of monetary gain for a prospect *A* can be represented by a probability function).

This silly interpretation of the term 'preference' ensures that "preferences" meet the conditions that are required by the EU theorists' axioms. For example, if you have a higher expectation of monetary gain for *A* than for *B*, and you also have a higher expectation of monetary gain for *B* than for *C*, then you must also have a higher expectation of monetary gain for *A* than for *C*. So, on this silly interpretation, your "preferences" will certainly be *transitive*. Similarly, it seems clear that on this silly interpretation, your "preferences" will satisfy the independence and monotonicity axioms as well.

However, this interpretation of "preferences" is obviously unacceptable. The most fundamental problem is that it makes it totally implausible that there is any fundamental connection between rational "preferences" and rational choice. For example, consider the St. Petersburg Paradox: a fair coin is going to be tossed repeatedly until it lands heads; and there is a ticket that will pay $2 if it lands heads on the first toss, $4 if it lands heads on the second toss, $8

if it lands heads on the third toss, and in general $\$2^k$ if it lands heads on the *k*th toss. How much should you be willing to pay for this ticket?[6]

If no rational agent ever chose any option *A* when they had a higher expectation of monetary gain for an alternative option *B*, there is *no* price that a rational agent would refuse to pay for this ticket. Similarly, if rational agents never chose any option when there was an alternative that had a higher expectation of monetary gain, they would never buy an insurance policy for more than its expected monetary payoff – which (at least if everyone was being guided by the same probability function) would make the insurance business impossible. These absurd results could all be avoided by a version of EU theory that detaches preferences from expected monetary gains, and allows that monetary gains have declining marginal utility (while monetary *losses* have *increasing* marginal disutility). The general point here is that no interpretation of "preferences" can serve the EU theorists' purposes unless it preserves the plausibility of this fundamental connection between rational preferences and rational choice.

Moreover, there is another reason why this interpretation of "preferences" cannot really serve the EU theorist's purposes. Although one could introduce a notion of utility as a way of measuring these "preferences", it seems quite superfluous to do so, since when the term 'preference' is interpreted in this silly way, the very nature of "preferences" provides an adequate measure already – namely, the notion of a prospect's *expected monetary gain*. So if "preferences" are interpreted in this admittedly silly way, utility falls out of the picture, as a completely redundant element. Assuming that it is an essential part of EU theory that utility, understood as a measure of preference, plays a crucial role in the theory, then it seems that the theory must start out with a more intuitive notion of "preferences". It cannot start out by

[6] For a discussion of the St. Petersburg Paradox, see especially Joyce (1999, 32–38).

stipulating a definition of "preferences" in terms of the expectation of some quantity (so that to prefer *A* over *B* is to have a higher expectation of the relevant quantity for *A* than for *B*): any such stipulative definition would deprive the notion of "utility" of any crucial role in the theory.

So, what is this more intuitive notion of "preference"? The natural place to look is to the concepts that are expressed by 'preference' in ordinary English. The dictionaries all agree that the verb 'prefer' has two main senses that are in common use today:

1. In the first sense, to "prefer" a particular prospect over another is to *like* it more (or to *dislike* it less).
2. In the second sense, to "prefer" a particular prospect over another is to *choose* the first prospect over the second.

These two senses of 'prefer' can obviously come apart. One might like having cream in one's coffee more than drinking it black; but if one is under doctor's orders to avoid cream, then one might choose to drink one's coffee black rather than with cream. *Liking* one item more than a second seems to involve being *pleased* more by the first item than by the second – where being pleased by an item involves having a *feeling* or *emotion* of a certain sort towards that item.

A *choice*, on the other hand, seems to involve forming an *intention*. That is, it involves *committing oneself* to carry out or execute the intention. Normally, unless one is somehow involuntarily prevented from executing an intention, one will attempt to *execute* the intention at some point, unless the intention is either abandoned or forgotten. Choices can be understood to include *conditional* choices: I might conditionally choose *A* – that is, conditionally on a certain condition's obtaining. One special case of a conditional choice is when I conditionally choose *A*, on the condition that I choose either *A* or *B*. (As one might say, "If I choose either *A* or *B*, it will be *A*.") So one interpretation of a "preference" for *A* over *B* is as a conditional choice – a choice

to do *A* rather than *B* if one does either.

In ordinary English, then, a "preference" seems to be either a kind of liking or else a kind of choice. However, there is a further complication in the everyday meaning of 'preference' that we need to avoid being confused by. In some contexts, to say that a person "prefers" wine over beer is to say that the person has a *tendency* or *disposition* to like drinking wine more than drinking beer; and in some contexts, to say that a person "prefers" travelling by train over driving is to say that the person has a *disposition* to choose to travel by train rather than to drive. Let us call the mental characteristics that are ascribed by the term 'prefer' in these contexts "preference-dispositions".

It seems that the objects of these preference-dispositions are not *particular* prospects, but general *types* of prospects. An example of a *particular* prospect might be: your drinking the particular glass of wine that is on the table in front of you right now. A particular prospect is realized either once or not at all; most *types* of prospect can be realized many times over. For example, the general type of prospect *your drinking some wine* is realized on every occasion when you drink some wine. To say that you have a preference-disposition for one type of prospect over another seems to be to make some kind of generalization over your preferences for particular prospects. Roughly, it is to say that, in *normal* conditions in which prospects of both types are available, you like the prospect of the first type more than the prospect of the second type (or choose the prospect of the first type over the prospect of the second type). In classical decision theory, the objects of preference are thought of as particular prospects (not as general types of prospects). For this reason, we should set aside the use of the word 'prefer' to ascribe preference-dispositions, and just focus on uses that ascribe preferences for particular prospects.

For this reason, it is important to distinguish between preference-dispositions and actual

preferences – just as it is important to distinguish between being disposed to dissolve (that is, being soluble) and actually dissolving. In view of this difference, it seems unlikely that merely having a disposition to prefer prospects of one type T_A over prospects of a second type T_B will be subject to the same constraints of rationality as actually preferring *A* over *B*. In a similar way, the requirements of rationality (if any) that apply to mere dispositions to *believe* are surely weaker than the requirements that apply to actual beliefs. It is not clearly irrational to have both a disposition to have a degree of belief of 0.7 in *p* and a disposition to have a degree of belief of 0.4 in '¬*p*', so long as you never *simultaneously manifest* both of these dispositions. On the other hand, it is plausibly a requirement of rationality that the set of beliefs that you *actually* have should be probabilistically coherent, and so that if you have a degree of belief in both *p* and '¬*p*', they should together add up to 1. Since EU theorists claim that preferences are subject to highly demanding requirements of rationality, it seems that these "preferences" cannot be mere dispositions to have preferences – they must be *actual* preferences (even if having these preferences essentially involves other dispositions of various sorts). So, EU theorists should not interpret the "preferences" that their theory is concerned with as mere dispositions to have states that would count as "preferences" in some ordinary sense of the term.

If we set these preference-dispositions aside, then in ordinary English the term 'preference' refers either to a kind of *liking* or to a kind of *choice*. Although these two meanings seem to be the only relevant senses of the term 'preference' in everyday English, some EU theorists seem to use the term in a third sense. In this sense, a "preference" is a kind of *value-judgment*. In effect, according to this interpretation, to prefer *A* over *B* is to have some kind of *belief* that one might express by saying that *A* is "better" or "more desirable" than *B*. For example, as James M. Joyce (1999, 40) puts it, "We think of [preferences] as an agent's 'all-

things-considered' judgments about the desirability of wagers." In my opinion, it is doubtful whether the term 'preference' ever refers to value-judgments of this kind in ordinary English. Nonetheless, I shall consider this third interpretation of "preferences" as well. In short, I shall consider these three different senses that the term 'preference' can have: the liking sense; the choice sense; and the value-judgment sense.

Some theorists have in effect suggest that EU theory uses 'preference' in a special technical sense, as a term for a *sui generis* type of mental attitude, distinct from any of the kinds of mental states that are commonly called "preferences" in everyday folk-psychological discourse. Thus, Daniel Hausman (2013, 34) proposes that decision theorists should interpret "preferences" as "total subjective comparative evaluations", while he recognizes that "this notion of 'preference' does not conform to the ordinary usage of the word" (35). However, Hausman never makes it clear how, if at all, the "evaluations" that he is speaking of differ from some kind of "value-judgment". The mere fact that he uses a different word does not guarantee that he is picking out a different mental state. The best way to make it clear what mental state one is picking out is by characterizing the mental state's functional role. The main functional role that Hausman (2013, 34) ascribes to preference is simply that (at least if the agent is rational) preference "satisfies the axioms of ordinal utility, and combines with beliefs to determine choices".[7] But Joyce would also argue that the "value-judgments" that he identifies with preferences plays this functional role as well. So it is not clear how, if at all, these evaluations

[7] Hausman (2013, 34) says that preferences are "motivational ('conative') comparative attitudes. But many philosophers would regard certain value-judgments as "motivational" or "conative" in a sense; see for example Wedgwood (2007). So this also fails to distinguish Hausman's interpretation of preferences as "evaluations" from the Joyce's interpretation of preferences as value-judgments.

differ from value-judgments. For this reason, I shall treat Hausman's interpretation of preferences as a notational variant of Joyce's value-judgment interpretation.

There is one way of understanding "preferences" that has played an important role in the history of decision theory. This is a broadly *behaviourist* or "*interpretivist*" understanding of preferences, according to which for an agent to have a preference just is for it to follow from the best possible interpretation of that agent's behaviour that he has that preference.[8] It is assumed that the best possible interpretation of an agent's behaviour will have to meet the following desiderata, at least as far as possible:

i. The interpretation is consistent with the agent's observed choices;
ii. The interpretation implies that the agent's preferences obey the axioms of EU theory, and that the agent never chooses an option if there is an available alternative that she prefers;
iii. The interpretation ascribes preferences to the agent that are broadly speaking normal for agents in her circumstances.

While I do not dispute that this understanding of "preferences" may be very useful for some purposes – such as the purposes of economics and other similar branches of social science – I do not believe that this conception is what we need for the purposes of a *philosophical* theory of *rational choice*. As I understand it, if rationality exists at all, it is a property of *psychologically real* mental states and mental processes. The grounds that behaviourist or "interpretivist" theorists take to justify the ascription of preferences do not ensure that the agent has any corresponding psychologically real mental states; at best, they only ensure that the agent's

[8] See especially Maher (1993, chap. 1) for a lucid presentation of this sort of understanding of preferences. For some effective criticisms of this sort of interpretivism, see Byrne (1998).

observable behaviour is *as if* she had such psychologically real preferences.

In general, the interpretivist's understanding of preferences seems to be at home in a rather distinctive sort of intellectual inquiry. For this sort of intellectual inquiry, all that matters is that ascribing preferences of this sort yields a mathematically tractable way of generating roughly correct predictions of human behaviour. If you are pursuing an intellectual inquiry of this sort, then it need not matter to you what are the real processes of thought that the agent is going through. It also need not matter whether preferences that meet the axioms of EU theory are "rational" in any sense that is broadly continuous with our everyday practice of evaluating choices or decisions as "reasonable" or "unreasonable", "wise" or "foolish", or as made "with good reason" or "for no good reason", or the like. If you are pursuing an inquiry of this sort, then it might be convenient for you to sum up the point that a certain agent's preferences meet these axioms by saying that these preferences are "rational", but in saying this, you need not be using the term 'rational' to express a normative concept. It is not crucial to you whether there is anything *defective* or *wrong* with preferences that do not meet the axioms.

By contrast, the inquiry that I am interested in is concerned to develop a theory that answers the very same questions that people ask when engaged in everyday evaluation of choices and decisions. This kind of everyday evaluation of choices focuses on the kinds of mental states and processes that are identified in everyday folk-psychological discourse – and assumes that these mental states and processes are psychologically real phenomena, and not just postulations that are convenient for the purposes of predicting behaviour. Moreover, I am using the term 'rational' to express an intrinsically normative concept: for me, it is a conceptual truth that irrationality is a kind of *vice* or *defect* of thought.

It is these features of the kind of inquiry that I am pursuing that explain why I should

assume that, if rationality exists at all, it is a property of psychological real mental states and processes, of the kinds that are identified by ordinary folk-psychological discourse. Thus, the only concepts of a "preference" that we need to investigate are the ones that are present in folk-psychological discourse.

For this reason, it seems reasonable for me to limit my attention here to interpretations of "preferences" that identify them with mental states that are recognized in ordinary folk-psychological discourse. As I have explained, I shall focus on three such interpretations: the *liking* interpretation; the *choice* interpretation; and the *value-judgment* interpretation.

3. The structure axioms

As we have seen, the "structure" axioms require that for any two prospects *A* and *B* within the relevant domain, the rational agent must either prefer *A* over *B*, or prefer *B* over *A*, or be indifferent between the two of them. By itself, this requirement might be acceptable if the relevant domain of prospects were restricted to prospects that the agent has actually thought about or has attitudes towards. It is not obviously wrong to claim that an (ideally) rational agent will make up her mind about how every single one of the prospects that she has considered compares with every other prospect that she has considered.

However, the structure axioms also require that the relevant domain of prospects must have a certain *structure*. The details of these axioms vary between different proofs of the representation theorem, but all of these different versions agree that the relevant domain of prospects must be *infinite*.[9] Since they imply that the agent must have a complete set of

[9] In von Neumann and Morgenstern's theory, the domain of "gambles" or "lotteries" must be infinite, because for every possible payoff, it must contain one "gamble" that yields that payoff in every state of affairs, and

preferences over this domain, and that this domain must be infinite, the structure axioms have the explosive result that every rational agent must have *infinitely many preferences*.

As we have seen, it could still be *true* of an agent that their preferences can be represented by means of a utility function, even if the agent only had finitely many preferences – so long as the agent's preferences meet the pure coherence conditions. Still, there is no hope of giving a *general proof* that a rational agent's preferences can *always* be measured by a unique utility function (at least up to an arbitrary choice of a unit and zero point) unless the rational agent's preferences always give a complete ranking of an infinite domain of prospects.

However, this implication of the structures – that every rational agent must have a set of preferences that gives a total ordering of this infinite domain of prospects – seems implausible to me, on any plausible understanding of what it means for an agent to be "rational" in the relevant sense. It is not clear that it is even *metaphysically possible* for us – given that we are essentially human beings, composed of flesh and blood and the like – to have infinitely many preferences in this way. A preference is presumably a mental state that involves the agent's taking the attitude of preference towards a pair of prospects each of which is in some way *represented*, by means of a structured representation composed of concepts. So, having an attitude towards a pair of prospects involves an investment of cognitive resources, where there is presumably some

---

for every pair of gambles that it contains, it also contains every possible "mixture" of those gambles (see Joyce 1999, 24 and 42). In Savage's (1954) theory, the domain of prospects (or "acts") has to be infinite for much the same reason: for every consequence, it must contain a "constant act" that has that consequence in every state, and it must be closed under "mixing" on any "event", and the theory guarantees that there are infinitely many "events" (see Joyce 1999, 83 and 92). According to Jeffrey's (1983) theory, the objects of preference (or "prospects") are simply propositions, and the theory requires that the relevant domain of propositions should be "atomless" and so infinite (see Joyce 1999, 133).

minimal level of investment that is involved in the representation of every prospect to which we have any attitude at all. Since our cognitive resources are finite, it does not seem possible for us to have infinitely many preferences.[10]

I am assuming here that the notion of rationality is a *normative* notion. This means that the notion of what is "rationally required" of us is a kind of 'ought'. There are controversies about whether it is always true that 'ought' implies 'can', but it is surely true that at least in this case, 'ought' implies at least *metaphysical* possibility. So it seems very doubtful whether it can be rationally required of us to have infinitely many preferences in this way.[11]

A second reason against regarding these "structure axioms" as requirements of rationality is emphasized by Joyce (1999, 99–101). Some goods seem to be incommensurable: there is no unitary scale or measure of goodness that can rank every good in comparison with every other. Sometimes it might be true neither that *A* is better than *B*, nor that *B* is better than *A*, nor that *A* and *B* are equally good. An agent who believed that he was confronted with a case of this kind, Joyce suggests, might quite rationally have no preference at all between *A* and *B* – neither preferring *A* over *B*, nor preferring *B* over *A*, nor being indifferent between the two.

Much more could be said about this question, about whether every rational set of preferences must give a complete ranking of an infinite domain of prospects in this way. But it seems at the very least a significant cost of the structure axioms that they have this implication.

---

[10] A further reason for doubting whether any of us could have infinitely many preferences is that each of us possesses only finitely many concepts, and there may well be an upper bound to the *complexity* of the conceptually structured thoughts that we are capable of having attitudes towards.

[11] For some further discussion of the significance of our cognitive limitations for the theory of rational choice, see Wedgwood (2011).

This is why EU theorists – such as Joyce (1999, 97–104) – agree that completeness (at least in the context of an infinitely large domain of prospects) is implausible as a requirement of rationality. These EU theorists suggest reinterpreting completeness and the other structure axioms as requirements of "coherent extendibility". That is, we should reinterpret these axioms so that they only require that the rational agent's preferences can be *coherently extended* to a set of preferences that gives a complete ordering of an infinite domain of prospects.

How much is conceded by this reinterpretation of the structure axioms? According to a "realist" or "substantive" interpretation, it could still be the case that the agent has a unique utility function, even if this cannot be proved to be the case by means of the representation theorem; but according to a "formalist" or "operational" interpretation, if it cannot be proved by means of the representation theorem that the agent has a unique utility function, the agent does not have a unique utility function at all.

Without the completeness axiom, the representation theorem would have to be weakened so that it no longer claims that the rational agent's actual preferences can be measured by means of a *unique* utility function. It may well be that for many rational agents, there are infinitely many ways of coherently extending the rational agent's preferences into an infinite set, and each of these different coherent extensions would yield a *different* utility function. So on a formalist or operational interpretation, we would have to concede that even a perfectly rational set of preferences may not be measurable by a unique utility function at all. On this view, to maintain that preferences can be measured by a unique utility function would involve committing what Mark Kaplan (1994, 23–31) has called the "sin of false precision".

The most that can be proved, on this interpretation, is that the rational agent's preferences can be measured by a *set* of utility functions – and if this set contains more than one utility

function, it must in fact contain infinitely many such functions. However, even if these preferences cannot be measured on a unique interval scale, perhaps they can in a sense be "measured" by the whole *set* of utility functions that can represent those preferences. It may be true that the degree to which the agent's preference for *A* over *B* is stronger or weaker than the agent's preference for *C* over *D* cannot be identified with the *unique* ratio of the difference between the utility of *A* and the utility of *B* to the difference between the utility of *C* and the utility of *D*. But it can at least be represented by the *set* of the ratios of these differences according to *each* of the many utility functions that can represent the agent's preferences.

This approach can still say that if an agent has a system of rational preferences of this sort, then it is rational for the agent to choose an option *A* if and only if there is no available alternative option *B* such that the agent prefers *B* over *A*. On this approach, the agent prefers *B* over *A* if and only if *all* of these utility functions assign a higher utility to *B* than to *A*; so it will follow that for the agent *not* to prefer *B* over *A* is for it *not* to be the case that all of these utility functions assign a higher utility to *B* than to *A* – that is, for it to be the case that *at least one* of these utility functions assigns a utility to *A* that is equal to or higher than the utility that it assigns to *B*. Thus, on this approach it is rational to choose an option *A* if and only if at least one of these utility functions assigns maximal utility to *A*.

In this way, a number of EU theorists have thought that, even on the formalist or operational interpretation of utility, they could reinterpret completeness and the other "structure" axioms as requirements of *coherent extendibility*. Still, even these EU theorists need it to be true that rational preferences must meet the axioms of "pure coherence". In the next three sections, however, I shall argue that on all the interpretations of "preferences" that do not render the notion of "utility" entirely redundant, the rational agent's preferences need not meet these pure

coherence axioms either.

4. Preferences as likings

In Section 2, I distinguished three possible interpretations of "preferences": the liking interpretation, the choice interpretation, and the value-judgment interpretation. In this section, I shall take each of these three interpretations in turn. We shall find that only one of these interpretations supports the EU theorists' claim that preferences can be measured by means of a utility function; but that interpretation undermines the neo-Humean definition of rational choices as choices that maximize expected utility.

I shall start by considering the "liking" interpretation of preferences. As we have noted, likings seem to be closely akin to desires and emotions. If you also experience *A* as actually being the case, you will typically be *pleased* that *A* rather than *B* is the case, while if you experience *B* as actually being the case, you will typically be *displeased* that *B* rather than *A* is the case – where being pleased or displeased that something is the case is broadly speaking a kind of *emotion*. Such likings also seem to involve *desires* of a sort that are closely related to emotions: if you like *A* more than *B*, then you will have a certain sort of *desire* for *A* rather than *B* to be the case, if either is case.

Emotions and desires of the sort that are closely related to emotions do not respond to *all* the relevant factors that bear on the choices that the agent has to make. By their nature, emotions are relatively automatic fast-track responses to information, resulting in changes in our attention, and priming us to react in ways that are often – but not always – appropriate to the whole truth about our situation. In this way, an emotion is in a sense a "partial" evaluation of its objects.

As decision theorists have recognized, if preferences are to play the role that EU theory demands of them, preferences must be *total* evaluations. As Hausman (2013, 34) says: "To say

that Jill prefers *x* to *y* is to say that when Jill has thought about everything she takes to bear on how much she values *x* and *y*, Jill ranks *x* above *y*. … Jill's total subjective ranking does not leave out anything that she regards as relevant to the evaluation of alternatives." Since emotions and desires that are closely akin to emotions are partial evaluations of their objects, they seem unlikely to be able to play the role of preferences.

One sign of this is that a person's likes and dislikes will very often conflict with each other: it seems that this is not irrational, but simply the natural condition of human life. I might both like and dislike a certain person's company – I like it to the extent that the person is clever, witty, and stimulating, but simultaneously dislike his company to the extent that the person is callous and heartless in his attitudes towards others. Similarly, if desires are closely related to likes and dislikes, then conflicting desires are not irrational. For example, you might be attracted to the prospect of reading John Rawls's *Theory of Justice* from cover to cover – the experience is sure to be deeply illuminating and stimulating; but you might simultaneously be somewhat repelled by the prospect – it will take all the will-power that you to possess to maintain your concentration over all of those hundreds of pages of earnestly wordy argument..[12]

These likings are not irrational, and yet they do not create any consistent ranking of their objects. You could (in one way) like *A* more than *B* at the same time as (in another way) liking *B*

[12] Of course, you might have an unalloyed attitude of *liking* the prospect of getting the philosophical illumination that Rawls's *Theory of Justice* could provide more than the prospect of *not* that philosophical illumination, and an unalloyed attitude of *disliking* the prospect of struggling through 600 pages of Rawls's earnest wordy style more than the prospect of not struggling through those 600 pages. But these are simply different pairs of prospects. The question remains: What is your attitude towards the pair of prospects that consist in (a) your reading *A Theory of Justice* from cover to cover, and (b) your not reading it from cover to cover? A rational agent surely might have thoroughly mixed feelings about this pair of prospects.

more than *A*. So even in a rational person, such likings need not satisfy the axioms of EU theory. Some philosophers might accept that preferences are a kind of liking or desire, but insist that preferences are "total" or "all-things-considered desires". But if preferences are "all-things-considered" conative states, they must surely be the result of some kind of *reasoning*. It seems essential to likings and emotions that they are not the result of such all-things-considered reasoning. In this respect, likings and emotions seem like experiences and sensations: they arise from a cognitively insulated mental module, in a way that is largely exogenous to our reasoning system. So if the idea of an "all-things-considered desire" makes sense, it seems that any such desire will have to be a kind of state that can result from reasoning of some sort. But this in effect makes such "all-things-considered desires" closer to a kind of *choice* or *judgment*. So the idea that preferences are "all-things-considered desires" seems to be a version of either the choice interpretation or the value-judgment interpretation of preferences.

5. Preferences as choices

Since the "liking" interpretation is so unpromising, I turn now to the "choice" interpretation. I have already explained why I believe that we cannot accept the interpretation of preferences as a mere *disposition* to make certain choices: a mere disposition to have mental states of a certain kind is not plausibly subject to the same requirements of rationality as an actual mental state of that kind. So it seems that the only version of the choice interpretation that we need to consider is the view according to which a preference is a *conditional choice*: that is, a preference for *A* over *B* is a conditional choice for *A* over *B* – the choice to go for *A* and not *B*, if one goes for either.

There is, however, an obvious problem with the choice view – a problem that is analogous to a well-known problem with the suggestion that the authoritative *test* for whether you prefer *A* over *B* is whether or not you choose *A* over *B*. To quote Leonard Savage (1954, 17):

> This procedure for testing preference is not entirely adequate, if only because it fails to take account of, or even define, the possibility that the person may not really have any preference between *A* and *B*, regarding them as equivalent; in which case his choice of *A* should not be regarded as significant.

Even if you are indifferent between *A* over *B*, you may still have to choose between *A* and *B*. If we claimed that an agent "preferred" *A* over *B* whenever the agent chose *A* over *B*, then these "preferences" would include some cases where the agent in fact regarded the two options as equally good in every way. It seems clear that "preferences" of this sort need not satisfy the axioms. If you regard *A* and *B* as equally good, and choose *A* over *B* simply as an arbitrary choice to enable you to resolve a "Buridan's Ass" problem, there is no reason to expect this arbitrary "preference" to satisfy the axiom of independence: because your choice of *A* over *B* is completely arbitrary, there is no reason to expect that you will also prefer the gamble "*A* if *p* and otherwise *C*" over the gamble "*B* if *p* and otherwise *C*".

In fact, however, there is an easy solution to this problem (although I do not know of any EU theorist who has identified this solution). We just have to give a more complicated account of choice. We could say that making a choice in fact involves two elements. First, out of the options that one has considered, one must identify a (proper or improper) subset of these options as the *choice set* – that is, as the set of *eligible* options. Then, if the choice set has more than one member, one must just arbitrarily *pick* a member of the choice set. The choice interpretation of preferences should define preferences in terms of this first element of choice rather than in terms of the second. So to prefer *A* over *B* is to make a conditional choice to put *A* in the choice set rather than *B* if one puts either in the choice set.

However, there is also a deeper problem with the choice interpretation of preferences. It

overlooks the fact that the way in which a rational agent will choose between two options *A* and *B* may depend on what other options are available. This point emerges especially clearly in the cases that are often supposed to illustrate the phenomenon of the "incommensurability" of values. Suppose that *A* and *B* are two options that are very different from each other. For example, following John Broome (1997), we might imagine that *A* is a career in the army, while *B* is a career in the church. In some cases, it might seem that it is not true – or at least not determinately true – either that *A* is better than *B* or that *B* is better than *A*. In this case, it might also seem true that *A* and *B* are not equally good either, since if we suppose that a third option *A*+ is available, where *A*+ is just like *A* except that it is "sweetened" in some way – it involves a salary that is $500 greater, for example – it seems undeniable that *A*+ as better than *A*, but it does not seem to follow from this that *A*+ is better than *B*.[13]

First, consider the choice situation in which the only available options are *A* and *B*. In this case, a rational agent would presumably assign both *A* and *B* to the choice set of eligible options; so on the latest version of the choice interpretation of "preference", the agent does not prefer either *A* or *B* over the other. But now consider a second choice situation in which all three options, *A*+, *A*, and *B*, are available. In this second choice situation, *A*+ and *B* are in the choice set, but *A* is not; *A* is ruled out as ineligible. So in this second choice situation, given the "choice" interpretation of preferences, the agent *does* prefer *B* over *A*.

This implication of the "choice" interpretation of preferences is fatal for EU theory. To capture the fact that *B* is preferred to *A* in the second choice situation, all utility functions capable of representing the agent's preferences must assign a *higher* utility to *B* than to *A*; but to capture the fact that that neither *A* nor *B* is preferred over the other in the first choice situation, it must

[13] For an illuminating discussion of these cases, see also Hare (2010).

not be true that all these utility functions assign a higher utility to *B* than to *A*. In short, EU theory must suppose that each prospect has a definite utility that is independent of the different choice situations in which the prospect is available; it cannot allow that one and the same prospect has one utility in one choice situation, but a quite different utility in the second choice situation.

Some EU theorists might suggest that prospects are in fact *individuated* by the choice situations in which they appear, so that it is in fact impossible for one and the same prospect to be available in more than one choice situation. That is, *A*-when-the-available-alternatives-are-*A*+-and-*B* is simply a different prospect from *A*-when-the-only-available-alternatives-is-*B*. But if the EU theorist accepts the formalist or operational interpretation of utility, this is also a fatal move for them to make. As we have seen, the proof of the representation theorem crucially relies on a "structure axiom" to the effect that the domain of prospects must include every possible outcome and every possible lottery over outcomes – where the outcomes are assumed to be the very same outcomes whatever lottery they are embedded in (whether the lottery in question is the "constant" lottery, which yields that outcome in every possible state of affairs, or a highly risky lottery that leads to dramatically different outcomes in different states of affairs). If it is in fact *impossible* to embed one and the same outcome into different lotteries, then it will not be possible "coherently to extend" the preferences of rational agents to a total ordering over this enormous domain of prospects and lotteries – since this domain of prospects and lotteries will not even exist. If the EU theorist cannot reinterpret the "structure axioms" at least as the requirement that the agent's preferences should be capable of being coherently extended in this way, there is no way of proving that the agent's preferences can even be represented by a set of utility functions.

At this point, the EU theorist might be tempted to abandon the formalist or operational interpretation of utility in favour of the rival realist or substantive interpretation. But this substantive interpretation is particularly implausible in the context of the "choice" interpretation of preference. It may be a real mental fact about one's desires that one desires one thing "more strongly" than another; but what sense can we attach to the idea that one chooses *A* over *B* "more strongly" than one chooses *C* over *D*? So the choice interpretation of preference cannot easily be combined with the realist or substantive interpretation of utility.

For these reasons, then, the choice interpretation of preferences also seems not to serve the purposes of the EU theorists. I shall now consider the remaining interpretation of what preferences are, according to which preferences are judgments of desirability – that is, value-judgments of some kind.

6. Preferences as value-judgments

According to the value-judgment interpretation, a "preference" for *A* over *B* is a judgment to the effect that *A* is in the relevant way *more desirable* or *better* than *B*. I shall assume here that these value-judgments are a kind of *belief*. So, according to this interpretation, to prefer *A* over *B* is to *believe* that *A* is better than *B*.

There is an obvious problem with the idea that a preference for *A* over *B* is a belief to the effect that *A* is better than *B*. Beliefs themselves come in degrees. So how strongly does one have to believe that *A* is better than *B*, in order to count as "preferring" *A* over *B*? We certainly cannot say that the degree to which one believes *A* to be better than *B* corresponds to the degree to which one prefers *A* over *B*: one might be utterly certain that *A* is very slightly better than *B* – which should surely count as a very weak preference for *A* over *B*, even though it involves the very highest possible degree of belief.

An alternative suggestion is that to prefer *A* over *B* is to have a higher degree of belief in the proposition that *A* is better than *B* than in the proposition that *A* is not better than *B*. But this suggestion also seems wrong. Suppose that although the rational agent prefers *A* over *B* in this sense, the agent has only a *very slightly* greater degree of belief in the proposition that *A* is better than *B* than in the proposition that *A* is not better than *B*; but the agent is also certain that if *A* is better, it is only *very slightly* better than *B*, whereas if *A* is not better, then *B* is *dramatically* better than *A*. Then it seems that the rational agent would choose *B* rather than *A*, even though the agent "prefers" *A* over *B*. But this contradicts the EU theorist's fundamental assumption that a rational agent will choose an option *B* only if there is no alternative *A* such that she prefers *A* over *B*. So this second version of the value-judgment interpretation of "preference" also cannot serve the EU theorist's purposes.

Some philosophers may be tempted to say that there must be some sense of the term 'better' on which, in the case that I have just described, *B* is better than *A*. (After all, the rational agent chooses *B* over *A* – so surely *B* is better in some sense?) Whatever exactly this sense is, the rational agent could presumably believe that *B* is better than *A* in this sense; and so perhaps we should identify the agent's preference with this belief? But this suggestion cannot solve the problem unless a rational agent can always be *certain* of whether or not *A* is (in the relevant sense) "better" than *B*. If an agent is ever rationally required to be less than perfectly certain of this proposition, we can imagine a case like the case that I described, where it is not rational for the agent to choose the option that she "prefers" in this sense. But the assumption that a rational agent can always be absolutely certain of such propositions seems extremely dubious.

Even if the rational agent is always certain of these propositions, why should these beliefs obey all the pure coherence axioms of EU theory? Only one answer suggests itself: in the

relevant sense, for one prospect *A* to be "better" than a second prospect *B* is for *A* to have a higher degree of *expected goodness* than *B*; for the rational agent to be *certain* of all the relevant truths about what is in this sense "better" than what, these expectations of goodness would have to be defined in terms of a probability function P such that for every relevant proposition *p*, and every number *n*, whenever P(*p*) = *n*, the agent is certain that P(*p*) = *n*. Given the controversial assumption that there is a probability function of this sort, we could now identify the state of preferring *A* over *B* with the belief (held with certainty) that *A* is in this sense better than *B*. However, this approach now seems to make the appeal to preferences, and to utility as a measure of preference, quite redundant: on this approach, rational choices are fundamentally guided by expected goodness; there is no need to bring in any talk of "preferences" or "utility" here at all.

The same result follows if the rational agent is *not* always certain of these propositions about which of the available prospects are in the relevant sense better than the alternatives. In that case, it seems that the agent's choices will have to be guided, not by beliefs that are held with certainty, but a range of *partial* beliefs about the degree of goodness that each of those prospects will have. The natural way for agents to be guided by this range of partial beliefs is for those agents to be guided by their *expectations* of these prospects' degree of goodness. Even if the rational agent is uncertain about exactly how good the various available prospects are, she may have various degrees of belief in various *hypotheses* about the degree of goodness that these prospects have. In particular, she may have various degrees of belief in certain hypotheses of this sort that collectively form a *partition* – that is, a set of exhaustive and mutually exclusive propositions such that the agent is quite certain that one and only one of these propositions is true. Then the degree of goodness that each of these prospects has according of each of these hypotheses can be weighted by the degree of belief that the agent has in the hypothesis; and the

weighted sum of these degrees of goodness can be identified with the prospect's degree of expected goodness. For example, if the agent has a 0.5 degree of belief that the prospect is good to degree 0, and a 0.5 degree of belief that the prospect is good to degree 10, then the agent's expectation of the prospect's degree of goodness will be 5.

A theorist might now identify the state of "preferring" *A* over *B* with the state of having a higher expectation of goodness for *A* than for *B*. This identification of "preferences" with this fact about the agent's expectations would clearly ensure that these "preferences" do indeed satisfy all the pure coherence axioms of EU theory. At least so long as every rational agent's degrees of belief are probabilistically coherent, this suggestion would indeed guarantee that "preferences" meet all the coherence axioms of EU theory. For example, if "preferences" are interpreted in this way, they will certainly be transitive: if a rational agent has a higher expectation of goodness for *A* than for *B*, and also has a higher expectation of goodness for *B* than for *C*, the agent must also have a higher expectation of goodness for *A* than for *C*. A similar point can be made about the other coherence axioms.

Equally clearly, however, this identification of preferences would make the appeal to preferences and to utility as a measure of preference quite redundant. On this interpretation, it is fundamentally the rational agent's expectation of goodness that guides her choices and intentions. In other words, this interpretation is not a version of the neo-Humean theory of rational choice; it is a version of the rival neo-Aristotelian theory of rational choice instead.

In this respect, this identification of preferences would be just like the move that we contemplated at the beginning of Section 2 – the move of identifying a preference for *A* over *B* with the state of having a higher expectation of monetary gain for *A* than for *B*. Admittedly, unlike that move, this suggestion would not obviously undermine the fundamental connection

between rational preference and rational choice. But it would still render the notion of a "preference" and of "utility" completely redundant and superfluous elements in the account of rational choice. On this picture, the key point is not that the rational agent has measurable preferences – whatever exactly they may be. The point is that the rational agent has rational degrees of belief in evaluative propositions, about the degree of goodness that the available prospects have. It is these degrees of belief in evaluative propositions that determine each of the relevant prospects' expected goodness, which is what guides the rational agent in making her choices, and in forming and revising her intentions. Preference and utility fall away, as entirely unnecessary elements of this account of rational choice and rational intention.

In this way, the value-judgment interpretation of preferences is just a prelude to the dénouement of our discussion, in which preferences and utility usher themselves off the scene. The only interpretation that we have found that makes it plausible that the rational agent's preferences must meet the coherence axioms of EU is also an interpretation that makes preferences and utility completely redundant as elements of the definition of rational choice. The correct definition of a rational choice is as a choice that maximizes some kind of *expected goodness* – not as a choice that maximizes expected utility.

7. Rational intentions maximize expected choiceworthiness

The results of the foregoing discussion are not merely negative. They make it plausible that we should abandon the broadly neo-Humean approach that interprets practical reasoning as pursuing the goal of preference-satisfaction. Instead, we should embrace a neo-Aristotelian view. According to this neo-Aristotelian view, there is some genuinely evaluative concept – the concept of a course of action that is *good* or *valuable* in the relevant way – such that the rational agent must form degrees of belief in various hypotheses about the extent to which the available

options are good or valuable in this way; then the rational agent will be guided by these degrees of belief, in such a way that she chooses a course of action that maximizes the expectation of the relevant sort of goodness or value.

What is the relevant concept of a "good course of action"? There are of course many concepts that can be expressed by the word 'good',[14] depending on the context in which the word appears, and it is certainly not every concept expressible by 'good' that can play this role in an account of rational choice. Indeed, we might wonder how any notion of goodness can play this role. For any notion of good that is prior to our theory of rational choice, it seems rationally permissible to be risk-averse about it: for instance, it seems rationally permissible to choose 100 units of this sort of good for sure over a gamble on 0 units and 201 units at equal odds.[15]

The answer must be that the relevant kind of goodness is somehow a special kind, which is somehow tailor-made for the purposes of a theory of rational choice. Just to give it a label, I shall call this kind of goodness "choiceworthiness". If there is such a notion of goodness, we could define a rational choice as one that *maximizes expected choiceworthiness*. So one way to make progress towards a better understanding of rational choice is to investigate what kind of concept "choiceworthiness" will have to be if it is capable of playing this role. I shall illustrate this point by giving one point about the nature of choiceworthiness.

In Section 5 above, we looked at an example of the kind that is often raised in discussions of "incommensurability". This example involves a contrasting pair of cases: in the first case, there are two radically different options *A* and *B*, that are valuable in profoundly different ways;

[14] This point is rightly emphasized by Thomson (1997).

[15] This point is one of the less controversial lessons of the debate about the "Allais paradox"; see e.g. Broome (1991) and Weber (1998).

and the second case is as much as possible like the first except that there is also a third available option *A*+, which is in effect a "sweetened" variant of *A*. We may suppose that there is no relevant uncertainty in either of these cases: so each option's degree of actual choiceworthiness is the same as its degree of expected choiceworthiness.

To deal with these cases, it must be that choiceworthiness sometimes gives only a partial ranking of the available options. So in the first case, neither *A* nor *B* is more choiceworthy than the other, although they are also not exactly equally choiceworthy either. However, since *A* and *B* are both maximally choiceworthy, they are both rational options to choose. In the second case, *A*+ and *B* are both maximally choiceworthy in the same way as *A* and *B* were in the first case; but in this second case, *A*+ and *B* are both *more* choiceworthy than *A*. So, it can happen that *B* is more choiceworthy than *A* in the second choice situation, but not in the first choice situation. In this way, the choiceworthiness of options is relative to choice situations.[16]

This is just one of the many points about choiceworthiness that can be derived from the premise that choiceworthiness is the evaluative concept that is capable of playing this sort of role in an account of rational choice. Much more investigation is required to develop a complete account of the nature of choiceworthiness; it would also be crucial to develop a complete account of the relevant sort of *expectation*, and to explain why it is this sort of expectation that plays a central role in the correct account of rational choice. Nonetheless, I believe that the discussion so far has already made it plausible that this neo-Aristotelian idea that a rational choice maximizes expected choiceworthiness is much more promising than the broadly neo-Humean idea of

[16] As I have argued elsewhere, in effect, this point – that the value that a rational choice must maximize the expectation of is relative to choice situations in this way – can help us to achieve a new and better solution to the Newcomb paradox; see Wedgwood (2013).

maximizing expected utility.

References

Bermudez, José Luis (2009). *Decision Theory and Rationality* (Oxford: Oxford University Press).

Broome, John (1991). "Rationality and the Sure-Thing Principle", in *Thoughtful Economic Man*, ed. Gay Meeks (Cambridge: Cambridge University Press): 74–102.

———— (1997). "Is Incommensurability Vagueness?" in *Incommensurability, Incomparability, and Practical Reason*, ed. Ruth Chang (Cambridge, MA: Harvard University Press).

Buchak, Lara (2013). *Risk and Rationality* (Oxford: Oxford University Press).

Byrne, Alex (1998). "Interpretivism", *European Review of Philosophy* 3, 199–223.

Hare, Caspar (2010). "Take the Sugar", *Analysis* 70: 237–47.

Hausman, Daniel M. (2012). *Preference, Value, Choice, and Welfare* (Cambridge: Cambridge University Press).

Jeffrey, Richard (1983). *The Logic of Decision*, 2nd edition (Chicago: University of Chicago Press).

Joyce, James M. (1999). *Foundations of Causal Decision Theory* (Cambridge: Cambridge University Press).

Kaplan, Mark (1996). *Decision Theory as Philosophy* (Cambridge: Cambridge University Press).

Lewis, David (1981) "Causal Decision Theory", *Australasian Journal of Philosophy* 59: 5–30. Reprinted in Lewis (1985, 305–337).

———— (1988). "Desire as Belief", *Mind* 97: 323–32. Reprinted in Lewis (2000, 42–54).

———— (1985). *Philosophical Papers*, Vol. II (Oxford: Clarendon Press).

———— (2000). *Papers in Ethics and Social Philosophy* (Cambridge: Cambridge University Press).

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

Savage, Leonard (1954). *Foundations of Statistics* (New York: John Wiley and Sons).

Thomson, Judith Jarvis (1997). "The Right and the Good", *Journal of Philosophy* 94: 273–98.

Weber, Michael (1998). "The Resilience of the Allais Paradox", *Ethics* 109: 94–118.

Wedgwood, Ralph (2003). "Choosing Rationally and Choosing Correctly", in *Weakness of Will and Practical Irrationality*, ed. Sarah Stroud and Christine Tappolet (Oxford: Oxford University Press): 201–229.

———— (2007). *The Nature of Normativity* (Oxford: Clarendon Press).

———— (2011). "Instrumental Rationality", *Oxford Studies in Metaethics*, ed. Russ Shafer-Landau, vol. 6: 280–309.

———— (2013). "Gandalf's Solution to the Newcomb Problem", *Synthese* 190, no. 14: 2643–2675.