RISK AND TRADEOFFS

Lara Buchak, UC Berkeley[1]

1. Introduction

Decision theories are theories of instrumental rationality: they formalize constraints of consistency between rational agents' ends and the means they take to arrive at these ends. We can model the possible actions an agent might take each as a gamble whose outcomes depend on the state of the world: for example, the action of not bringing an umbrella is a gamble that results in getting wet if it rains and staying dry if it doesn't. Decision theory places constraints on the structure of rational agents' preferences among the actions available to them and, as a result, can represent the beliefs and desires of any agent who meets these constraints by precise numerical values.

The prevailing view is that *subjective expected utility theory*, which dictates that agents prefer the gamble with the highest expected utility, is the correct theory of instrumental rationality. Subjective expected utility theory (hereafter, EU theory) is thought to characterize the preferences of all rational decision makers. And yet, there are some preferences that violate EU theory that seem both intuitively appealing and prima facie consistent. An important group of these preferences stem from how ordinary decision makers take risk into account: ordinary decision makers seem to care about "global" properties of gambles, but EU theory rules out their doing so.

If one is sympathetic to the general aim of decision theory, there are three potential lines of response to the fact that EU theory does not capture the way that many people take risk into account in their preferences among gambles. The first is to claim that contrary to initial appearances, expected utility theory can represent agents who care about global properties, by re-describing the outcomes that these agents face. The second response is to claim that while many people care about global properties (and that these patterns of preferences cannot be represented by the theory), these people are simply not rational in doing so. I think that neither of these responses can succeed. I advocate a third response: modifying our normative theory to broaden the range of rationally permissible preferences. In particular, I advocate broadening the set of attitudes towards risk that count as rationally permissible. Although I won't directly argue against the first two responses here, formulating an alternative will be important to evaluating them. We need to know what it is that agents are doing when they systematically violate EU theory in order to discover whether doing what they are doing constitutes taking the means to their ends in a rationally permissible way. In this paper, I will explain the alternative to EU theory that I favor, and I will in particular explain how it does a better job at explicating the components of instrumental rationality than does EU theory.

[1] Penultimate version. Forthcoming in *Erkenntnis (FEW 2012 special issue)*: please cite that version. This article is excerpted in large part from Lara Buchak (2013), Risk and Rationality, Oxford University Press.

2. Risk, EU Theory, and Instrumental Rationality

I begin by briefly explaining subjective expected utility theory; explaining how it must analyze the phenomenon of risk aversion; and showing that as a result, EU theory cannot capture certain preferences that many people have. I will then argue that this problem arises for EU theory because it neglects an important component of instrumental rationality.

EU theory says that rational agents maximize expected utility: they prefer the act with the highest mathematical expectation of utility, relative to their utility and credence (subjective probability) functions. So, if we think of an act as a gamble that yields a particular outcome in each state of the world—for example, g = {E_1, x_1; E_2, x_2; …; E_n, x_n} is the act that yields x_i if E_i obtains, for each *i* —then the value of this act is:

$$EU(g) = \sum_{i=1}^{n} p(E_i)\, u(x_i)$$

According to EU theory, a rational agent strictly prefers *f* to *g* if and only if EU(f) > EU(g), and if she weakly prefers *f* to *g* if and only if EU(f) ≥ EU(g). So utility and credence are linked to rational preferences in the following way: if we know what an agent's utility function and credence function are, we can say what her preferences ought to be. They are also linked in another way that will be of central interest in this paper: if we know an agent's preferences, and if these preferences conform to the axioms of EU theory, then we can determine her credence function uniquely and her utility function uniquely up to positive affine transformation:[2] we can represent her as an expected utility maximizer relative to a some particular *p* and *u*. It is crucial for the EU theorist that the preferences of all rational agents can be represented in this way.

It is uncontroversial that many people's preferences display risk aversion in the following sense: an individual would rather have $50 than a fair coin-flip between $0 and $100, and, in general, would prefer to receive $*z* rather than to receive a gamble that will yield $*z* on average.[3] If such an agent is representable as an EU maximizer and the preference for $50 rather than the coin-flip is strict, then it must be that u($50) – u($0) > u($100) – u($50), i.e., that getting the first $50 makes more of a utility

[2] To say that two utility functions u(x) and u'(x) are equivalent up to positive affine transformation means that there are some constants *a* and *b* where *a* is positive and au(x) + b = u'(x).

[3] In all these examples, I will assume that probabilities are given, to simplify the discussion. But the probabilities involved should be assumed to be the agent's subjective probabilities. Moreover, I will use the term "risk-averse" neutrally: an agent is risk averse with respect to some good (say, money) iff she prefers a sure-thing amount of that good to a gamble with an equivalent mathematical expectation of that good. For a more general definition of risk aversion that is compatible with what I say here and that captures the idea that a risk-averse person prefers a gamble that is less spread out, see M. Rothschild and J. Stiglitz (1970), "Increasing Risk: I. A Definition," *Journal of Economic Theory.*

difference than getting the second \$50 does. More generally, her utility function in money must diminish marginally:

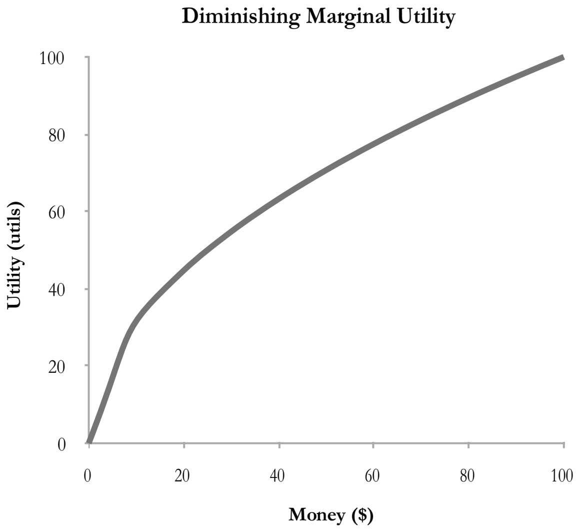

Diagram 1: Diminishing Marginal Utility

Diagram 1: Diminishing Marginal Utility

The converse holds as well: when a utility function is concave, a gamble will always have an expected utility that is no greater than the utility of its expected dollar value.[4] On EU theory, then, aversion to risk is equivalent to diminishing marginal utility.

Intuitively, though, there are two different psychological phenomena that could give rise to risk-averse behavior. On the one hand, how much one values additional amounts of money might diminish the more money one has. As an extreme case of this, consider Alice, who needs exactly \$50 for a bus ticket, and doesn't have much to buy beyond that. On the other hand, one might value small amounts of money linearly, but care about other properties of the gamble besides its average value: for example, the

[4] A real-valued function f is concave in some interval C iff $\forall x,y\in C$ and all $\alpha\in[0, 1]$, $f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y)$. A continuous function is concave in some interval C iff $\forall x,y\in C$, $f((x + y)/2) \geq (f(x) + f(y))/2$. Concavity is strict if the inequality is strict for $x \neq y$ and $\alpha\in(0, 1)$. Definitions of convexity and strict convexity are given by reversing the inequalities. Throughout this paper, I will continue to use "concave" to refer to the definition that uses weak inequalities, so that linearity is a degenerate case of concavity. This are the standard usage. Jensen's Inequality (Jenson 1906) tells us that if u is concave, then the expected utility of a random variable (in this case, amounts of money) is no greater than the utility of its expected monetary value: if f is a gamble, then $EU(f) \leq u(E(f))$, where $E(f)$ is the expected monetary value of f. And if u is convex, $EU(f) \geq u(E(f))$. Chateauneuf and Cohen (1994: 82) note that a preference for a guaranteed E(f) rather than f itself implies a concave utility function. But I do not know where this was originally shown. J.L.W.V. Jensen (1906). "Sur les fonctions convexes et les inégalités entre les valeurs moyennes." *Acta Mathematica* 30(1): 175-193. Alain Chateauneuf and Michèle Cohen (1994). "Risk-seeking with Diminishing Marginal Utility in a Non-Expected Utility Model." *Journal of Risk and Uncertainty* 9: 77-91.

minimum value it might yield, the maximum, or the spread of possible outcomes. In other words, one might be sensitive to *global properties*. Consider Bob, who gets as much pleasure out of the first $50 as the second, but would rather guarantee himself $50 than risk having nothing for the possibility of $100. Both Alice and Bob prefer $50 to the coin-flip, and the EU theorist must interpret both agents as having a non-linear utility function, on the basis of this preference.

What is the relationship between an agent's psychology and the utility function that is derived from her preferences? There are two views about this: on the *realistic* picture, the utility function represents some pre-existing value that the agent has; furthermore, this value is independent from preferences, so could in principle be known apart from preferences. For the realist, EU theory will have misinterpreted Bob, since Bob's strength of desire for money is linear. On the *constructivist* picture, which seems to have more widespread endorsement among contemporary philosophers, utility cannot be determined independently of what the agent prefers.[5] The utility function is a construction from preferences: it is the quantity whose mathematical expectation an agent maximizes. Utility may or may not correspond to some psychological quantity, but even if it does, it is not a quantity that we can give content to apart from preferences. So the constructivist won't necessarily care about the agent's own claims about how much she values various outcomes. There is nothing yet in Alice's and Bob's preferences to distinguish them from the point of view of constructivist EU theory. However, an agent's reasons for the preferences she has will turn out to matter, because if the constructivist EU theorist accidentally commits himself to something false about an agent's reasons, then although he will be able to explain an isolated preference (such as that for $50 over the $0/$100 coin-flip), his explanation will commit the agent to having preferences that she does not in fact have and will therefore fail to represent her.

Matthew Rabin presents a "calibration theorem" to show that in order to describe the preferences of decision makers that display risk aversion in modest-stakes gambles, EU theory is committed to absurd conclusions about preferences between gambles when the stakes are higher (absurd in the sense that no one actually has these preferences). As mentioned, on EU theory, modest stakes risk-aversion entails a concave utility function. Rabin's results assume nothing about the utility function except that it continues to be concave in higher stakes, and so doesn't, for example, have an inflection point above which it increases marginally. Here are some examples of the sorts of conclusions EU theory is committed to.[6] If

[5] Particularly clear expositions of this view appear in the following: (1) Patrick Maher (1993), Betting on Theories, Cambridge: Cambridge University Press; (2) John Broome (1999), "Utility," in Ethics out of Economics, Port Chester, NY, USA: Cambridge University Press; (3) James Dreier (2004), "Decision Theory and Morality," Chapter 9 of Oxford Handbook of Rationality, eds. Alfred R. Mele and Piers Rawling, Oxford University Press.

[6] Matthew Rabin (2000), "Risk Aversion and Expected Utility Theory: A Calibration Theorem," *Econometrica* 68(5): 1281-1292. Results on p. 1282.

an agent prefers not to take a fair coin-flip between losing \$100 and gaining \$110 (that is, if she prefers a sure-thing \$0 to the -\$100/\$110 coin-flip), regardless of her initial wealth level, then she must also prefer not to take a coin-flip between losing \$1,000 and gaining *any amount* of money. Similarly, if an agent prefers not to take a coin-flip between losing \$1,000 and gaining \$1050 for any initial wealth level, then she will also prefer not to take a coin-flip between losing \$20,000 and gaining any amount of money. Furthermore, if an agent prefers not to take a coin-flip between losing \$100 and gaining \$105 as long as her lifetime wealth is less than \$350,000, then from an initial wealth level of \$340,000, she will turn down a coin-flip between losing \$4,000 and gaining \$635,670. In other words, she will prefer a sure-thing \$340,000 to the gamble {\$339,600, 0.5; \$975,670, 0.5}.[7]

Rabin's results are problematic for both the realist and constructivist EU theorist: if most people have the modest-stakes preferences but lack the high-stakes preferences that 'follow' from them, then EU theory with a diminishing marginal utility function will fail to represent most people.

In case the reader is worried that Rabin's results rely on knowing a lot of the agent's preferences, there are also examples of preferences that EU theory (under either interpretation) cannot account for that involve very few preferences. One example is Allais's famous paradox. Consider Maurice, who is presented with two hypothetical choices, each between two gambles.[8] He is first asked whether he would rather have L_1 or L_2:

L_1: 10% chance of \$5,000,000, 90% chance of \$0.

L_2: 11% chance of \$1,000,000, 89% chance of \$0.

He reasons that the minimum he stands to walk away with is the same either way, and there's not much difference in his chances of winning *some* money. So, since L_1 yields much higher winnings at only slightly lower odds, he decides he would rather have L_1. He is then asked whether he would rather have L_3 or L_4:

L_3: 89% chance of \$1,000,000, 10% chance of \$5,000,000, 1% chance of \$0.

L_4: 100% chance of \$1,000,000.

He reasons that the minimum amount that he stands to win in L_4 is a great deal higher than the minimum amount he stands to win in L_3, and that although L_3 comes with the possibility of much higher winnings, this fact is not enough to offset the possibility of choosing L_3 and ending up with nothing. So he decides he would rather have L_4. Most people, like Maurice, prefer L_1 to L_2 and L_4 to L_3. However, there is no

[7] Rabin states his results in terms of changes from initial wealth levels because he hypothesizes that part of the correct explanation for people's risk aversion in modest stakes is *loss aversion* of the kind discussed in Kahneman, Daniel and Amos Tversky (1979), "Prospect Theory: An Analysis of Decision under Risk," *Econometrica* 47: 263-291.

[8] Example due to Maurice Allais (1953), "Criticisms of the postulates and axioms of the American School," reprinted in Rationality in Action: Contemporary Approaches, Paul K. Moser, ed., Cambridge University Press, 1990. Amounts of money used in the presentation of this paradox vary.

way to assign utility values to \$0, \$1m, and \$5m such that L_1 has a higher expected utility than L_2 and L_4 has a higher expected utility than L_3; therefore these preferences cannot be represented as maximizing expected utility.[9] Allais's example does not require any assumptions about an agent's psychology – it relies only on the agent having the two preferences mentioned – and so again presents a problem for both the realist and the constructivist EU theorist.

Most people have preferences like those that Allais and Rabin show cannot be captured by EU theory; and there are many other examples of preferences that EU theory cannot capture. The reason EU theory fails to capture the preferences in the Rabin and Allais examples is that it fails to separate two different sorts of reasons for risk averse preferences: local considerations about outcomes, like those that Alice advanced in order to determine that she prefers \$50 ("this particular amount of money is more valuable…") and global considerations about gambles as a whole, like those that Bob advanced in order to determine that he prefers \$50 ("I would rather be guaranteed \$50 than risk getting less for the possibility of getting more").

Why would an agent find this second kind of consideration relevant to decision making? Let us examine the idea that decision theory formalizes and precisifies means-ends rationality. We are presented with an agent who wants some particular end and can achieve that end through a particular means. Or, more precisely, with an agent who is faced with a choice among means that lead to different ends, which he values to different degrees. To figure out what to do, the agent must make a judgment about which ends he cares about, and how much: this is what the utility function captures.[10] In typical cases, none of the means available to the agent will lead with certainty some particular end, so he must also make a judgment about the likely result of each of his possible actions. This judgment is captured by the subjective probability function. Expected utility theory makes precise these two components of means-ends reasoning: how much one values various ends, and which courses of action are likely to realize these ends.

But this can't be the whole story: what we've said so far is not enough for an agent to reason to a unique decision, and so we can't have captured all that's relevant to decision making. An agent might be

[9] For if L_1 is preferred to L_2, then we have $0.1(u(\$5m)) + 0.9(u(\$0)) > 0.11(u(\$1m)) + 0.89(u(\$0))$. Equivalently, $0.1(u(\$5m)) + 0.01(u(\$0)) > 0.11(u(\$1m))$. And if L_4 is preferred to L_3, then we have $u(\$1m) > 0.89(u(\$1m)) + 0.1(u(\$5m)) + 0.01(u(\$0))$. Equivalently, $0.11(u(\$1m)) > 0.1(u(\$5m)) + 0.01(u(\$0))$. These two contradict; so there is no utility assignment that allows for the common Allais preferences.

[10] This talk may faze a certain kind of constructivist. We could recast it in terms that are acceptable to the constructivist as follows. If risk-preferences are based only on local considerations so that the agent obeys the axioms of EU theory, then the utility function as determined by EU theory will reflect these even if it doesn't correspond to anything 'real.' If risk-preferences are based on both kinds of considerations so that the agent doesn't obey the axioms of EU theory, then constructivist EU theory will read the agent as not having a utility function. However, if we can define the utility function from suitable preference axioms that these preferences do obey, then the utility function will again reflect the local considerations, as we will see in section 5.

faced with a choice between one action that guarantees that he will get something he desires somewhat and another action that might lead to something he strongly desires, but which is by no means guaranteed to do so.[11] Knowing how much he values the various ends involved is not enough to determine what the agent should do in these cases: the agent must make a judgment not only about how much he cares about *particular* ends, and how likely his actions are to realize *each* of these ends, but about what *strategy* to take towards realizing his ends *as a whole*. The agent must determine how to structure the potential realization of the various aims he has. This involves deciding whether to prioritize definitely ending up with something of some value or instead to prioritize potentially ending up with something of very high value, and by how much: specifically, he must decide the extent to which he is generally willing to accept a risk of something worse in exchange for a chance of something better. This judgment corresponds to considering global or structural properties of gambles.

How should an agent trade off the fact that one act will bring about some outcome for sure against the fact that another act has some small probability of bringing about some different outcome that he cares about more? This question won't be answered by consulting the probabilities of states or the utilities of outcomes. Two agents could attach the very same values to particular outcomes (various sums of money, say), and they could have the same beliefs about how likely various acts are to result in these outcomes. And yet, one agent might hold that his preferred strategy for achieving his general goal of getting as much money as he can involves taking a gamble that has a small chance of a very high payoff, whereas the other might hold that he can more effectively achieve *this same general goal* by taking a gamble with a high chance of a moderate payoff. Knowing they can only achieve some of their aims, these agents have two different ways to structure the potential realization of them.

This dimension of instrumental reasoning is the dimension of evaluation that standard decision theory has ignored. To be precise, it hasn't ignored it but rather supposed that there is a single correct answer for all rational agents: one ought to take actions that have higher utility on average, regardless of the spread of possibilities. There may or may not be good arguments for this, but we are not in a position to address them before we get clear on what exactly agents are doing when they answer the question differently, and how this relates to instrumental reasoning. The aim of this paper is to make this clear.

3. An Alternative Theory

To explain the alternative theory of instrumental rationality I endorse, I will start with the case of gambles with only two outcomes: gambles of the form $\{\bar{E}, x_1; E, x_2\}$. As mentioned, the EU of such a gamble is $p(\bar{E})u(x_1) + p(E)u(x_2)$. We can state this equivalently as $u(x_1) + p(E)[u(x_2) – u(x_1)]$. Taking x_2

[11] For the skeptical constructivist of the previous footnote: local considerations might point in the direction of one act, and considerations about the likelihood of realizing various ends might point in the direction of another.

to be weakly preferred to x_1, this is equivalent to taking the minimum utility value the gamble might yield, and adding to it the potential gain above the minimum – the difference between the high value and the low value – weighted by the probability of the event in which that gain is realized. For example, the value of the \$0/\$100 coin-flip will be u(\$0) + (0.5)[u(\$100) – u(\$0)].

The value of a gamble is its **instrumental value**, a measure of how the agent rates it in terms of satisfying her aims: we might say, a measure of its effectiveness as a means to her various ends. To review, on EU theory, while it is up to agents themselves how valuable each outcome is and how likely they believe each event is to obtain, these two evaluations are of set significance to the instrumental value of a gamble. If two decision makers agree about the values of various outcomes and on the probabilities involved, they must evaluate gambles in exactly the same way: they must have identical preference orderings.[12] They must agree about how the gambles rank in terms of satisfying their aims.

However, it is plausible to think that some people are more concerned with the worst-case scenario than others, again, for purely instrumental reasons: because they think that *guaranteeing* themselves something of moderate value is a better way to satisfy their general aim of getting some of the things that they value than is making something of very high value merely possible. More realistically, the minimum value won't always trump the maximum in their considerations, but it will weigh more heavily. Alternatively, an agent might be more concerned with the best-case scenario: the maximum weighs more heavily in the estimation of a gamble's value than the minimum does, even if these two outcomes are equally likely. So, it is plausible that two agents who attach the same values as each other to \$100 and \$0 will not both attach the same value to a coin-flip between \$0 and \$100. One agent will take the fact that he has a 50% chance of winning the better prize to be a weaker consideration than it is for the other. Thus, in addition to having different attitudes towards outcomes and different evaluations of likelihoods, two agents might have different attitudes towards some way of potentially attaining some of these outcomes.

A natural way to interpret these different attitudes is to postulate that different decision makers take the fact that they might improve over the minimum to be a more or less important consideration in evaluating a gamble. Formally, they weight the potential gain above the minimum differently from each other. In EU theory, the instrumental value of a gamble is at least its minimum utility value, and the potential utility gain above the minimum is weighted by the probability of attaining the higher value. But this latter feature is too restrictive: a potential gain over the minimum *might* increase a gamble's instrumental value over the minimum value by the size of the gain multiplied by the probability of realizing that gain, but it might instead improve it by more or by less, depending on what the agent cares

[12] For the constructivist: if two decision-makers share the same local considerations and agree on the probabilities involved, they must have identical preference orderings.

about. Of course, the probability and the size of the improvement will be relevant: the higher the probability of some particular gain or the larger the size of a gain with some particular probability, the better. Therefore, I propose that the possibility of a potential utility gain over the minimum improves the gamble above its minimum utility value by the size of the gain multiplied by a **function of** the probability of realizing that gain, instead of by the bare probability. This function represents the agent's attitude towards risk in the "global properties" sense. Put formally, we might calculate the **risk-weighted expected utility (REU)** of a gamble $\{\bar{E}, x_1; E, x_2\}$, where $u(x_1) \leq u(x_2)$, to be $u(x_1)$ + $\mathbf{r(}p(E)\mathbf{)}[u(x_2) - u(x_1)]$, where *r* is the agent's "risk function," adhering to the constraints $0 \leq r(p) \leq 1$ for all *p*; $r(0) = 0$; $r(1) = 1$; and *r* is non-decreasing. The equation says that the instrumental value of a two-outcome gamble will be its low value plus the interval between the low value and the high value, weighted by the output of the risk function when the input is the probability of getting the high value.

The equation is equivalent to $r(p(E))u(x_2) + (1 - r(p(E))u(x_1)$. So we can think of r(p) either as the weight a particular improvement-possibility gets when this possibility has probability *p*, or as the weight that the better outcome gets when this outcome has probability *p*. If the risk function has a high value for some value *p*, then the value of the better outcome will count for a lot in the agent's evaluation of the gamble, and if it has a low value for some value *p*, then the value of the worse outcome will count for a lot. This formulation also makes it clear how an agent's evaluation of gambles rests on factors that are irreducibly global: the amount by which each outcome gets weighted will depend on which outcome is the minimum.[13]

For example, for an agent who values money linearly and has a risk function of $r(p) = p^2$, the coin-flip will be worth \$25: u({HEADS, \$0; TAILS, \$100}) = u(\$0) + $(0.5)^2$[u(\$100) – u(\$0)] = u(\$25).

[13] If contra our supposition, $u(x_2) \leq u(x_1)$, then the value of the gamble would be $r(p(\bar{E}))u(x_1) + (1 - r(p(\bar{E}))u(x_2)$, i.e. $r(1 - p(E))u(x_1) + (1 - r(1 - p(E))u(x_2)$, which need not be equivalent to $r(p(E))u(x_2) + (1 - r(p(E))u(x_1)$.

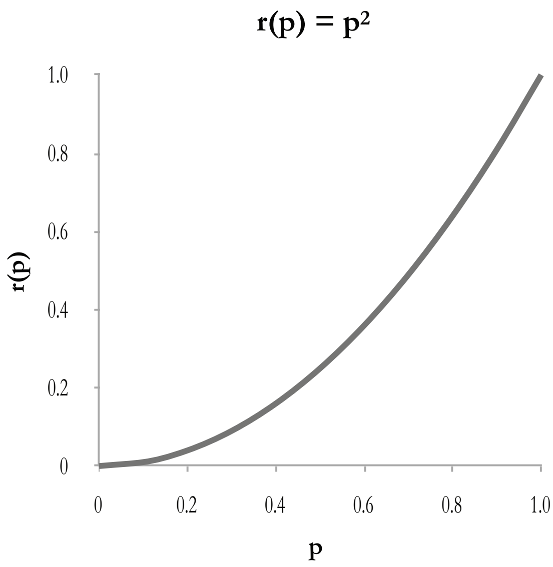

Diagram 4: Sample Risk Function: $r(p) = p^2$

Diagram 2: Sample Risk Function: $r(p) = p^2$

And here we come to the crux of the difference between how EU theory represents risk aversion and what I think instead merits the term. *On EU theory, to be risk averse is to have a concave utility function. On a theory like mine, to be risk averse is to have a convex risk function.*[14] The intuition behind the diminishing marginal utility analysis of risk aversion was that adding money to an outcome is of less value the more money the outcome already contains. The intuition behind the present analysis of risk aversion is that adding *probability* to an outcome is of more value the more likely that outcome already is to obtain. Risk averters prefer to "get to certainty," so to speak. Of course, theories like mine allow that the utility function is concave (or, indeed, any shape). But I claim that this feature, which describes how an agent evaluates outcomes, pulls apart from his attitude towards risk properly called. So I claim that what we might appropriately describe as an agent's attitude towards risk is captured by the shape of his risk function.

There is a natural way to extend this theory to gambles with more than two possible outcomes. The way I've set up the risk-weighted expected utility equation emphasizes that an agent considers his possible gain above the minimum (the interval between the low outcome and the high outcome), and weights that gain by a factor which is a function of the probability of obtaining it, a function that depends on how he regards risk. Now consider a situation in which a gamble might result in one of *more than two* possible outcomes. It seems natural that the agent should consider the possible gain between each

[14] For further discussion of this point, see Lara Buchak (2013), Risk and Rationality, Oxford University Press.

neighboring pair of outcomes and his chance of attaining the higher outcome or better, and, again, subjectively determine how much that chance of attaining this adds to the instrumental value of the gamble.

One way to state the value of a gamble with more than two outcomes for a standard EU maximizer is as follows. Start with the minimum value. Next, add the interval difference between this value and the next highest value, weighted by the probability of getting at least that higher value. Then add the interval difference between this value and the next highest value, weighted by the probability of getting at least *that* value. And so forth. Just as we replaced subjective probabilities by subjective weights of subjective probabilities in the two-outcome case, we can do so in this case. So the value of a gamble for the REU maximizer will be determined by following this same procedure but instead weighting by a function of the probability at each juncture.

For example, consider the gamble that yields \$1 with probability ½, \$2 with probability ¼, and \$4 with probability ¼. The agent will get at least \$1 for certain, and he has a ½ probability of making at least \$1 more. Furthermore, he has a ¼ probability of making at least \$2 beyond *that*. So the EU of the gamble is u(\$1) + ½[u(\$2) – u(\$1)] + ¼[u(\$4) – u(\$2)], and the REU of the gamble is u(\$1) + **r**(½)[u(\$2) – u(\$1)] + **r**(¼)[u(\$4) – u(\$2)].

So the gamble *g* = {$E_{1,}$ x_1; $E_{2,}$ x_2; … ; E_n, x_n}, where $u(x_1) \leq \ldots \leq u(x_n)$, is valued under expected utility theory as $\sum_{i=1}^{n} p(E_i)u(x_i)$, which is equivalent to:

EU(g) =

$$u(x_1) + (\sum_{i=2}^{n} p(E_i))(u(x_2) - u(x_1)) + (\sum_{i=3}^{n} p(E_i))(u(x_3) - u(x_2)) + ... + p(E_n)(u(x_n) - u(x_{n-1}))$$

And that same gamble will be valued under risk-weighted expected utility theory as follows:

REU(g) =

$$u(x_1) + r(\sum_{i=2}^{n} p(E_i))(u(x_2) - u(x_1)) + r(\sum_{i=3}^{n} p(E_i))(u(x_3) - u(x_2)) + ... + r(p(E_n))(u(x_n) - u(x_{n-1}))$$

We can now see how the standard Allais preferences are captured by REU theory: they maximize risk-weighted expected utility only if *r* is convex.[15]

This functional form is an instance of the "rank-dependent" approach in non-expected utility theory, discovered by economists around the 1980s, in which the agent maximizes a sum of utility values of outcomes, weighted by a factor that is related to the probability of that outcome but that depends on the

[15] $L_1 > L_2$ ⇔ r(.1)[u(\$5m) – u(\$0)] > r(.11)[u(\$1m) – u(\$0)].
$L_4 > L_3$ ⇔ (1 – r(.99))[u(\$1m) – u(\$0)] > r(.1)[u(\$5m) – u(\$1m)].
These inequalities hold jointly only if r(0.11) – r(0.1) < 1 – r(0.99).

outcome's rank among possible outcomes. In particular, two of these theories are formally equivalent to REU theory when we abstract away from what their "weighting factor" is a function of. The first is Choquet expected utility (CEU), due to David Schmeidler and Itzhak Gilboa,[16] and the second is anticipated utility (AU), due to John Quiggin.[17] However, CEU employs a weighting function of states, not of probabilities of states: it does not include an agent's judgments about probabilities at all. Indeed, it is meant to apply to decision making under uncertainty, in which agents do not always have sharp probability judgments.[18] AU does attach decision weights to probabilities, but it uses an "objective" probability function: it takes the probabilities as given.

What is missing from each of these theories, for the purposes of philosophers interested in instrumental rationality, is a way to separate *beliefs* from *decision weights*: each theory contains only one subjective parameter that goes into fixing the weight an outcome gets in the "expectation" equation. Therefore, neither theory can formalize what it is to have a belief separately from what it is to have a decision weight. But as I argued in my discussion of instrumental rationality, the question of which strategy to employ is entirely separate from the question of what one believes. The REU formulation allows that an agent attaches subjective probabilities to states and then employs a weighting function of these probabilities. Furthermore, the theorem presented later in the paper will allow us to derive each of these uniquely from an agent's preferences, so we can see exactly how each contributes to the evaluation of a gamble. Nonetheless, it might be helpful to the reader to think of REU theory as a generalization of AU theory to decision making with subjective probabilities: i.e., as "subjective" anticipated utility theory. Alternatively, one could think of REU theory as a restriction of CEU to the case in which agents are probabilistically sophisticated and whose decision weights are a function of their subjective probabilities: i.e., as "CEU with probabilistic sophistication."[19] This accurately represents the formalism, if not the philosophical commitments.

If we set $r(p) = p$, we get the standard subjective expected utility equation. And again, for agents who care more about the worst-case scenario – agents with convex *r*-functions – the possibility of getting more than the minimum will contribute to the value of the gamble less than it will for the expected utility

[16] David Schmeidler (1989), "Subjective Probability and Expected Utility without Additivity," *Econometrica* 57: 571-587. Itzhak Gilboa (1987), "Expected Utility with Purely Subjective Non-Additive Probabilities." *Journal of Mathematical Economics* 16: 65-88.

[17] John Quiggin (1982), "A Theory of Anticipated Utility," *Journal of Economic Behavior and Organization* 3: 323-343.

[18] Schmeidler's (1989) version includes some objective probabilities to derive the decision weights.

[19] For further discussion of the relationship of REU theory to these theories and of other non-expected utility theories present in the economics and psychology literature, see Buchak (2013). For other surveys of non-expected utility theories, see Robert Sugden (2004), "Alternatives to Expected Utility: Foundations," Chapter 14 of Handbook of Utility Theory, eds. Salvador Barberà, Peter J. Hammond, and Christian Seidl. Boston: Kluwer Academic Publishers, 685-755; and Chris Starmer (2000), "Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk," *Journal of Economic Literature* 38: 332-382.

maximizer. The most extreme case of this is the maximinimizer, who simply takes the gamble with the highest minimum. He can be represented using r(p) = {0 if p ≠ 1, 1 if p = 1}. And for agents who care more about the best-case scenario – agents with concave *r*-functions – the possibility of getting higher than the minimum will contribute to the value of the gamble more. The maximaximizer, who takes the gamble with the highest maximum, can be represented using r(p) = {0 if p = 0, 1 if p ≠ 0}. The REU equation also ensures that the value of a gamble is always at least its minimum and at most its maximum, and, since *r* is non-decreasing, that increasing the probability of getting a good outcome will never make a gamble worse (preferences respect weak stochastic dominance). If we require in addition that *r* is increasing, then increasing the probability of getting a good outcome will always make a gamble better (preferences respect strict stochastic dominance).

What ingredient of instrumental rationality does the risk function represent? The utility function is traditionally supposed to represent desire, and the probability function belief – both familiar propositional attitudes. We try to make beliefs "fit the world," and we try to make the world fit our desires. But the risk function is neither of these things: it does not quantify how we see the world – it does not, for example, measure the strength of an agent's belief that things will go well or poorly for him – and it does not describe how we would like the world to be. It is not a belief about how much risk one should tolerate, nor is it a desire for more or less risk. The risk function corresponds to neither beliefs nor desires. Instead, it measures how an agent structures the realization of his aims. We will see in the remainder of this paper exactly how it does this.

On REU theory, the agent subjectively determines *three* things: which outcomes he prefers, how likely various acts are to result in various outcomes, and the extent to which he is generally willing to accept the risk of something worse in exchange for the possibility of something better. First, like EU theory, REU theory allows agents to attach subjective values to outcomes. It is up to agents themselves to choose their ends, and hence, REU theory includes a subjective utility function, which is not necessarily linear in money (or any other good). Second, also like EU theory, REU theory allows agents to set their credences, and hence, REU theory includes a subjective probability function. Third, *unlike* EU theory, REU theory allows agents to decide how to structure the potential realization of the outcomes they care about. It allows them to judge which gamble is better from the point of view of their general aim of getting more money (which includes their particular aims of, say, getting $50 or of getting $100, which they may value twice as much). It is up to them whether they will better fulfill their general aim by prioritizing a high minimum or by prioritizing a high maximum. And it is up to them exactly how these two features of gambles trade off, e.g., how much possibly doing better than the minimum is worth, or how much weight to put on what happens in the top *p* portion of outcomes. Hence, REU theory includes a subjective risk function, which is not necessarily linear in probability.

Every agent has beliefs and desires, and determines for himself a norm for translating these two things into preferences. EU theory claims that r(p) = p is the correct norm: that we ought to be globally neutral. But just as rationality does not dictate a single utility function or credence function for all agents, I claim that it also does not dictate a unique norm.

4. From Preferences to Beliefs and Desires

I have so far been focusing on the question of how an agent might aggregate her beliefs (credence function) and desires (utility function) to arrive at a single value for an act. I've claimed that agents need not aggregate according to expected utility, but instead might weight utility intervals by a function of their probabilities. Thus we can model decision makers as using a more general decision rule, which includes a utility function, a credence function, and a risk function. However, the question that has received the most attention in philosophy is not how an agent might arrive at preferences given her beliefs and desires, but rather how we might determine an agent's beliefs and desires from her preferences. Specifically, decision theorists are interested in what restrictions on preferences will allow us to fix unique beliefs and desires. So the question that arises for REU theory is what restrictions are needed in order to extract beliefs, desires, and attitudes towards risk.

As mentioned, a utility and probability function represent an agent under EU theory just in case for all acts *f* and *g*, the agent weakly prefers *f* to *g* iff EU(g) ≤ EU(f), where expected utility is calculated relative to the agent's subjective probability function of states. A *representation theorem* for EU theory spells out a set of axioms such that if an agent's preferences obey these axioms, then she will be representable under EU theory by a unique probability function and a utility function that is unique up to positive affine transformation: she will be an EU maximizer relative to these functions.[20]

Representation theorems are important in decision theory, but their upshot depends on the use to which decision theory is put. There are at least two very different ways in which decision theory has been used, which I refer to as the *prescriptive* use and the *interpretive* use.[21]

When the theory is taken prescriptively, an agent uses it to identify the choice he should make or the preferences he should have; or the decision theorist uses the theory to assess whether the agent's

[20] See Leonard Savage (original 1954, second edition 1972), The Foundations of Statistics, New York: Dover Publications, Inc. See also Frank P. Ramsey (1926), "Truth and Probability," in Ramsey, 1931, The Foundations of Mathematics and other Logical Essays, Ch. VII, pp 156-198, edited by R.B. Braithwaite, London: Kegan, Paul, Trench, Trubner & Co., New York: Harcourt, Brace and Company. For a survey of representation theorems for EU theory, see Peter Fishburn (1981), "Subjective Expected Utility: A Review of Normative Theories," *Theory and Decision* 13. A different sort of representation theorem is due to Richard Jeffrey (1965), The Logic of Decision, McGraw Hill. For Jeffrey, the state space and the outcome space are the same, and each outcome is a gamble over other outcomes, i.e., there are no "final outcomes." As a result, his uniqueness result for the utility function is weaker.

[21] A third use is the *descriptive* use, which I won't discuss here.

choices and preferences are rational. The agent himself can use decision theory prescriptively in at least two ways. First, if he has already formed preferences over enough items, he can look to decision theory to tell him the preferences he should have over other items: to tell him how to attain his ends of getting, on balance, things that he (already) more strongly prefers. (Under realist decision theory, knowing his utility and probability values will also be enough for decision theory to tell him the preferences he should have.) Second, if an agent realizes that his preferences are not in accord with decision theory, then he can conclude that he has done something wrong and, insofar as he is rational, that he should alter his preferences so that they do so accord. In addition, the *theorist* using decision theory prescriptively ascertains whether the agent's choices in fact accord with the theory, and it is by this criterion that she judges whether the agent's preferences are rational.

Representation theorems state the conditions under which an agent can count as an EU maximizer, and thus the conditions under which an agent's preferences are rational (according to the standard theory). Therefore, they are useful for prescriptive decision theory because they provide a criterion for determining when an agent has irrational preferences that doesn't require knowing his precise numerical values. Furthermore, this criterion can be employed if we think there are no precise numerical values to know aside from those that result from the theorem, so they are especially useful to the constructivist. For the constructivist, rationality just is conformity to the axioms of decision theory, and it is a convenience that this also guarantees representability as an expected utility maximizer. Thus, representation theorems are useful because they allow us to refocus the debate about rationality: instead of arguing that a rational agent ought to maximize expected utility because, say, he ought to care only about average utility value, the EU theorist can argue that a rational agent ought to conform to the axioms.

In contrast to prescriptive decision theory, a portion of the modern philosophical literature treats decision theory *interpretively*: not as a useful guide to an agent's own decisions, but rather as a framework to interpret an agent's desires, his beliefs, and perhaps even the options that he takes himself to be deciding among. The interpretive use of decision theory arises in response to a worry about how to discover what an agent believes and desires, given that we have no direct access to these mental states – and, if constructivism (or a version of realism on which one's desires are opaque) is true, neither do the agents themselves, since these states cannot be discovered by introspection. However, it seems that it is relatively easy to discover agents' preferences: preferences do manifest themselves directly, if not perfectly, in behavior, and are ordinarily open to introspection.

It should be clear how representation theorems are useful to interpretive theorists. If an agent's preferences obey the axioms of EU theory, then the interpretive theorist can start with the agent's (observable) preferences and derive the probability function and utility function that represent her. It should also be clear why it is important that the theorems result in *unique* probability and utility functions. If there were multiple <p, u> pairs that each could represent the agent as an expected utility maximizer, we wouldn't know which way of representing the agent accurately captures "her" beliefs and desires.[22]

We can see that representation theorems are crucial to decision theory, so any alternative to EU theory needs a representation theorem if it can hope to serve the purposes EU theory is traditionally put to. Furthermore, comparing the axioms of an EU representation theorem to those of an alternative will allow us to see the difference between what each theory requires of rational agents. In the remainder of this paper, I explain the representation theorem for REU theory. This theorem presents a set of axioms such that if a decision maker's preferences obey these axioms, we can determine a unique probability function, a unique risk function, and a utility function that is unique up to positive affine transformation such that the agent maximizes REU relative to these three functions. One thing this will illuminate is what attitude the risk function corresponds to at the level of preferences, and exactly how it differs from a credence function. My main aim here is to show how REU theory captures what it is to be risk-averse, by contrasting the axioms of the REU representation theorem and the stronger axioms of the analogous representation theorem for EU theory. This will provide a way to frame the debate about whether REU maximizers are rational around the question of whether agents ought to obey the axioms of EU theory or only the weaker axioms of REU theory.

5. The Axiomatic Dispute

What is crucial to the present paper is presenting the axiomatic difference between REU theory and EU theory in a way that explains how the difference between the axioms of REU theory and the stronger axioms of EU theory amounts to the difference between allowing agents to care about global properties – or to determine for themselves the third component of instrumental rationality – and prohibiting them from doing so. Thus, I will leave the technical presentation of the axioms and theorem, as well as a more formal discussion of them, to the appendix. I will here concentrate on explaining the crucial differences between the theories.[23]

First, though, let me briefly say something about the relationship of the theorem here to other related theorems. The theorem here draws on two other results, one by Veronika Köbberling and Peter

[22] Recall that the utility function is only unique up to positive affine transformation. Therefore, only the facts that are common to all of the utility functions, e.g., the relative size of the utility intervals between outcomes, are rightly called facts about the agent's utilities.

[23] The proof of the theorem is found in Buchak (2013).

Wakker and the other by Mark Machina and David Schmeidler.[24] Köbberling and Wakker prove a representation theorem for another rank-dependent theory, CEU theory. CEU, like REU theory, applies to preferences over acts, in which outcomes are tied to events whose probabilities are not given. However, as mentioned, CEU does not represent the agent as having a function that assigns *additive probabilities* to events, and thus the representation theorem for CEU does not provide us with a way of extracting a rational agent's degrees of belief from his preferences. Machina and Schmeidler give conditions under which an agent can be represented as probabilistically sophisticated – as having a unique probability function relative to which his preferences respect stochastic dominance – and as maximizing *some* value function, but their result does not allow us to determine the values of outcomes aside from the gambles they are embedded in. Combining their results, I give conditions under which we can represent an agent as a probabilistically sophisticated decision maker maximizing the specific function that this paper is concerned with: conditions under which we can extract from an agent's preferences a probability function, a utility function, and a *function that represents how he structures the realization of his aims in the face of risk.* Thus the set of axioms in the REU representation theorem are a combination of Köbberling and Wakker's and Machina and Schmeidler's axioms, strictly stronger than either set of axioms.

The crux of the disagreement between REU theory and EU theory concerns the status of an axiom known as Unrestricted Tradeoff Consistency. According to the EU theorist, rational preferences must satisfy this condition. But according to the REU theorist, rational preferences must only satisfy two jointly weaker conditions, Comonotonic Tradeoff Consistency (from Köbberling's and Wakker's axiomatization) and Strong Comparative Probability (from Machina's and Schmeider's axiomatization).[25]

We can see the difference between these commitments by observing the pattern of preferences that Unrestricted Tradeoff Consistency (UTC) rules out but that REU theory allows. The basic idea behind UTC is that which tradeoffs you are willing to accept – tradeoffs about what happens in various states – reveals your values. For example, if it is worth it to you to make one state worse by $10 in order to make another state better by $20, that shows that the utility difference $10 makes to the first state is equivalent to the utility difference $20 makes to the second state, as long as the two states are equally likely. But things are slightly more complicated in the formal theory: since we are trying to extract both utilities and probabilities from preferences, we cannot use the terms "equally likely" or "utility

[24] Veronika Köbberling and Peter Wakker (2003), "Preference Foundations for Non-expected Utility: A Generalized and Simplified Technique," *Mathematics of Operations Research* 28, 395-423. Mark J. Machina and David Schmeidler (1992), "A More Robust Definition of Subjective Probability," *Econometrica* 60(4): 745-780.

[25] Comonotonic Tradeoff Consistency follows directly from Unrestricted Tradeoff Consistency. In the presence of the other axioms of EU or REU theory, UTC is strictly stronger than the combination of CTC and Strong Comparative Probability.

difference” in our axioms themselves. So EU theorists have to capture the spirit of this idea just using preferences.

To capture this idea, the EU theorist defines a precise notion of *equal tradeoffs*. Let us say that *x* rather than *y* in E is an appropriate tradeoff for *f* rather than *g* in Ē if an agent is indifferent between the gamble that yields *x* if event E obtains and *f* if event Ē obtains and the gamble that yields *y* if event E obtains and *g* if event Ē obtains.[26] For example, if the agent is indifferent between these two gambles:

{$10, HEADS; $150, TAILS}
{$0, HEADS; $170, TAILS}

Then $10 rather than $0 in HEADS is an appropriate tradeoff for $150 rather than $170 in TAILS. The idea is: the improvement in the HEADS state ($10 rather than $0) exactly compensates for the devaluation in the TAILS state ($150 rather than $170). Now let us assume that the agent is also indifferent between these two gambles:

{$110, HEADS; $150, TAILS}
{$100, HEADS; $170, TAILS}

Then $110 rather than $100 in HEADS is an appropriate tradeoff for $150 rather than $170 in TAILS. If $10 rather than $0 in HEADS and $110 rather than $100 in HEADS are both appropriate tradeoffs in the same event for the same pair ($150 and $170 in TAILS), then they *play the same compensatory role*. We can say that $10 rather than $0 is “tradeoff equal” to $110 rather than $100 because there is some situation in which they play the same compensatory role. The EU theorist wants tradeoff equality to imply that the value difference between the outcomes is the same: u($10) – u($0) = u($110) – u($100). And the way to capture this thought axiomatically – mentioning only preferences, rather than utilities or probabilities – is to dictate that whenever $10 rather than $0 is an appropriate tradeoff in some event for some pair, so too is $110 rather than $100 an appropriate tradeoff in that event for that pair. If $10 rather than $0 plays the same compensatory role in some event in some pair of gambles as $110 rather than $100, they play the same compensatory role in every event in every pair of gambles. This is essentially what the Unrestricted Tradeoff Consistency Axiom says.[27]

[26] In this discussion, *x* and *y* stand for outcomes and *f* and *g* for gambles (including the degenerate gamble which yields the same outcome in every state), but the difference won’t matter for understanding the discussion here.

[27] The axiom is actually framed as: if $110 rather than $100 is tradeoff equal to $10 rather than $0 (there is some event and some gamble pair in which they are both appropriate tradeoffs), and we have some outcome y’ that is preferred to $110, then $y’ rather than $100 is *not* tradeoff equal to $10 rather than $0 (there is no event and no gamble pair in which they are both appropriate tradeoffs). But the difference is not important to our informal discussion.

For example, let us assume the agent has the above preferences, and let us also assume the agent is indifferent between these gambles:

{\$10, HEADS; \$50, TAILS}
{\$0, HEADS; \$70, TAILS}

Again, \$10 rather than \$0 in HEADS is an appropriate tradeoff for \$50 rather than \$70 in TAILS. UTC implies that the agent must also be indifferent between the following two gambles, since we know from the agent's other preferences that \$10 rather than \$0 is an appropriate tradeoff for the same event in the same pair as \$110 rather than \$100:

{\$110, HEADS; \$50, TAILS}
{\$100, HEADS, \$70, TAILS}

Violating UTC is supposed to amount to not evaluating outcomes consistently: to allowing the value difference between a pair of outcomes to vary depending on the circumstances. But what UTC neglects is that an agent might evaluate outcomes consistently, but allow that an outcome makes a different contribution to the value of a gamble when that outcome occurs in a different structural part of that gamble. An agent might care about where in the structure of the gamble a tradeoff occurs. For example, we can give a rationale for an agent's having the following four preferences, which I've just said UTC rules out:[28]

{\$10, HEADS; \$150, TAILS} ~ {\$0, HEADS; \$170, TAILS}
{\$110, HEADS; \$150, TAILS} ~ {\$100, HEADS; \$170, TAILS}
{\$10, HEADS; \$50, TAILS} ~ {\$0, HEADS; \$70, TAILS}
{\$110, HEADS; \$50, TAILS} < {\$100, HEADS; \$70, TAILS}

Here is the rationale. Making the HEADS state better by \$10 might exactly compensate for making the TAILS state worse by \$20 *only when the HEADS state is the worst state*. When the TAILS state is the worst state, making the HEADS state better by \$10 might not compensate for making the TAILS state worse by \$20. As an analogy: we might think that giving \$10 to one person and taking \$20 from another preserves the overall value of a social distribution only if we are taking \$20 from the best-off person and giving the \$10 to the worst-off person, since we might think it's okay to take more from the top in order to give something to the bottom, but not vice versa.

If which tradeoffs an agent is willing to accept is sensitive to *where in the structure of the gamble these tradeoffs occur* (in the worst state or the best state, for example), then facts about which tradeoffs an agent considers appropriate will only reveal utility differences *when we are dealing with gambles that put the events in the same order*. By putting the events in the same order I mean: if event *E* is at least as good

[28] It rules them out, in the presence of the Ordering Axiom, because if the agent is indifferent in the first three pairs of gambles, UTC implies that she is indifferent in the fourth pair, but she in fact has a preference between them.

as event *F* for one gamble, then event *E* is at least as good as event *F* for another gamble. For example, the gambles in the first three pairs above are all such that the TAILS state is at least as good at the HEADS state, but this is not so in the gambles in the last pair. The technical term for gambles that order the events in the same way is *comonotonic* gambles, and a set of gambles that all order the events in a particular way is called a *comoncone*.

The Comonotonic Tradeoff Consistency Axiom, the axiom that the REU theorist accepts, dictates that tradeoffs reveal utility differences only when the gambles order events in the same way. It does this by restricting the Unrestricted Tradeoff Consistency Axiom to hold only when we are dealing with comonotonic gambles. In short: if two pairs ($10 rather than $0 and $110 rather than $100) are both appropriate tradeoffs in the same event in some pair of gambles, then the utility value difference between them must be the same *unless the tradeoffs occur in different structural parts of the gamble.* If $10 rather than $0 in HEADS and $110 rather than $100 in HEADS are both appropriate tradeoffs in the same event for the same pair, *and* they are tradeoffs in the same structural part of the gamble – if they are "comonotonic tradeoff equal" – then the utility value difference between them is the same: u($10) – u($0) = u($110) – u($100). Again, we can put this in terms of preferences without mentioning utility: if $10 rather than $0 plays the same compensatory role in some event in some pair of gambles as $110 rather than $100 plays in that event in that pair of gambles and all four of these gambles order the events in the same way as each other, then $10 rather than $0 plays the same compensatory role as $110 rather than $100 in any event and any gamble such that all four of the gambles order the events in the same way as each other. This is essentially what Comonotonic Tradeoff Consistency says (with variables instead of specific amounts of money, of course).[29] So Comonotonic Tradeoff Consistency limits the situations in which we can infer utility differences from which tradeoffs an agent is willing to accept.

In sum, Unrestricted Tradeoff Consistency says that if one pair plays the same compensatory role in some event in some gamble as another pair plays, then the two pairs must play the same compensatory role in every event in every gamble. Comonotonic Tradeoff Consistency says that this holds only when the two compensating pairs occur in the same structural part of the gamble as each other, for example, when the tradeoffs both occur in the worst outcome or both occur in the best outcome. Here is another way to put this. Unrestricted Tradeoff Consistency entails that *the utility contribution made by each outcome is separable from what happens in other states.* In other words, y-in-E rather than x-in-E makes

[29] As above, the axiom is actually framed as: if $110 rather than $100 is comonotonic tradeoff equal to $10 rather than $0 (there is some event and some gamble pair in which they are both appropriate tradeoffs, and all four gambles are comonotonic), and we have some outcome y' that is preferred to $110, then $y' rather than $100 is *not* comonotonic tradeoff equal to $10 rather than $0 (there is no event and no gamble pair in which they are both appropriate tradeoffs and all four gambles are comonotonic). But, again, the difference is not important to our informal discussion.

the same difference to the overall gamble (it exactly compensates for the same subgambles) regardless of what happens in $\bar{E}$. Furthermore, *y* rather than *x* makes the same value difference regardless of which event the substitution occurs in – not in terms of absolute value, but in terms of which other tradeoffs it is equivalent to.[30] Comonotonic Tradeoff Consistency entails that *the utility contribution made by each outcome is separable from what happens in other states if and only if we stay within a comoncone.* In other words, y-in-E rather than x-in-E makes the same difference to the overall gamble regardless of what happens in $\bar{E}$, as long as *E* occupies the same position in the "event ordering" in each relevant gamble. But still, if we remain in the same comoncone, then which event *E* is will not matter, so the value difference a tradeoff makes will be relativized to a gamble, but not to an event.

Why might it make a difference which structural part of the gamble a tradeoff occurs in? For example, why does $10 rather than $0 in HEADS play the same compensatory role as $110 rather than $100 in HEADS when HEADS is the worst outcome in both cases, but $10 rather than $0 in HEADS when HEADS is the worst outcome doesn't play the same compensatory role as $110 rather than $100 when HEADS is the best outcome? There are two possibilities. The first is that the agent considers HEADS more likely when it has a worse outcome associated with it, and less likely when it has a better outcome associated with it. He cares more about what happens in the worst possible state because the worst possible state is more likely. In this case, the agent would not have a fixed view of the likelihood of events but would instead be *pessimistic*: he would consider an event less likely simply because its obtaining would be good for him. But another axiom of REU theory, the axiom of Strong Comparative Probability, rules this interpretation out: in the presence of the other axioms, it entails that an agent has a stable probability function *p* of events. Aside from having the standard properties of a probability function, it is an important feature of *p* that it takes a higher value for one event than another just in case the agent would always rather put a better outcome on the first event than the second, holding fixed what happens in the rest of the gamble. This is reason to take the function *p* rather than the function *r* to reflect an agent's beliefs.

The second possibility is that what happens in *E* matters less to the agent when *E* is higher in the structural ordering not because she considers *E* itself less likely, but because this feature of the overall gamble plays a smaller role in the agent's considerations. If an agent is more concerned with guaranteeing himself a higher minimum, for example, then tradeoffs that raise the minimum are going to

[30] To clarify: substituting *y* rather than *x* into a gamble will make a different value difference depending on the event the substitution occurs in, merely because the more probable the event, the bigger value difference it will make; however, if substituting *y* rather than *x* for some event in some gamble makes the same difference as substituting *w* rather than *z* for that same event in that same gamble, then for *any* event and any gamble, substituting *y* rather than *x* in that event in that gamble makes the same value differences as substituting *w* rather than *z* in that event and that gamble.

matter more than tradeoffs that raise the maximum. I stressed that one thing an agent must determine in instrumental reasoning is the extent to which he is willing to trade off a guarantee of realizing some minimum value against the possibility of getting something of much higher value: that is, the extent to which he is willing to trade raising the minimum against raising the maximum. And, again, this is because agents must determine for themselves how to structure their goals.

So we can now see that restricting Tradeoff Consistency to gambles with the same structural properties – gambles that order the events in the same way – captures the idea that agents who are risk-averse in the sense of caring about global or structural properties are structuring their goals differently than EU maximizers. Unrestricted Tradeoff Consistency says that substituting one outcome for another must make the same value difference to a gamble regardless of how these outcomes feature into the structural properties of the gamble. But Comonotonic Tradeoff Consistency says that the difference a substitution makes depends not just on the difference in value between the outcomes in the particular state, but on where in the structure of the gamble this substitution occurs. If the agent cares about these structural properties then he will only obey the comonotonic version of the axiom. Furthermore, if he has stable beliefs about the state of the world, beliefs that don't depend on how good various events are for him, then he will obey the comonotonic version of the axiom not because he is pessimistic but because he structures his goals so as to place more importance on what happens in the worst possible state.

6. Conclusion

I have proposed a theory on which agents subjectively determine the three elements of instrumental rationality: their utilities, their credences, and the tradeoffs they are willing to make in the face of risk. In this paper I have explained how allowing agents to subjectively determine which sorts of tradeoffs they are willing to make corresponds to adopting a weaker set of axioms on preferences than those endorsed by the EU theorist. On EU theory, which tradeoffs an agent is willing to make must be determined solely by the outcomes and events those tradeoffs involve. This means that lowering the value of what happens in an event has the same effect on the value of a gamble regardless of what happens in the rest of the gamble. However, on REU theory, agents can care about where in the structure of the gamble the tradeoffs occur. Therefore, the effect on the value of the gamble can depend on whether it is the value of the minimum or maximum that is lowered. Furthermore, even if the agent assigns the same probability to two events *E* and *F*, she needn't think that lowering the value of *E* in exchange for raising the value of *F* (by the same utility) is an acceptable tradeoff. In particular, if the worst-case scenario is proportionately more important to her than the best-case scenario, this may not be an acceptable tradeoff when the prize in *E* is already worse than the prize in *F*. What *r* represents, then, is the extent to which an agent privileges what happens in various structural parts of the gamble: whether she is prudent in making sure the minimum value is high, or venturesome in making sure the maximum value is high.

Now that we've seen the difference between what EU theory requires of agents and what the more permissive REU theory requires of them, we can properly address the question of which theory captures the requirements of instrumental rationality. Since we can see what decision-makers who are supposedly irrational are doing in terms of taking the means to their ends, the burden will be on the defender of EU theory to show why individuals ought to adopt a very particular strategy for attaining their goals: averaging the utility values without regard to the spread of possibilities, or ignoring global considerations when deciding which tradeoffs to make. I contend that EU theory will not be able to meet this burden, and that it is rational to be sensitive to global properties of gambles in the way I suggest here. But that is a discussion for another time.

APPENDIX: REPRESENTATION THEOREM AND TECHINAL DISCUSSION

The theorem here draws on two other results, one by Veronika Köbberling and Peter Wakker, and the other by Mark Machina and David Schmeidler.[31] The set of axioms I use in the REU representation theorem are a combination of Köbberling and Wakker's and Machina and Schmeidler's axioms, strictly stronger than either set of axioms.

I start by explaining the spaces and relations we are dealing with.[32] The **state space** is a set of states SS = {…, s, …}, whose subsets are called events. The **event space**, EE, is the set of all subsets of SS. Since we want to represent agents who have preferences over not just monetary outcomes but discrete goods and, indeed, over outcomes described to include every feature of the world that the agent cares about, it is important that the outcome space be general. Thus, the **outcome space** is a set of outcomes XX = {…, x, …}. I follow Savage (1954/1972) in defining the entities an agent has preferences over as "acts" that yield a known outcome in each state. The **act space** AA = {…, f(.), g(.), …} is thus the set of all finite-valued functions from SS to XX, where the inverse of each outcome $f^{-1}(x)$ is the set of states that yields that outcome: $f^{-1}(x) \in$ EE. So for any act f$\in$AA, there is some partition of the state space SS into {E_1, … E_n} and some finite set of outcomes Y $\subseteq$ XX such that *f* can be thought of as a

[31] Köbberling and Wakker (2003). Machina and Schmeidler (1992). See Buchak (2013) for a discussion of related results. Particularly noteworthy is a similar theorem due to Nakamura in: Yutaka (1995). "Probabilistically sophisticated rank dependent utility." *Economic Letters* 48: 441-447.

[32] In the denotation of the spaces, I follow Machina and Schmeidler (1992).

member of Y^n. And as long as f(s) is the same for all $s \in E_i$, we can write $f(E_i)$ as shorthand for "f(s) such that $s \in E_i$."

For any fixed finite partition of events $M = \{E_1, \ldots, E_n\}$, all the acts on those events will form a subset $A_M \subseteq AA$. Thus, A_M is defined to contain all the acts that yield for each event in the partition, the same act for all states in that event: $A_M = \{f \in AA \mid (\forall E_i \in M)(\exists x \in XX)(\forall s \in E_i)(f(s) = x)\}$. An upshot is that for all acts in A_M, we can determine the outcome of the act by knowing which event in M obtains: we needn't know the state of the world in a more fine-grained way.

The **preference relation** $\geq$ is a two-place relation over the act space. This gives rise to the indifference relation and the strict preference relation: $f \sim g$ iff $f \geq g$ and $g \geq f$; and $f > g$ if $f \geq g$ and $\neg(g \geq f)$.

For all $x \in XX$, $\underline{x}$ denotes the constant act $\{f(s) = x$ for all $s \in SS\}$. The relata of the preference relation must be acts, but it will be useful to talk about preferences between outcomes. Thus, we will define an auxiliary preference relation over outcomes:

$x \geq y$ iff $\underline{x} \geq \underline{y}$ (for $x, y \in XX$)

where indifference and strict preferences are defined as above. It will be useful to talk about preferences between outcomes of particular acts, so, following the above definition, $f(s) \geq g(s)$ holds iff $\underline{f(s)} \geq \underline{g(s)}$, the constant act that yields f(s) in every state is weakly preferred to the constant act that yields g(s) in every state. Furthermore, $x_E f$ denotes the act that agrees with *f* on all states not contained in *E*, and yields *x* on any state contained in *E*: $x_E f(s) = \{x$ if $s \in E$; f(s) if $s \notin E\}$. Likewise, for disjoint E_1 and E_2 in EE, $x_{E1}y_{E2}f$ is the act that agrees with *f* on all states not contained in E_1 and E_2, and yields *x* on E_1 and *y* on E_2: $x_{E1}y_{E2}f(s) = \{x$ if $s \in E_1$, y if $s \in E_2$, f(s) if $s \notin E_1 \cup E_2\}$. Similarly, $g_E f$ is the act that agrees with *g* on all states contained in *E* and agrees with *f* on all states not contained in *E*: $g_E f(s) = \{g(s)$ if $s \in E$; f(s) if $s \notin E\}$. We say that an event E is **null** on $F \subseteq AA$ just in case the agent is indifferent between any pair of acts which differ only on E: $g_E f \sim f$ for all $g_E f$, $f \in F$.[33]

The concepts in this paragraph and the next are important in Köbberling and Wakker's result. The first, comonotonicity, was introduced by Schmeidler (1989). Two acts f and g are **comonotonic** if there are no states $s_1, s_2 \in SS$ such that $f(s_1) > f(s_2)$ and $g(s_1) < g(s_2)$. This is equivalent to the claim that for any partition A_M of acts such that $f, g \in A_M$, there are no events $E_1, E_2 \in M$ such that $f(E_1) > f(E_2)$ and $g(E_1) < g(E_2)$. The acts f and g order the states (and, consequently, the events) in the same way: if s_1 leads to a strictly preferred outcome to that of s_2 for act f, then s_1 leads to a weakly preferred outcome to that of s_2 for act g. We say that a subset C of some A_M is a **comoncone** if all the acts in C order the events in the same way: for example, the set of all acts on coin-flips in which the heads outcome is as good as or better

[33] Machina and Schmeidler (1992: 749).

than the tails outcome forms a comoncone. Formally, as Köbberling and Wakker define it, take any fixed partition of events $M = \{E_1, \ldots, E_n\}$. A permutation ρ from $\{1, \ldots, n\}$ to $\{1, \ldots, n\}$ is a *rank-ordering* permutation of f if $f(E_{\rho(1)}) \geq \ldots \geq f(E_{\rho(n)})$. So a comoncone is a subset C_ρ of A_M that is rank-ordered by a given permutation: $C_\rho = \{f \in A_M \mid f(E_{\rho(1)}) \geq \ldots \geq f(E_{\rho(n)})\}$ for some ρ. For each fixed partition of events of size n, there are n! comoncones.[34]

Here is an example to illustrate the idea of a comoncone. Consider the following gambles:

f = {HEADS, $50; TAILS, $0}	g = {HEADS, $100; TAILS, $99}
h = {HEADS, $0; TAILS, $50}	j = {HEADS or TAILS, $70}

The set [f, g, j] forms a comoncone, because for each gamble in the set, the heads outcome is weakly preferred to the tails outcome. The set [h, j] forms a comoncone, because for each gamble in the set, the tails outcome is weakly preferred to the heads outcome.

We say that outcomes x^1, x^2, … form a **standard sequence** on $F \subseteq AA$ if there exist an act $f \in F$, events $E_i \neq E_j$ that are non-null on F, and outcomes y, z with $\neg(y \sim z)$ such that for all k, $(x^{k+1})_{Ei}(y)_{Ej}f \sim (x^k)_{Ei}(z)_{Ej}f$, with all acts $(x^k)_{Ei}(y)_{Ej}f$, $(x^k)_{Ei}(z)_{Ej}f \in F$.[35] The intended interpretation is that the set of outcomes x^1, x^2, x^3, …, will be "equally spaced." Since the agent is indifferent for each pair of gambles, and since each pair of gambles differs only in that the "left-hand" gamble offers y rather than z if E_j obtains, and offers x^{k+1} rather than x^k if E_i obtains, the latter tradeoff must exactly make up for the former. And since the possibility of x^{k+1} rather than x^k (if E_i) makes up for y rather than z (if E_j) for each k, the difference between each x^{k+1} and x^k must be constant. Note that a standard sequence can be increasing or decreasing, and will be increasing if $z > y$ and decreasing if $y > z$. A standard sequence is bounded if there exist outcomes v and w such that $\forall i(v \geq x^i \geq w)$.

We are now in a position to define a relation that is important for Köbberling and Wakker's result and that also makes use of the idea that one tradeoff exactly makes up for another. For each partition M, we define the relation $\sim^*(F)$ for $F \subseteq A_M$ and outcomes x, y, z, $w \in XX$ as follows:

xy $\sim^*(F)$ zw

iff $\exists f,g \in F$ and $\exists E \in EE$ that is non-null on F such that $x_E f \sim y_E g$ and $z_E f \sim w_E g$,

where all four acts are contained in F.[36] Köbberling and Wakker explain the relation $\sim^*(F)$ as follows: "The interpretation is that receiving *x* instead of *y* apparently does the same as receiving *z* instead of *w*;

[34] Köbberling and Wakker (2003: 400). On p. 403, Köbberling and Wakker (2003) point out that we can also define a comoncone on an infinite state space, although this is not necessary for our purposes.

[35] Köbberling and Wakker (2003: 398). The concept of a standard sequence, however, does not originate with them.

[36] Köbberling and Wakker (2003: 396-7).

i.e. it exactly offsets the receipt of the [f's] instead of the [g's] contingent on [$\bar{E}$]."[37] The idea here is that if one gamble offers *f* if $\bar{E}$ obtains, whereas another gamble offers *g* if $\bar{E}$ obtains, then this is a point in favor of (let's say) the first gamble. So in order for an agent to be indifferent between the two gambles, there has to be some compensating point in favor of the second gamble: it has to offer a better outcome if E obtains. And it has to offer an outcome that is better by the right amount to exactly offset this point. Now let's assume that offering *y* rather than *x* (on E), and offering *w* rather than *z* (on E) both have this feature: they both exactly offset the fact that a gamble offers *f* rather than *g* (on $\bar{E}$). That is, if one gamble offers *f* if $\bar{E}$, and a second gamble offers *g* if $\bar{E}$, then this positive feature of the first gamble would be exactly offset if the first offered *x* if E and the second offered *y* if E – and it would be exactly offset if instead the first offered *z* if E and the second offered *w* if E. If this is the case, then there is some important relationship between *x* and *y* on the one hand and *z* and *w* on the other: there is a situation in which having the first member of each pair rather than the second both play **the same compensatory role**. This relationship ~* is called **tradeoff equality**. We write xy ~*(C) zw if there exists a comoncone $F \subseteq A_M$ such that xy ~*(F) zw: that is, if *x* and *y* play the same compensatory role as *z* and *w* in some gambles *f* and *g* where all of the modified gambles after *x, y, z,* and *w* have been substituted in are in the same comoncone.

The relation ~*(F), and particularly ~*(C), features centrally in the representation theorem, because one important axiom places restrictions on when it can hold: on when pairs of outcomes can play the same compensatory role. This relation also plays a crucial role in determining the cardinal utility difference between outcomes using ordinal preferences. What we are interested in is the utility contribution each outcome makes to each gamble it is part of: this will help us determine the utility values of outcomes. More precisely, since utility *differences* are what matter, we are interested in the utility contribution that *x* rather than *y* makes to each gamble. And tradeoff equality gives us a way to begin to determine this: if getting *y* rather than *x* in event *E* and getting *z* rather than *w* in event *E* both exactly compensate for getting *f* rather than *g* in event $\bar{E}$, then *y* rather than *x* and *z* rather than *w* make the same difference in utility contribution in event *E* in those gamble pairs. In order to get from these differences in utility contributions to utility full stop, we need to fix when it is that two pairs making the same difference in utility contribution means that they have the same difference in utility. And to do this, we identify the conditions under which if two pairs have the same difference in utility (full stop), they must make the same difference in utility contribution; and we constrain the rational agent to treat a pair

[37] Köbberling and Wakker (2003: 397). Note the similarity to the four-place relation "=" in Ramsey's (1926) axiomatization of EU theory. In Ramsey's axiomatization, "xy = zw" holds when the agent is indifferent between {E, x; $\bar{E}$, w} and {E, y; $\bar{E}$, z} for any "ethically neutral" proposition *E* believed to degree 0.5.

consistently in these situations – to consistently make tradeoffs. Tradeoff consistency axioms provide such a constraint.

With the preliminaries out of the way, I can now present the axioms of REU theory, side-by-side with those of the analogous representation theorem for EU theory that Köbberling and Wakker spell out.[38]

[38] Mostly following Köbberling and Wakker (2003). I alter their axioms slightly, to make the comparison clear. See Appendix B of Buchak (2013).

EXPECTED UTILITY THEORY

A1. Ordering: $\geq$ is complete, reflexive, and transitive.

A2. State-wise dominance: If $f(s) \geq g(s)$ for all $s \in SS$, then $f \geq g$. If $f(s) \geq g(s)$ for all $s \in SS$ and $f(s) > g(s)$ for all $s \in E \subseteq SS$, where E is non-null on AA, then $f > g$.

A3. Preference Richness:
(i) There exist outcomes x and y such that $x > y$.
(ii) For any fixed partition of events $E_1, \ldots, E_n$, and for all acts $f(E_1, \ldots, E_n)$, $g(E_1, \ldots, E_n)$ on those events, outcomes x, y, and events E_i with $x_{Ei}f > g > y_{Ei}f$, there exists an "intermediate" outcome z such that $z_{Ei}f \sim g$.

A4. Small Event Continuity:
For all acts $f > g$ and any outcome x, there exists a finite partition of events $\{E_1, \ldots, E_n\}$ such that for all i, $f > x_{Ei}g$ and $x_{Ei}f > g$

B5. Archimedean Axiom: Every bounded standard sequence on AA is finite.

B6. Unrestricted Tradeoff Consistency: For all $A_M \subseteq AA$, improving an outcome in any $\sim^*(A_M)$ relationship breaks that relationship. In other words, $xy \sim^*(A_M)\ zw$ and $y' > y$ entails $\neg(xy' \sim^*(A_M)\ zw)$.

RISK-WEIGHTED EXPECTED UTILITY

A1. Ordering: $\geq$ is complete, reflexive, and transitive.

A2. State-wise dominance: If $f(s) \geq g(s)$ for all $s \in SS$, then $f \geq g$. If $f(s) \geq g(s)$ for all $s \in SS$ and $f(s) > g(s)$ for all $s \in E \subseteq SS$, where E is non-null on AA, then $f > g$.

A3. Preference Richness:
(i) There exist outcomes x and y such that $x > y$.
(ii) For any fixed partition of events $E_1, \ldots, E_n$, and for all acts $f(E_1, \ldots, E_n)$, $g(E_1, \ldots, E_n)$ on those events, outcomes x, y, and events E_i with $x_{Ei}f > g > y_{Ei}f$, there exists an "intermediate" outcome z such that $z_{Ei}f \sim g$.

A4. Small Event Continuity:
For all acts $f > g$ and any outcome x, there exists a finite partition of events $\{E_1, \ldots, E_n\}$ such that for all i, $f > x_{Ei}g$ and $x_{Ei}f > g$

A5. Comonotonic Archimedean Axiom:
For each comoncone $F \subseteq A_M \subseteq AA$, every bounded standard sequence on F is finite.

A6. Comonotonic Tradeoff Consistency:
Improving an outcome in any $\sim^*(C)$ relationship breaks that relationship. In other words, $xy \sim^*(C)\ zw$ and $y' > y$ entails $\neg(xy' \sim^*(C)\ zw)$.

A7. Strong Comparative Probability: For all pairs of disjoint events E_1 and E_2, all outcomes $x' > x$ and $y' > y$, and all acts $g,h \in AA$,
$x'_{E1}x_{E2}g \geq x_{E1}x'_{E2}g \quad \Rightarrow \quad y'_{E1}y_{E2}h \geq y_{E1}y'_{E2}$

Any agent whose preferences obey the axioms in the left-hand column maximizes expected utility relative to a unique probability function and a utility function unique up to positive affine transformation. Furthermore, in the presence of (A3), any agent who maximizes expected utility will satisfy the remaining axioms.

Analogously, if a preference relation ≥ on AA satisfies (A1) through (A7), then there exist (i) a unique finitely additive, non-atomic probability function p: EE → [0, 1]; (ii) a unique risk function r: [0, 1] → [0, 1]; and (iii) a utility function unique up to positive affine transformation such that REU represents the preference relation ≥. If there are three such functions so that REU(f) represents the preference relation, we say that REU holds. Thus, if ≥ satisfies (A1) through (A7), then REU holds. Furthermore, in the presence of (A3), if REU holds with a continuous r-function, then the remaining axioms are satisfied.

Therefore, if we assume preference richness (A3), we have:

(A1), (A2), (A4), (A5), (A6), (A7) are sufficient conditions for REU.
(A1), (A2), (A4), (A5), (A6), (A7) are necessary conditions for REU with continuous r-function.

The proof of this theorem, with references to details found in Köbberling and Wakker and in Machina and Schmeidler, can be found in Buchak (2013).