Risk Aversion and Rationality

Lara Buchak, July 2009

0 Introduction

Ralph has the opportunity to participate in two gambles. In the first, a referee will flip a coin, and if it lands heads, Ralph will receive a towel that Elvis once wiped his face on (Ralph is a big Elvis fan). In the second, the referee will flip a different coin, and if it lands tails, Ralph will receive a pair of gloves. Ralph believes both coins to be fair. Now a trickster comes along, and offers Ralph a sort of insurance: for a few cents, she will rig the game so that the first coin determines both outcomes – if it lands heads, he gets the Elvis towel, and if it lands tails, he gets the gloves – and therefore that Ralph is guaranteed to receive some prize. Ralph values the two goods *independently* in the sense that having one does not add to or decrease from the value of having the other; they are not like, say, a left-hand glove and a right-hand glove. Receiving a prize does not have any value apart from the value of the prizes themselves: Ralph does not like winning for its own sake. And he is not the sort of person who experiences regret or disappointment when he might have gotten something but didn't; he only cares about what he actually has. He decides that the trickster's deal is worthwhile – it would be nice to guarantee that he gets something no matter what – so he decides to pay a few cents to rig the game. We can represent his options schematically as follows:

	HH	HT	TH	TT
Deal 1	*Elvis towel*	*Elvis towel and gloves*	*Nothing*	*Gloves*
Deal 2	*Elvis towel*	*Elvis towel*	*Gloves*	*Gloves*

Ralph prefers deal 2 to deal 1. This seems very reasonable. Many of us would have a similar preference. And yet, standard decision theory rules this out: it cannot represent his preferences, and therefore judges Ralph to be irrational.[1]

Margaret is very clear on how she values small amounts of money. Receiving \$50 is just the same to her whether she starts with \$0 or \$50, and she feels similarly about all small increments of money. We might say that she values money *linearly*: every dollar that she receives is worth as much to her as the previous ones, at least for amounts of money less than,

[1] He is irrational if he has any preference (besides indifference) between these two deals. Note that this is a general schema for a counterexample to the theory, so if the reader is not motivated by the example with these goods, he can substitute in some other goods. Of course, the above qualifications – that the goods are independent and that the agent is not the sort who experiences regret or likes winning for its own sake – are essential. Sections of this paper will be devoted to showing that these assumptions are plausible. Note also that in saying that the goods are independent for Ralph, I am not assuming that utilities can be defined apart from an agent's preferences; I will discuss this in detail later in the paper.

say, $200. And Margaret is equally clear on how she values bets: she prefers $50 to a coin flip between $0 and $100. If she takes the former, she will certainly get $50, and the possibility of getting $100 is not enough to make up for the (equally likely) possibility of getting $0 – she would rather guarantee herself $50 than take that chance. Again, these preferences seem appealing to many people, and are at least understandable. But standard decision theory cannot represent Margaret's preferences, and judges her – like Ralph – to be irrational.

Finally, in a classic example due to Maurice Allais, commonly known as the Allais paradox, people are presented with a choice between L_1 and L_2 and a choice between L_3 and L_4, where the gambles are as follows:

L_1: $5,000,000 with probability 0.1, $0 otherwise.
L_2: $1,000,000 with probability 0.11, $0 otherwise.
L_3: $1,000,000 with probability 0.89, $5,000,000 with probability 0.1, $0 otherwise.
L_4: $1,000,000 with probability 1.

People tend to choose L_1 over L_2, and L_4 over L_3: in the first pair, the minimum amount that one stands to walk away with is the same for either gamble, and there is not much difference in one's chances of winning *some* money – but L_1 yields higher winnings; in the second pair, however, the minimum amount that L_4 yields is a great deal higher than the minimum that L_3 yields.[2] Again, these preferences are understandable (most people express them), but standard decision theory cannot accommodate them, and, again, must judge them to be irrational.

In this paper, I defend the rationality of certain preferences – like Ralph's, Margaret's, and the Allais choosers' – that decision theory treats as irrational, and I offer a more permissive theory of rational decision making. What the preferences I defend have in common is that they all stem from the decision maker's attitude towards *risk*. The standard theory (expected utility theory) rules out caring about how the possible outcomes in a gamble relate to each other: for example, about the best or worst prize an agent might receive, about the difference in value between these two, about the variance among outcomes, or about the proportion of outcomes reaching a certain threshold. I show that we get an interesting and more intuitive version of utility theory if we relax the assumption that rules out these preferences, and I show that three classic arguments against these preferences fail: contra the standard theory, these preferences are in fact rational.

The first section is devoted to explicating the standard theory and the constraints it places on how agents treat risk. I show that the decision-theoretic notion of risk aversion is not the commonsense one, and that this leads to decision theory's failure to capture some intuitive

[2] Allais (1953), pg. 132. Amounts of money used in the presentation of this paradox vary.

preferences – though I do not yet argue that these preferences are rational. Along the way, I will flesh out the assumptions in the examples mentioned. I then formalize what I take to be the more intuitive way to think about risk. In my preferred theory, possible outcomes combine to yield the value of a gamble in a way that is sensitive not just to what happens in each state, but also to "global" properties of gambles, e.g. the variance among the states, that contribute to what is more naturally called the riskiness of a gamble.

The main difference between my theory and the standard theory is that while the standard theory has two parameters that determine an agent's preferences – a subjective utility function and a subjective probability function – mine adds an additional parameter: a subjective risk function. Furthermore, my theory lacks the axiom known as the sure-thing principle (hereafter, STP). The remaining three sections are devoted to exploring whether STP, or expected utility theory more generally, is a requirement on rational preferences; if it is, then it will be irrational for agents to care about risk in the way I describe. So my theory may be of technical and psychological interest, but it will be of no interest to agents trying to determine what they should do, or to theorists trying to determine which sets of preferences are rational. I consider three arguments purporting to show that agents are rationally required to obey STP and maximize expected utility:[3] that STP follows from an undeniable intuition, that agents who violate the theory do worse over the long run, and that it is conceptually impossible for rational agents to violate STP because the (rational) reasons for their preferences render purported violations not violations at all. I claim that none of these arguments succeeds. I conclude that there are attitudes towards risk that give rise to preferences that are not capturable by standard decision theory, but are nonetheless rational.

First, a note about terminology: I will use the term *option* to refer to the things amongst which the decision maker must choose. I will use the term *outcome* to refer to any of the final results of an option. *States* are the various contingencies which might obtain. A *gamble* is a function from states to outcomes; that is, a gamble specifies which outcome obtains in each possible state of the world – it might be the same outcome in every state, or different outcomes in some states than others. As an example, consider an agent deciding between two options: not bringing an umbrella and bringing an umbrella. The relevant states are "it rains" and "it does not rain" and the outcomes are (O_1) not carrying an umbrella and getting wet, (O_2) not carrying an umbrella and not getting wet, and (O_3) carrying an umbrella and not getting wet. The option "no

[3] Obeying STP and maximizing expected utility are equivalent in the presence of the other axioms of expected utility theory, which I take to be non-controversial for the purposes of this paper. The first and third argument are specifically a defense of STP, and the second of EU theory in general (that is, of the conjunction of all the axioms).

umbrella" can be thought of as the gamble that yields O_1 if it rains and O_2 if it does not rain, and the option "umbrella" as the 'gamble' that yields O_3 either way. We can represent his decision problem in a chart:

	Rain	No rain
No umbrella	Not carry, wet	Not carry, dry
Umbrella	Carry, dry	Carry, dry

In general, we have schematically:

	State	State	State	State
Option	*Outcome*	*Outcome*	*Outcome*	*Outcome*
Option	*G*	*A M*	*B L*	*E*

The distinction between states and outcomes does not matter for our purposes. However, in making a distinction between gambles and outcomes, I am following Savage rather than Jeffrey.[4] I leave it open what outcomes are – I'll sometimes speak of receiving a prize, and sometimes of making a proposition true (e.g. the proposition that Ralph gets a pair of gloves). It should be obvious when I'm using each, and it makes no difference to my argument.

1. Risk, intuitively

In expected utility (EU) theory, the *structure* of people's preferences is fixed: as long as a decision maker's preferences obey certain basic axioms,[5] there is some quantity (utility) such that an agent prefers to maximize the expectation of that quantity. The values of all the possible outcomes, and therefore the utility function, are determined by plugging an agent's ordinal preferences into that structure.[6] On this picture, utility is not meant to be a measure of goodness, or of any quantity out there in the world; rather, it is just whatever quantity plays the correct role in the theory: the quantity whose mathematical expectation an agent maximizes. Thus, properly understood, we do not have intuitions about (cardinal) utility values. We say that a utility function *represents* an agent's preferences (under expected utility theory) if the following holds: whenever the agent prefers X to Y, the function assigns a higher expected utility to X than to Y,

[4] Savage (1954), pp 13-17. I use the term *gambles*, where as Savage uses the term *acts*; the differences are not important. Jeffrey (1965), pp 145-150. It might be that the formalism of my theory can ultimately do without specifying beforehand what counts as an outcome and what counts as a gamble, but I certainly need the distinction between outcomes and gambles for the arguments of this paper to be tractable. I take it the distinction is at least intuitive.

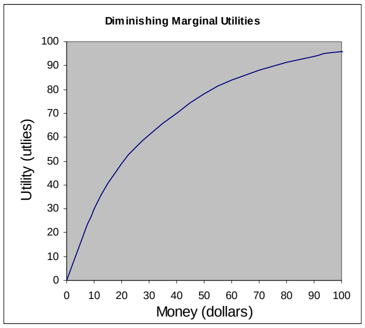
[5] Including, among others, completeness, transitivity and STP.

[6] That this is possible for agents who meet the axioms is shown by various representation theorems. E.g. Savage (1954).

and vice versa.[7] Actually, this needn't be strictly true: if an agent occasionally deviates from what the function entails because of an inability to make precise calculations, or because of a brief lapse in judgment, then we can say that the function represents him because it represents his ideal preferences (or approximates his actual preferences) – but there cannot be *widespread* deviation that an agent would *endorse* even at his most clearheaded. If an agent (or an idealized version of that agent) is not representable by a utility function then he is considered irrational – or incoherent.

For agents who are representable, any variation in preferences over gambles among these agents is going to show up as variation in utility values of the outcomes,[8] since the value of a gamble is simply the mathematical expectation of its outcomes. And the fact that some agent does not like to take risks – and has preferences that reflect this – is going to show up as her having a particular utility function; again, it is going to show up in the values her utility function assigns to *outcomes*.

For example, many people are risk averse in the following sense: they would rather have $50 than a coin flip between $0 and $100, and, in general, they would prefer to receive $*z* rather than to receive a gamble between $*x* and $*y* that will yield $*z* on average.[9] And when these preferences are plugged into the structure of EU theory, we get that utility is a function of monetary value, but one that diminishes marginally: as people think about the prospect of having more money, additional amounts of money add less value. If an agent is going to get $50, then getting an extra $50 doesn't increase the value of the prospect as much as it does when he is initially not going to get anything. Thus, preferences that are risk averse with respect to money imply, on the standard theory, a concave utility function.



[7] And similarly for indifference: the function assigns the same expected utility to X and Y iff the agent is indifferent between X and Y.

[8] We assume for this discussion that the probabilities of the outcomes in the gambles are fixed and known to the agents.

[9] For the purposes of this paper, 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.

And vice versa: when a utility function is concave, a risky gamble (say, the coin flip between $0 and $100) will always have an expected utility that is less than the utility of its expected dollar value ($50). This is because the difference between the utility of $0 and the utility of $50 is larger than the difference between the utility of $50 and the utility of $100, so the value of the higher outcome does not entirely make up for the value of the lower outcome: the higher outcome is not better than $50 by as much as the lower outcome is worse than $50.[10]

On the standard theory, then, aversion to risk is equivalent to diminishing marginal utility. Intuitively, however, these two phenomena are quite different. Consider the following example:[11] as an Elvis aficionado, I place a certain value on a towel that Elvis once wiped his face on. But I would not value 100 towels that Elvis wiped his face on 100 times more than one towel; once I have one towel, getting more towels is of little value to me. It is clear that Elvis towels have diminishing marginal value for me. And, according to the standard theory – on which risk aversion and diminishing marginal utility are the same phenomenon – I am therefore risk averse with respect to Elvis towels. But this does not seem like an accurate description of my preferences: it's not that I have any particular attitude towards *risk*, it's that I only want one Elvis towel.[12]

Furthermore, it seems that an agent can care about risk without a good having diminishing marginal value for her. In the example above, Margaret claimed that she values small amounts of money linearly: $100 really is worth twice as much to her as $50.[13] As evidence of this, Margaret would be willing to do an unpleasant task for exactly $50, she would be willing to do an unrelated unpleasant task for exactly $50, and she would be willing to do them both for $100; that is, whether she is willing to do one for $50 does not depend on whether she

[10] To keep things simple, I am talking about (fair) coin flips, where an agent has the same chance of getting each outcome. But my remarks are intended to be general: the value of the higher outcome (weighted by the probability of getting it) does not make up for the value of the lower outcome (weighted by the probability of getting it). That is, the higher value does not raise the mean by as much as the lower value lowers the mean.

[11] This type of example was suggested to me by Adam Elga, though it is not his own. Unfortunately, its original source is unknown to me.

[12] The Elvis aficionado would have preferences that display risk aversion in the following sense: he would rather have 1 towel than a coin flip between 0 towels and 2 towels. For now, however, I am just talking about intuitive explanations.

[13] This assumes that $0 – or more accurately, the status quo – is worth nothing to her. When I say that X is worth twice as much as Y, this is shorthand for the claim that the value of X above the status quo is twice the value of Y above the status quo. This shorthand makes no difference to the discussion, so I will use it for ease of exposition. Furthermore, throughout this paper, I will make the harmless assumption that the status quo has utility 0.

agrees to do the other.[14] But she would still rather have $50 than a coin flip between $0 and $100, because she prefers a sure thing to a risky bet.

The intuitive reason that the gamble is worse than the sure thing for Margaret is not that an increase from $0 to $50 is worth more than an increase from $50 to $100; rather, it is that when Margaret evaluates the attractiveness of a gamble, she does not just care about the final dollar amounts she might receive, in isolation – she also cares about how these amounts are arranged across the possibility space, since she does not yet know which possibility will obtain. In other words, she cares not just about the components of a gamble, but about global features of the gamble, e.g. the minimum value, the maximum value, the interval between these two values, the variance, and so forth. The possibility that she might get the higher outcome does not entirely make up for the possibility that she might get the lower outcome: adding a good possibility (with some specific probability) to the mix does not make up for adding a bad possibility (with some specific probability) to the mix, or for lowering the minimum value of the gamble, even if the mean remains unchanged. I will call this explanation of risk aversion *global sensitivity*.

The global sensitivity explanation says that even holding fixed how much one likes various outcomes, and thus holding fixed the average value of a gamble, which outcomes constitute the gamble and how they are arranged matter. That is, properties of gambles that do not supervene on any particular outcome or chance of getting that outcome – global properties of gambles – matter. This does not rule out marginal diminishment in the value of goods. People who have diminishing marginal utility functions can still be sensitive to global properties; I suspect that most people's utility functions do diminish marginally for very large quantities of money, and yet that many people are globally sensitive.[15] So, whereas the expected utility theorist claims that risk averse behavior always entails a diminishing marginal utility function, I claim that it might indicate global sensitivity instead (or some combination of the two).

[14] Technically, while this is evidence for the claim that her utilities are linear in money, it is not enough to *entail* that her utilities are linear in money. After all, her utilities might diminish marginally in money, and also diminish marginally (along the negative axis) in the performance of unpleasant tasks: all we know is that the utility of tasks is linear in the utility of money. Expected utility theory cannot fix the differences in value between outcomes without knowing how an agent feels about gambles. However, if it is genuinely true that she cares about risk in the way I will suggest – she cares about the way outcomes are arranged across the possibility space, or the variance among outcomes – then expected utility theory will be unable to capture her preferences. There will be no utility function such that her preferences maximize the expectation of that function. The description of this example is meant to illustrate the intuitive reason that there will not be such a utility function.

[15] So far, you cannot have diminishing marginal *utilities* unless you obey the standard axioms of decision theory, since utility functions are defined only for people who obey these axioms. What I mean to suggest is that the following two psychological facts can hold of an agent simultaneously: the value of goods to an agent can diminish marginally and global properties can matter to how she evaluates gambles.

What my explanation does is separate the two dimensions of evaluation (of options) that get folded together by the standard theory: (1) how much an agent values certain ends and (2) how effective he rates various means of arriving at these ends to be.[16] This isn't exactly the distinction between how an agent values ends and how an agent *values* means. I am not pointing out, for example, that of two agents who share the same ends, one may think that hard work is a better way to his ends while the other thinks that schmoozing is a better way because of the intrinsic values of these means. Indeed, since ends in this case are total worlds, the value of getting good things through hard work and the value of getting them through schmoozing will be accounted for in the value the agent assigns to ends, because total worlds include facts about how an agent obtained his ends. Instead, when I talk about the agent rating the effectiveness of various means, I am talking about determining which means he thinks will most effectively achieve his ends, with only the goal of achieving those ends in mind. What I am pointing out is this: two agents could attach the very same values to certain ends (various sums of money, say). And yet, one agent might think that he can more effectively achieve his ends of getting as much money as he can by taking a gamble that has a small chance of a very high payoff, whereas the other might think that he can more effectively achieve *these same ends* by taking a gamble with a high chance of a lower payoff.

Expected utility theory claims that as long as the average values of the prizes in two gambles are the same, rationality requires that an agent does not care how the possibilities are arranged to yield that average: he must be indifferent as to whether all the possible value is concentrated in an outcome with small probability or whether the value is spread evenly over states. Another way of explaining the constraint that expected utility theory places on evaluating gambles is as follows: we can think of a gamble as a probability distribution over utilities. This distribution will have many features: a minimum, a mean value, and a variance, to name a few. According to expected utility theory, the *only* feature of this distribution that can (rationally) matter when evaluating a gamble is its mean value (i.e. its expected value); we are meant to throw out any other information as useless to determining the value of the gamble.[17] We will find preferences that violate the theory when we find agents who are sensitive to some of these other features of distributions. And since decision theory is supposed to represent all rational preferences, all of these globally sensitive agents are deemed irrational on the standard theory.

[16] Agents also evaluate the subjective probabilities of states; this might be thought of as a third dimension of evaluation. However, this does not make a difference to my arguments: as stated, for the purposes of this paper, we can take the probabilities to be known or already determined by the agent. I claim that two agents who assign the same values to outcomes and the same probabilities to states may rate the same gamble differently in terms of its effectiveness of realizing their ends.

[17] This way of making the point was suggested to me by Branden Fitelson.

So, as mentioned, a rational agent cannot value money linearly and still prefer \$50 to a coin flip between \$0 and \$100. Nor can Ralph, if he is rational, prefer deal 2 to deal 1, if my assumptions in the case are correct.[18] I will spell out one of these assumptions now: that the values of the outcomes (specifically, an Elvis towel and gloves) are independent. I mentioned that having one should not add to or decrease from the value of having the other; that is, receiving A is just as valuable whether or not one already has B, and vice versa. In utility terms, the values of each prize individually sum to the value of both prizes together: $u(A \& \sim B) + u(B \& \sim A) = u(A \& B)$.[19] Examples of goods that are obviously not independent include a left-hand glove and a right-hand glove: the right-hand glove is more valuable to me if I have the left-hand glove, and vice versa.[20] An Elvis towel and a different Elvis towel are also not independent: one is less valuable to me if I already have the other.[21] As a special case of outcome independence, if the utility of money is taken to be linear in small amounts, then receiving one dollar is valued independently of receiving another dollar: $u(\$1) + u(\$1) = u(\$2)$. So the fact that Margaret values money linearly is the fact that, for her, dollar amounts are valued independently. If the goods in Ralph's decision problem are independent, then the utility of deal 1 must be the utility of deal 2.[22]

The assumptions that Margaret values money linearly and that there are some goods that Ralph values independently do not seem controversial. However, in case the reader is worried that we too quickly assume that the outcomes in question are independent, Allais's example does

[18] Some of this paper will be devoted to critiquing an attempt, on behalf of the standard decision theorist, to show that they cannot be correct; see section 4.

[19] That is, $u(A \& B) – u(\sim A \& B) = u(A \& \sim B) – u(\sim A \& \sim B)$: the difference the presence of A makes is the same whether or not B is present. This is equivalent to $u(A \& \sim B) + u(B \& \sim A) = u(A \& B) + u(\sim A \& \sim B)$, which is equivalent to the above when we set the utility of the status quo (in which the agent has neither A nor B) to 0.

[20] $u(r \& \sim l)$ and $u(l \& \sim r)$ are each very small, so $u(r \& \sim l) + u(l \& \sim r) < u(l \& r)$.

[21] $u(E1 \& \sim E2)$ and $u(E2 \& \sim E1)$ are each almost as large as $u(E1 \& E2)$, so $u(E1 \& \sim E2) + u(E2 \& \sim E1) > u(E1 \& E2)$.

[22] Since the results of the coin flips are probabilistically independent, each state has probability ¼. So:

$$\begin{aligned} u(\text{deal 1}) &= \tfrac{1}{4}u(\text{towel and gloves}) + \tfrac{1}{4}u(\text{towel}) + \tfrac{1}{4}u(\text{gloves}) \\ &= \tfrac{1}{4}[u(\text{towel}) + u(\text{gloves})] + \tfrac{1}{4}u(\text{towel}) + \tfrac{1}{4}u(\text{gloves}) \text{ if the towel and gloves are independent} \\ &= \tfrac{1}{2}u(\text{towel}) + \tfrac{1}{2}u(\text{gloves}) \\ &= u(\text{deal 2}) \end{aligned}$$

That the expected utilities of the two deals are equivalent is perhaps easier to see if we evaluate the expected utility of the first coin flip, and then add it to the expected utility of the second. However, I wanted to flag exactly where the assumption that the goods were independent entered in: we can only evaluate the utility of two gambles by adding their utilities if they are independent in the relevant sense anyway. Of course, I said that we do not have intuitions about utility, so I cannot actually describe the utility of goods apart from an agent's preferences. But there is a way to get around this problem: a subset of the agent's preferences should be enough to fix the utility function, if one exists, as valuing the goods as independent (or we could use a subset of the agent's preferences to find *some* two goods that are independent, and use those in the example, in place of the Elvis towel and the gloves). If we fix the utility function from this subset of the agent's preferences, then if the agent prefers deal 2 to deal 1, there will be *no* utility function that represents his preferences.

not require any background information about how the outcomes are related in order to get a contradiction with expected utility theory. All we need to know is that a decision maker prefers L_1 to L_2 and L_4 to L_3: there is no assignment of utility values to the outcomes \$0, \$1m, and \$5m such that $u(L_1) > u(L_2)$ and $u(L_3) < u(L_4)$.[23]

To illustrate my point that Ralph and Margaret care about global features of gambles, we can consider Ralph's choice as a choice about which of the (equally likely) states he would like to enhance with the possibility of a pair of gloves, when presented with the following initial setup:

	HH	HT	TH	TT
Deal	*Elvis towel*	*Elvis towel*	*Nothing*	*Gloves*

If the possibility of gloves is added to the HT state, we get deal 1, and if it is added to the TH state, we get deal 2:

	HH	HT	TH	TT
Deal 1	*Elvis towel*	*Elvis towel and gloves*	*Nothing*	*Gloves*
Deal 2	*Elvis towel*	*Elvis towel*	*Gloves*	*Gloves*

Since Ralph prefers deal 2, he prefers that the gloves be added to the TH state. Thus, he has a preference about how values are arranged across the possibility space – and note that the fact that the goods are independent implies that the (utility) *value* added to each state is the same. He would like to add the gloves to a state in such a way that he increases the minimum he might receive from the gamble. Similarly, we can consider Margaret as faced with an initial gamble that yields \$0 if a fair coin lands heads and \$50 if it lands tails. Her choice is whether to add \$50 to the heads state or to the tails state – and if she would be adding the same value to the state in either case (that is, if her utility function does not diminish marginally), but prefers to add it to the heads state, then she cares about how the values are arranged across the possible states.

A standard utility theorist could question the assumption that the goods are really independent, or that the utility function is really linear, and I will address versions of these objections in section five: in particular, I will address the objection that the outcomes are not independent *in these particular deals*, because it is relevant to the value of the outcomes that things could have turned out otherwise. But I hope the standard theorist will admit that there are

[23] 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.

two different reasons that a person might prefer to add a good to one possible state rather than another (of equal probability), setting aside whether these reasons are rationally defensible: the goods might not be independent, so the value added might be different depending on the state to which it is added, *or* the value added would be the same, but the person cares about global properties of the gamble.

The most important criterion in a decision-theoretic model is that is accurately represents agents' preferences in the sense mentioned above: if the model says that the agent prefers X to Y (whatever the explanation for this preference), then the agent, or some suitably idealized version of that agent, does prefer X to Y. All the EU theorist is concerned with is that an agent's preferences maximize expected utility. And since utility is calculated from preferences, he is just concerned that for each agent, there is *some* utility function such that the agent maximizes expected utility.[24] The explanation does not matter: if I say that diminishing marginal utility is intuitively different from caring about risk, but there is still some utility function that can represent agents who care about risk in my intuitive sense, then what I say cannot be a strike against the expected utility model. But we have just shown that there are several cases in which people have preferences, stemming from how they treat risk, that are not capturable by any *expectational* utility function. This is (I claim) because the standard theory folds together how much an agent values particular ends and how he rates the effectiveness of various ways of arriving at these ends. Let us now turn to an alternate theory that allows for a wider range of attitudes towards risky gambles to count as rational.

2 Formal representation of the two dimensions of evaluation

In expected utility theory, utility functions vary from agent to agent, but utilities and probabilities interact in a set way. The utility of a gamble is its mathematical expectation: the utility of each possible outcome weighted by the probability of obtaining that outcome. So, the utility of a gamble between A and B, where the agent has a probability p of getting A and $1 - p$ of getting B, is $p(u(A)) + (1 - p)(u(B))$. This is equivalent to $u(B) + p[u(A) - u(B)]$. Taking B to be the less (or equally) desirable option, this latter formulation merely says that the value of a gamble will be the minimum value it guarantees plus the amount by which the agent might do better, weighted by the probability of doing that much better. So the possibility of getting a better outcome will increase the value of a gamble above its minimum value in a set way.

[24] More formally, that for each agent *S*, there is some function *u* that assigns (utility) values to outcomes such that for all gambles X and Y, *S* prefers X to Y (is indifferent between the two) iff *u* assigns a higher expected utility to X than to Y (the same utility to both).

Thus, as pointed out, in expected utility theory, agents evaluate options along a single dimension: they attach values to various outcomes.[25] But probabilities are still of set significance: e.g., if the probability of getting a good outcome (instead of nothing) is doubled, the value of a gamble (above the status quo) is doubled, even if the new probability is 1. As mentioned above, the value of a gamble that yields a good outcome (rather than nothing) with probability 0.5 must be half the value of that good outcome. Furthermore, once two decision makers agree on the values of various outcomes, they must evaluate gambles in exactly the same way: their preference ordering must be exactly the same, and their utilities must be exactly the same (up to unit and scale). However, it is plausible to think that some people are more cautious than others.[26] It is plausible that two agents who attach the same value as each other to $100 and $0 will not both attach the same value to a coin flip between $0 and $100; one may think the fact that he only has a 50% chance of winning the better prize diminishes the worth of the gamble more than the other does. In other words, aside from having different attitudes towards outcomes, two agents might have different attitudes towards potential ways of attaining some of these outcomes. And to repeat, these attitudes are not most plausibly read as 'actually' having different attitudes about the value differences between $0 and $50 and between $50 and $100. Rather, we should think the agents have different attitudes towards global properties.

One way these different attitudes can show up is that decision makers who are sensitive to global properties of gambles will weight the interval by which they could improve over the minimum by different amounts. For some, the possibility (with some specific probability) of improving over the minimum by some amount might not count heavily in the evaluation of the gamble – on the contrary, the minimum they are guaranteed to receive will be more important. These people will display risk averse behavior: if two gambles yield the same monetary value on average, but one yields a specific amount no matter what (and therefore that gamble has a higher minimum), they will prefer the latter. For others, the possibility of improving over the minimum will count for a lot – and the maximum they might receive will count heavily in the evaluation of the gamble. These people will display risk seeking behavior. And each agent might assess gambles as he does (either weighting the minimum or the maximum very highly) solely in service of his goal of getting what he values – that is, solely in service of ending up with the outcome he rates as best. He might think that taking a gamble with a higher minimum is the most effective

[25] Agents also attach (subjective) probabilities to various outcomes. Again, I am assuming for this paper that probabilities are fixed, or at least that we know the probabilities agents attach to each outcome.

[26] Can some people be more cautious than others without any of them being irrational? The answer is not obviously in the negative, and in chapters three through five, I will argue that it should be in the affirmative.

way to end up with what he wants, or he might think that taking a gamble with a higher maximum is the most effective way.

More generally, the amount by which any agent might improve above the minimum will be scaled by a *function* of the probability of doing that much better. That is, agents think about gambles as yielding at least their minimum, and then consider, and possibly discount (or amplify), the interval by which they might improve over the minimum and the probability of so improving.

We can state this formally as follows. Agents who are sensitive to global properties will have preferences that accord with a *weighted* expected utility function: the desirability of a gamble {A with probability *p*, B with probability 1 – *p*}, where A is at least as good as B, will be u(B) + **r(p)**[u(A) – u(B)], where r is the agent's "risk function" or "weighting function," adhering to the constraints r(0) = 0, r(1) = 1, r is non-decreasing, and $0 \leq r(p) \leq 1$ for all p.[27] In effect, the interval by which the agent might improve her lot above what she is guaranteed to get shrinks not by her chance of getting the better prize, but by a function of this chance, which reflects her attitude towards risk.[28] Thus the value of a gamble will be the minimum value guaranteed plus

[27] For now, I am only speaking of gambles between two options. See Appendix A for gambles between multiple options. My theory is closely related to two existing theories in the literature, Schmeidler's Choquet expected utility (described in Hong and Wakker (1996) and Kobberling and Wakker (2003)), and Quiggin's anticipated utility (described by Quiggin (1982); later called rank-dependent utility or cumulative utility) – and can be make equivalent to these theories, each under a particular assumption. Like these theories, my theory employs a weighting function that is sensitive to the ordering of outcomes. However, Choquet utility employs a weighting function of states, not of *probabilities* of states – i.e. it does not directly include an agent's judgments about probability at all. And anticipated utility uses an "objective" probability function, and imposes additional constraints. For example, Quiggin assumed – in my terminology – that r(½) = ½ for a gamble with two outcomes, though this assumption can be relaxed; indeed, if we relax this assumption, then the first formulation I present, with "objective" probabilities, is equivalent to versions of anticipated utility in which the weighting function is non-decreasing. My formulation entails that an agent attaches *subjective probabilities* to states, which is important for decision theorists who are interested in extracting beliefs (as well as desires) from preferences, though admittedly this difference is not essential to the arguments in this paper. For more on the relationship between this theory and other non-expected utility theories, see my book manuscript. Note further that I call this theory risk-weighted expected utility theory; but it bears no important relationship to another theory in the literature known as weighted expected utility theory. Finally, note also that I still use "u" and "utility" function in my proposal, even though my 'utility' function is not defined in the same way as a standard utility function and thus does not refer to the same quantity. I keep the terminology to suggest that my *u* function is trying to capture the same thing as the standard *u* function: the values of various outcomes. If the reader objects, he can call my function *u** and the quantity utility*, but I take "utility*" to be explicating the same psychological notion that expected utility theorists take "utility" to be explicating.

[28] The reader might wonder what exactly the risk function is supposed to 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 it is a desire for more or less risk (if it was the latter, then we would be able to account for it on the standard

the amount by which the agent could do better, weighted by this function of the probability of doing that much better.[29] If this function has a low value, then any improvement over the minimum will be heavily discounted; thus, the minimum will weigh heavily in the evaluation of the gamble. If it has a high value, then any improvement over the minimum will be amplified; thus, the maximum will weigh heavily. Since this theory includes a weighting function that reflects the agent's attitude towards risk, I will call it **risk-weighted expected utility theory** (hereafter, **REU**); but there is one more point about the I need to clarify.

A feature of the standard theory that I have not yet discussed but that is important to preserve is that gambles are not usually stated in terms of the *probabilities* of attaining each outcome, but rather in terms of the states of the world that result in each outcome. According to the tradition following Savage, gambles are functions from states to outcomes: each gamble is a complete list of which outcome results from that gamble in each state. So, we might construct a gamble that yields some outcome if a coin lands heads and yields some other outcome if the coin lands tails, but we would not assume in the setup that the gamble yields a 50% probability of each outcome. This is because probabilities are *subjective*: they are degrees of belief, and an agent may have any degree of belief in the proposition that the coin lands heads. It is assumed that an agent's degree of belief function, if he can be interpreted as having one, will behave like a probability function, but it is not generally assumed that the theorist knows the agent's degrees of belief. Even if the theorist states that the probability of some event is *x*, we cannot assume that the agent has degree of belief *x* in that event (he may think the theorist is lying, or he may think he believes the theorist while his decision making implies otherwise).

Thus, on the standard theory, the proper way of stating the utility of a gamble is as follows: the utility of the gamble {A if event E obtains, B if event ~E obtains}, where A is at least as desirable as B, u is the agent's utility function, and p is the agent's subjective probability function, is u(B) + p(E)[u(A) – u(B)]. And on my theory, we have the risk-weighted expected utility of the gamble:

$$REU(\{A \text{ if } E, B \text{ if } {\sim}E\} = u(B) + r(p(E))[u(A) - u(B)],$$

theory, as I discuss in section five). I admit that I do not have a satisfactory answer to this question. The most informative thing I can say is that the risk function is *not* a propositional attitude, but rather a structural feature (or relational property) of how beliefs and desires interact with preferences. Standard decision theory itself employs a structural assumption: an *expectational* function that explains the relationship between preferences on the one hand and beliefs and desires on the other. And if I am on the hook for saying what the relationship is between the structural feature I assume and our folk psychological concepts, then so is the standard theorist. However, standard decision theory assumes that the structure is the same across agents. So I have an additional problem: given that the risk function may be different in different agents, I must say what explains this difference. At the moment, I leave this question open to further speculation.

[29] For an example of this, see Appendix B.

where A is at least as desirable as B, and u, p, and r are the agent's subjective utility, probability, and risk functions.

When it is not crucial to the argument, I will sometimes speak of a gamble that yields an outcome with some specific probability, where I mean the probability that the *agent* attaches to the event that produces that outcome.

My proposal allows that agents evaluate gambles along three dimensions. First, like the standard theory, it allows them to attach subjective values to outcomes: it is up to agents to choose their ends – hence, a subjective utility function. Second, again like the standard theory, it allows them to gauge the relationship between various means and their ends: it allows them to gauge the likelihood of some particular gamble leading to some particular result – hence, a subjective probability (degree of belief) function. Third, *unlike* the standard theory, it allows them to subjectively judge which means are more *effective* at arriving at their ends. It is up to them to judge which gamble better realizes their end of getting more money (or their particular ends of getting $50, or, better by twice, getting $100). It is up to them whether they will better fulfill their goals by guaranteeing themselves a high minimum or by allowing themselves the possibility of some high maximum – and it is up to them how these two features of gambles trade off.

Let me briefly say something about why *this* modification of the standard theory, rather than a different modification that takes account of global properties. First, I want to remain within a consequentialist framework – a framework that says that only the outcomes matter – while allowing (unlike the standard theorist) that there are different ways of aggregating possible consequences. Second, I find my proposal to accord with how I intuitively think about risk, and experimental research shows that I am not alone.[30] Finally, though I will not discuss it in this paper, my theory follows from a set of axioms that are themselves intuitive constraints on rational preference – more intuitive, I claim, than those of the standard theory.[31]

[30] Kahneman and Tversky (1979). Kahneman's and Tversky's project is descriptive: they want to formalize how people actually make decisions. They also introduce a weighting function of probabilities, along with other modifications of the standard theory, such as a tendency to overvalue the status quo – to value losses relative to the status quo more negatively than gains relative to the status quo are valued positively – and a sensitivity to how decisions are framed. Usually, these empirical results are taken to show the varieties of irrationality that actual people exhibit, and my proposal does not allow for many of these tendencies because I think they are genuine instances of irrationality. However, I think the way in which actual people take risk into account has been mislabeled as an instance of irrationality, so I think that while rational agents should not, e.g., be sensitive to framing effects, they should be allowed to assign weights to probabilities. I take my theory to be a (maximal) theory of rationally permissible preferences.

[31] See my book manuscript (Buchak 2009). In particular, I accept a relaxation of the sure-thing principle identified by Hong and Wakker (1996) and Kobberling and Wakker (2003).

On my theory, the standard Allais preferences are perfectly acceptable, as are Ralph's and Margaret's preferences.[32] The new equation may look unfamiliar; but, in fact, under certain r-functions, it gives us familiar decision rules. For example, if $r(p) = \{0, p \neq 1; 1, p = 1\}$ for some agent, she uses the maximin decision rule:[33] she pays attention only to the minimum, and any possibility of doing better than the minimum adds no value to the gamble. Similarly simple r-functions produce the maximax rule and the "Hurwicz criteria" for decisions under uncertainty.[34] If you think that these policies are rationally permissible, then you have some reason to be sympathetic to my proposal.[35] More importantly, if $r(p) = p$ for some agent, she is a standard expected utility maximizer: she weights an improvement above the minimum by precisely her probability of realizing it.

An agent will display risk averse behavior, even if her utility function is linear (or if goods are independent), if her r-function is convex.[36] And here we come to the crux of the difference between how the standard theory explains risk averse behavior and how my preferred theory explains it: *on the standard theory, to be risk averse is to have a concave utility function (or, in the general case, a utility function that displays non-independence). On my theory, to be risk averse is to have a convex risk function*. The intuition behind the diminishing marginal utility explanation of risk aversion was that adding money to an outcome is of less value the more money it already has. The intuition behind my explanation of risk aversion is that adding to the probability of realizing an outcome is of more value the more probability that outcome already has of obtaining. In other words, risk averse people have increasing marginal risk functions: they prefer to make the outcome set less diverse, so to speak. Of course, my theory allows that the utility function is concave (or, indeed, any shape) – but on my theory, this feature, which describes how an agent evaluates outcomes, pulls apart from her attitude towards *risk*.

I will not argue extensively for my positive theory here, beyond pointing out that it is intuitive, that it takes care of the counterexamples, and that familiar theories – including the standard theory – are special cases of it.[37] The main purpose of introducing the theory in this paper is to set up a concrete opponent for the EU theorist. In the remainder of this paper, I will

[32] See Appendix C.

[33] This rule says "choose the gamble with the highest minimum." It is usually discussed in the context of decisions under uncertainty, i.e. when the agent does not have a precise probability function.

[34] See Appendix D.

[35] Even if you don't, my proposal suggests a way to argue about them, as well as a way to talk about considerations that might count in against them (e.g. that *r* must be continuous).

[36] Note that an agent will display risk seeking behavior, even if her utility function is linear, if her r-function is concave.

[37] For a more extensive treatment of my positive proposal, including comparisons to other non-expected utility theories, see Buchak (2009), Risk and Rationality. Manuscript.

consider arguments on behalf of EU theory: in the next two sections, arguments that agents who violate it are thereby irrational, and in the final section, an argument that it is conceptually impossible for rational agents to violate it.

3 The Sure-thing principle

Since risk-weighted expected utility theory is more permissive than expected utility theory (and generalizes expected utility theory), it must lack as least one constraint. Before I present this constraint formally, let me present the intuition it is supposed to formalize.

Suppose I am deciding between two options, say, driving on the highway and driving on the parkway; and I am uncertain about whether it will rain. I might then consult my preferences among the options in the event of rain, and in the event of no rain. I discover that if it rains, I am indifferent between the two options, since rain will prevent me from enjoying any scenery anyway – on either route, the outcome is "a drive without nice scenery." Then I can simplify my decision by only consulting my preferences about what will happen if it does not rain, and letting this preference determine my overall preference. In general, if I have to make a decision, and I don't know what will happen, one way to simplify my decision is to ignore all eventualities that result in the same outcome no matter what I choose. I will call this s*ure-thing reasoning*: if two gambles agree on what happens if one event obtains, then my preferences between them must depend only on my preferences between what they yield if this event does not obtain.

Savage's "sure-thing principle" formalizes this reasoning as follows:

> **Sure-Thing Principle (STP) For all X and Y (and for all E): If E and ~E are mutually exclusive and exhaustive events (sets of states), where s is not the null event, then, for all Z and W, I prefer {X if E, Z if ~E} to {Y if E, Z if ~E} iff I prefer {X if E, W if ~E} to {Y if E, W if ~E}.**[38]

It is helpful to represent this schematically:

Event	E	~E
Deal A	*X*	*Z*
Deal B	*Y*	*Z*

Event	E	~E
Deal C	*X*	*W*
Deal D	*Y*	*W*

The sure-thing principle says that as long as there is some possibility of *E* happening, then if I prefer deal A to deal B, I must prefer deal C to deal D, and vice versa.[39] If I prefer X to Y, then I prefer a gamble that yields X in some states to one that yields Y in those states, as long as the gambles agree on what happens otherwise. This implies that what happens in each state makes a contribution to the overall value of the gamble that is *separable* from what happens in the other

[38] Originally due to Savage (1954; 1972). This is the formulation that Susan Hurley (1989) uses, pg. 81.

[39] Since preferences are complete, this also implies that if I am indifferent between A and B, then I am indifferent between C and D.

states: each outcome makes the same contribution to the value of a gamble regardless of which other outcomes comprise that gamble. Thus the sure-thing principle ensures that preferences among gambles do not depend on irreducibly global properties of gambles, or on the relationships between the outcomes in various states. They do not depend on, e.g., the interval between the best outcome and the worst outcome, the uniformity of the outcomes across states, or the proportion of outcomes reaching a certain threshold. If they did, an outcome's contribution to the value of a gamble would depend on which other outcomes comprise that gamble.[40]

Any risk-weighted expected utility maximizer for whom $r(p) \neq p$ violates the sure-thing principle.[41] As an example, the common preferences in the Allais paradox violate STP. Consider the following schematic diagram of the choice the Allais agent must make, where the states are the drawings of one of a hundred equiprobable lottery tickets:

Ticket	1	2-11	12-100		Ticket	1	2-11	12-100
L_1	*$0*	*$5m*	*$0*		L_3	*$0*	*$5m*	*$1m*
L_2	*$1m*	*$1m*	*$0*		L_4	*$1m*	*$1m*	*$1m*

Take *E* to be the event in which ticket 1-11 is drawn. L_1 and L_2 agree on what happens in ~*E*, as do L_3 and L_4. So, according to STP, one's preferences between L_1 and L_2, and between L_3 and L_4, should just depend on whether one prefers $5m if tickets 1-10 are drawn and $0 if ticket 11 is drawn (that is, a 10/11 chance of $5m) or $1m for certain: if the former is preferred, L_1 and L_3 should be preferred, and if the latter is preferred, L_2 and L_4 should be preferred. But, as noted, L_1 and L_4 are commonly preferred.

Is this principle a constraint on rationality? On the one hand, it does not seem intuitive that rationality should rule out caring about irreducibly global properties; but on the other hand, the sure-thing principle seems intuitively compelling.[42] So how should we resolve this dilemma? I submit that the sure-thing principle only seems compelling because we are confusing it with another, more plausible principle.

To evaluate the plausibility of STP, you may have substituted various goods for each variable in the schema, such as:

[40] This is because whether an outcome increases the size of the interval between best and worst outcome, or increases the uniformity, or increases the proportion of outcomes reaching a certain threshold depends on what the other outcomes are.

[41] We know this is the case because risk-weighted expected utility theory, conjoined with the sure-thing principle, entails $r(p) = p$.

[42] Although it is important to note that it doesn't have the same intuitive appeal as, say, transitivity. Patrick Maher mentions that people don't usually change their preferences when told they violate STP, but they do when told they violate transitivity. For the former claim, he cites studies by MacCrimmon (1968), Slovic and Tversky (1974), and MacCrimmon and Larsson (1979); for the latter claim, he cites MacCrimmon (1968). Maher (1993), pg 65, 34.

Event	*Heads*	*Tails*
Deal A	Ice cream	Liver
Deal B	Pizza	Liver

Event	*Heads*	*Tails*
Deal C	Ice cream	Fruit
Deal D	Pizza	Fruit

And you probably thought something like: of course it doesn't matter – when deciding between A and B, or between C and D – what item is in the "tails" column, since if tails comes up, I'll get that no matter what; so my preferences between A and B (and between C and D) should just depend on my preferences between ice cream and pizza. And they should depend on them for the following reason: if I prefer ice cream to pizza, *I'll do at least as well by taking deal A (C) as by taking B (D), no matter how the coin comes up* (and vice versa); and this is true no matter what goods we plug into the tails column. And if I'll do at least as well by taking one deal rather than another, and possibly better – regardless of which state obtains – then I should prefer that deal to the other. This last claim is a generally accepted principle:

> **State-wise Dominance: if A is preferred or indifferent to B in all possible states of the world, and preferred in at least one state, then A is preferred to B.**

I have no immediate quarrel with this principle; and if any principle in the vicinity can be rightly called a restriction on rational preferences, this is it. So if the sure-thing principle follows from it – or is equivalent to it – then the sure-thing principle is itself a restriction on rational preferences. But STP does not follow from it, as I will now show.

State-wise dominance reasoning cannot be faulted in the example given. However, it fails to apply to all examples in which STP needs to hold, and so STP does not follow from it. In the example, we plugged in (non-risky) outcomes for X, Y, Z, and W; but for STP, these variables actually range not only over outcomes, but also over *gambles*. And when gambles are plugged in for the variables in the STP schema, there is not always one option that is better in every state, so we cannot always apply dominance reasoning.[43] For example, we might have the following possible deals:

Event	Heads		Tails	
	Red card	Black card	Red card	Black card
Deal A	Ice cream	Liver	Pizza	
Deal B	Pizza		Pizza	

Event	Heads		Tails	
	Red card	Black card	Red card	Black card
Deal C	Ice cream	Liver	Fruit	Liver
Deal D	Pizza		Fruit	Liver

In order for STP to be satisfied, an agent's preference between deal A and deal B must be the same as that between deal C and deal D, since these deals fit the above schema, where event *E* is a coin landing heads and event ~*E* is the coin landing tails. Now assume the agent prefers ice cream to pizza to fruit to liver. Then neither deal B nor deal A (and neither C nor D) have the

[43] Edward McClennen also makes this point in McClennen (1983), pg. 176-8.

feature that she'll do at least as well in every state as in the other deal; for deal A is better than deal B if the coin lands heads and a red card is drawn, but deal B is better than deal A if the coin lands heads and a black card is drawn (and similarly for C and D). Thus, we cannot invoke the state-wise dominance principle to prevent the agent from preferring, say, deal B to deal A because she does not want to run the risk of getting liver but deal C to deal D because as long as there is already some risk of getting liver, she does not mind increasing that risk in exchange for some possibility of ice cream. If an agent cares about global properties, then what happens in the tails event is relevant to her preference between two options, even if what happens is the same either way.[44]

The reason, I suspect, that STP initially sounds so intuitive is because the applications we immediately call to mind are those in which the variables range over outcomes. However, in restricting our attention to these applications, we are actually considering a much weaker principle, SWD. STP does not follow from SWD, because there are some applications of STP in which neither option is such that the agent will do at least as well in every state by taking that option as by taking the other. Interestingly enough, risk-weighted expected utility maximizers obey a principle intermediate between the two, which seems to capture both the idea that agents should use sure-thing reasoning when there are no global differences between gambles and the idea that agents can care about global properties of gambles. That is, risk-weighted expected utility maximizers obey a restricted version of STP.[45]

So what is at issue is whether rational agents are required to obey STP even when gambles differ in global properties. Therefore, to see whether STP should be endorsed in full, we cannot just consider its plausibility in a few cases. We need to consider the full range of cases, and construct a general argument for the sure-thing principle that applies even when gambles are plugged in for the variables in the principle. I will not take up that discussion here.[46] At the very least, I hope I have shown that STP is not an obvious principle of rationality.

[44] It might sound like I am begging the question in favor of those who care about global properties, but I am not; I pointed out that state-wise dominance does not entail STP, and caring about global properties is simply a reason for preferences that would obey state-wise dominance but not STP. In any case, it is certainly not obvious (without philosophical argument) that our preferences should be A > B & C > D or B > A & D > C, as it does seem obvious in the case in which outcomes were substituted in for the variables.

[45] In particular, they obey what Hong and Wakker (1996) call the Comonotonic Sure-Thing Principle. This principle is the restriction of the sure-thing principle to gambles within the same comoncone; that is, gambles such that the order in which the agent ranks the events (in terms of which prizes they yield) is the same – gambles that order the events in the same way as each other.
For a discussion of the relationship of this principle to global properties, see Buchak 2009.

[46] I discuss three important arguments for STP, the argument from the non-aversiveness of knowledge and the arguments from consistency over time and from consistency at a time, in Buchak 2009. These arguments are presented by Mark Machina in Machina (1991) and Patrick Maher in Maher (1993).

4 Should we care about expectation?

Let me turn to an argument that rational agents must obey (non-risk-weighted) expected utility theory as a whole; of course, this theory entails the sure-thing principle, so if this defense of expected utility theory succeeds, then as a consequence, rational agents must obey the sure-thing principle.

This defense tries to make explicit the connection between utility and *expected* utility. As mentioned in section 1, utility is not a real quantity. Rather, utility is just whatever quantity we want more of: whatever utility is (if it can be defined), an agent prefers an outcome that has higher utility to one that has lower utility. In this section, I will explore possible arguments for the claim that given that an agent prefers the outcome with the highest utility, an agent must prefer the *option* with the highest *expected* utility. I call this claim the Expecting claim:

> **(EX) If option A has higher *expected* utility than option B for a rational agent, then that agent prefers option A to option B.**

Note that I am not assuming that utility is defined independently of an agent's choices. If utility cannot be defined except via some theory that links preferences with utility, then we can read the Expecting claim as stating that rational agents must conform their preferences to a theory that identifies preference with expected utility. In other words, that the theory from which value (or utility) is derived must have it that we are indifferent between a basic outcome and a gamble that, on average, yields something equally valuable – or between two gambles that yield the same goods on average. The Expecting claim links *expected* utility and *actual* utility, regardless of whether this link is used to necessitate that an agent makes certain choices or to constrain how a correct theory defines his utilities.

The Expecting claim is historically significant: it can be found in the earliest formulations of decision theory by Blaise Pascal, and both Maurice Allais and Hilary Putnam take it to be the background to their arguments.[47] Pascal's original inception of decision theory had it that the *monetary value* of a stake in a game should be equal to its expected value.[48] Joyce notes two attractive features of Pascal's proposal that if a game ends prematurely, each gambler should be given as his share of the pot the monetary expectation of his stake.[49] The first is that a player's share of the pot depends only on two (objective) features of his situation: the amounts of money

However, Machina and I have different responses to the arguments. Another interesting discussion of whether STP is a rational requirement on preferences can be found in McClennen (1983).

[47] See Joyce (1999); Allais (1953); Putnam (1986).

[48] Interestingly enough, the move from talking about expected value to expected *utility* came in response to noted widespread aversion to risk in money.

[49] Joyce (1999), pg. 9-11.

he might have won and the probabilities he had of winning those amounts. The second is that the amounts of money that each of the players gets add up to exactly the amount of money in the pot. Neither of these features entails that the players should be indifferent between receiving their share of this division and continuing the game, although it may be difficult to argue that the division favors any particular player. More importantly, though, neither of these features is preserved when we move from the expected monetary value theory to expected *utility* theory: whereas the expected monetary value of a player's stake depended only on objective amounts of money and probabilities, the expected utility of her stake also depends on the (subjective) utility she assigns to each amount of money; and whereas the expected monetary values of each stake always add up to the total amount of money in the pot, the expected utilities of each stake clearly do not sum to the utility of the money in the pot.[50]

So why do we have reason to prefer the gamble with the higher expected utility? It cannot be because we are likely to do better in any particular instance by choosing a gamble with higher expected utility than we would by choosing a gamble with lower expected utility; we are not. For example, if we are choosing between a gamble that yields a prize worth 200 utiles if ticket 1 is drawn, and nothing if tickets 2-100 are drawn, and a gamble that yields a prize worth 1 utile if tickets 1-100 are drawn, then the former has a higher expected utility than the latter, even though I am *likely* to do better by taking the second gamble; that is, in 99 of the 100 (equally likely) possible states of the world, I will do better by taking the second gamble than by taking the first. Higher expected utility does not entail a likelihood of higher utility, because expected utility is not just responsive to the possibility of doing better: the size of the amount by which I stand to do better can outweigh the low probability of obtaining it.

Since expected utility is just an average of the utility of the gamble in all possible states of the world, the reason to care about it must make reference to more than one outcome – e.g. to what will happen over the long run or to what will happen in all possible worlds. I set this latter possibility aside[51] and consider the claim that rational agents should make decisions that maximize expected utility because *over the long run, an expected utility maximizer will end up with more utility than someone following another strategy, e.g., than someone who uses*

[50] Unless utility is linear in money and equivalent for every player. There is a problem about how to compare utility values across people: since utility is fixed only up to scale and unit for each individual, it is difficult to find a criteria for selecting a member of the class of utility functions that represent me and a member of the class of utility functions that represent you that have the "same" scale and unit. Therefore it does not even make sense to "sum" the expected (subjective) utilities of each player's stake.

[51] In Buchak 2009, I consider (and reject) the "possible worlds" defense of EU theory: that rational agents should make decisions that maximize expected utility because this amounts to making decisions that maximize the total utility across possible worlds.

maximin.[52] More precisely, assume that an agent faces a series of choices identical to the original decision problem, and that he must choose the exact same gamble each time. Then, as the number of (identical) choice situations becomes larger and larger, it grows more and more probable that taking the gamble with the highest expected utility (each time) will lead to a higher amount of average utility than taking any other gamble (each time); this is true by the law of large numbers.[53] If Ralph were to participate in identical instances of the referee's gamble ad infinitum, then over the long run, he'd (with near certainty) get the exact same average number of gloves and Elvis towels by taking deal 1 as by taking deal 2 – either way, he'd average half an Elvis towel and half a pair of gloves per instance of the gamble. So he shouldn't prefer one deal to the other or pay to have one deal rather than the other. Similarly for Margaret: over the long run, she'd average 50 utiles on each instance of the gamble whether she took $50 or the coin flip between $0 and $100; so she shouldn't prefer the former to the latter.[54] And for the Allais agent: depending on how he values the monetary outcomes, either L_1 and L_3 will lead to a higher average utility over the long run than L_2 and L_4, respectively, or L_2 and L_4 will.

So if one expects to participate in a gamble repeated ad infinitum, then maximizing expected utility will lead (with near certainty) to having a higher average utility from each instance of the gamble (and so to a higher total utility). In other words it will lead to doing better for oneself. I will first respond to this fact by showing that it does not prove enough: it does not show that one should maximize expected utility in any particular instance. I will then argue that if this fact proves anything, it proves too much: if we can justify expected utility maximization by reference to what will happen in the long run, then we can also justify other decision-making strategies, such as maximin, in this same way; and furthermore, we can even show that agents should follow the original expected *value* theory instead of expected utility theory.

For my first response, I will grant to the standard theorist that *near* certainty of doing better is enough to entail that if a situation is to be repeated ad infinitum, then one should adopt the policy of maximizing expected utility. Still, it is not obvious that one should care about maximizing expected utility in situations that are not to be repeated ad infinitum. What will happen in a very long amount of time need not resemble what will happen in a much shorter time.

[52] Recall that a person who uses maximin always chooses the gamble with the highest minimum.

[53] The law of large numbers states that as the number of trials of a random variable approaches infinity, the probability that the average of the trials is more than ε from the mean value of the variable approaches zero. By average utility, I mean the sum of the utility values that result from each gamble divided by the total number of gambles.

[54] Taking for granted that Margaret values money linearly, and setting the unit and scale (harmlessly) so that $50 is worth 50 utiles, $0 is worth 0 utiles, and $100 is worth 100 utiles.

Maurice Allais, who himself rejects the expectational theory in favor of a theory that takes into account the spread of possible outcomes resulting from an action, is responding to the claim that the expectational theory is the correct theory of rationality simply because it is expectational when he writes: "It would be improper to brand a cautious man irrational because he is willing to accept a lower psycho-mathematical expectation as the price of lower dispersion. Nor can a person who enjoys risk as such be labeled as irrational, merely because he accepts a lower psycho-mathematical expectation as the counterpart of the availability of some chance of extremely high psychological gain."[55] He immediately goes on to argue that expected value is not a good measure of actual value to the agent.[56]

The main points from Allais's argument can be summarized as follows: first, no agent can actually participate in an unlimited number of gambles with an unlimited bankroll. Second, many or most decisions will be isolated events, either unrepeatable by nature or unlikely to be repeated. Hilary Putnam also makes this second criticism in his "Rationality in Decision Theory and Ethics."[57] He asks us to consider whether someone who does not have much life left to live should be swayed by considerations of what would happen in the long run, and concludes that "If my *only* reason for believing that I should be reasonable were my beliefs about what will happen *in the long run* if I act or behave reasonably, then I would have absolutely *no* reason…to think it better to be reasonable in an unrepeatable single case."[58] Putnam still thinks that maximizing expected utility is the only rational thing to do, even if the situation is unrepeatable; however, he thinks that the only response available to the standard decision theorist is to say that this is true but that he cannot and need not justify it. Though this response might satisfy a die-hard believer in EU theory, it won't convince the agnostic to accept it.

The EU theorist might try to respond to both criticisms by pointing out that even though an agent will not face the exact same decision over and over again, he will most likely face an extraordinary number of decisions throughout his lifetime, and adopting a policy of maximizing expected utility will, with increasing likelihood, get him the most average utility. However, this is not what the law of large numbers says: it only talks about the expected value of a single

[55] Allais (1953), pg. 115. Note that Allais identifies utility with "psychological gain," to contrast it with purely monetary gain.

[56] Allais (1953), pp. 116-118.

[57] Putnam (1986), pp. 3-16.

[58] Putnam (1986), pg. 9, 12. Italics Putnam's.

variable. Moreover, refusing to maximize expected utility in a single instance will "wash out" in the long run, so there is no reason to be always bound by the policy.[59]

It seems this line of reasoning will not work for the defender of EU theory. But let us grant, for the sake of argument, that we can justify a single instance of a decision-making strategy (such as expected utility maximization) by showing that if the situation were to be repeated, it would be preferable in the long run to adopt that strategy each time. Then, I argue, we can prove too much: for if this argument (from what would happen in the long run to what an agent should do in some particular instance) is valid, then advocates of other decision rules can use a parallel argument to justify *those* rules.

I've already pointed out that because of the way utility is defined, everyone has a preference for getting more utility rather than less. And the expected utility maximizer interprets taking the most effective means to satisfying this preference to require that the agent choose so that, with near certainty, the gambles he takes average out to the highest utility. However, there are other standards of success. The maximinimizer can interpret taking the most effective means to require that the agent get the most he can be *guaranteed* to end up with – and, in the long run, choosing the gambles that a maximin strategy recommends will give him as high a guarantee (and usually higher) as choosing any other gambles. Unless the expected utility maximizer literally takes infinity instances of a gamble, there is still some (vanishingly small) chance that things will turn out as badly as they could turn out in every instance. After any finite number of trials, an agent might end up with the minimum value of his chosen deal in every trial – the coins may land TH each time, and Ralph may end up with nothing – and the minimum (or guaranteed) value of the maximinimizer's chosen deal will be at least as high as, and sometimes higher than, that of the expected utility maximizer's chosen deal. In other words, the expected utility maximizer is succeeding by his own interpretation of the standards, and the maximinimizer is succeeding by *his*.

The second point I want to make towards showing that this argument proves too much is that even if we accept the EU maximizer's interpretation of the standards, we can produce an argument of the same form but for an incompatible conclusion. This argument shows that an agent who cares about getting the highest average of the quantity he prefers (were the gamble repeated ad infinitum) should not be maximizing expected utility at all – rather, he should be maximizing expected *monetary value*. The argument runs as follows: assume an agent cares

[59] Granted, this problem comes close to problems like the "tragedy of the commons," in which it is the case that any particular instance of deviation will wash out, but widespread deviation will not. Perhaps it is too much to require that the EU theorist solve this problem in order to justify his theory using this defense.

about doing better on average, and that he prefers larger amounts of money to smaller. If the gamble is to be repeated ad infinitum, then, again by the law of large numbers, it will grow increasingly likely that taking the gamble with the highest expected *monetary value* leads to a larger average amount of *money* than taking any other gamble. And since the agent prefers larger amounts of money to smaller (indeed, larger amounts of money have higher utility), he should maximize expected value.

The original argument and my argument of the same form have contradictory conclusions. It seems we have shown that agents interested in getting higher utility in repeated gambles over the long run should both maximize expected utility and maximize expected monetary value. But these do not come to the same thing (unless utility is linear in monetary value). To spell out this contradiction more precisely, the law of large numbers implies that taking the gamble with the highest expected utility leads to the highest average utility (per trial), and that taking the gamble with the highest expected monetary value leads to the largest average amount of money (per trial).[60] And since the average amount of quantity *x* in *n* trials is just the total amount of *x* divided by *n*, then after *n* trials, taking the gamble with the highest expected utility should lead to the highest total utility, and taking the gamble with the highest expected monetary value should lead to the largest amount of money. But larger amounts of money have more utility; so we have an argument that two incompatible strategies each lead to higher total utility than any other strategy (including each other).

What is going on here? As it turns out, a gamble can yield the highest average utility on each trial without having the highest total utility – if by average we mean we *first* take the utility of the monetary prize a gamble yields on *each* trial (ignoring which other prizes the agent has won on previous trials), and *then* sum these utilities (then divide by the total number of trials), and by total we mean *first* add the monetary prizes together and *then* take the utility of that number. That a repeated gamble yields a higher average utility on each trial (without regard to what else the agent has) does not entail that someone taking this repeated gamble will be left with a better collection of prizes at the end. This is because, as already stressed, the utility of two prizes together is not always the sum of their utilities (e.g., u($50) + u($50) ≠ u($100)). Prizes are not *always* independent. Utility values of prizes, unlike monetary values, are different depending on how one has done in past gambles (especially if utilities diminish marginally).

To make it concrete: let us assume an agent's utility function assigns a utility of 0 to $0, 0.6 to $25, and 1 to $100. And he is offered a choice between option A ($25), which has an

[60] I will omit the caveat that this is only *nearly* certain (i.e. only with increasing likelihood), because it makes no difference to what follows.

expected utility of 0.6, and option B, a fair coin flip between $0 and $100, which has expected utility 0.5. Option A had a higher expected utility, and, after, say, 1000 trials, it will yield an average utility of 0.6, a total monetary amount of $25,000, and a total utility of u($25,000). Option B has a higher expected *value*, and after 1000 trials, it will yield an average utility of 0.5, an average monetary value of (roughly) $50, a total monetary amount of (roughly) $50,000, and a total utility of u($50,000) – a higher total utility than Option A.

So the expected utility maximizer only does better with respect to the *average* utility (the average of the utilities of each trial) over the long run; he cannot claim that his winnings, when taken together, will have a higher *total* utility in the long run. Nor can he claim they will be guaranteed higher than by following a different strategy, such as maximin. The standard theorist still has a long way to go to justify EU maximization as the uniquely rational strategy.

This does not mean, incidentally, that the expected utility maximizer will do worse than the expected value maximizer when he encounters a gamble that is *actually* to be repeated. When faced with a choice among gambles of which one gamble is to be repeated ad infinitum, the expected utility maximizer will calculate the expected utility of each collection he is considering (that is, each collection that consists of repeated instances of one gamble). As emphasized above, since the utility of two prizes together is not always the sum of their individual utilities, the expected utility of this collection will not be the number of gambles multiplied by the expected utility of an individual gamble. The gamble he will choose to be repeated is the gamble corresponding to the *collection* that has the highest expected utility. If the gamble is to be repeated ad infinitum, this will be the collection corresponding to the gamble with highest expected monetary value. So my counterargument does not show that the expected utility maximizer is behaving irrationally by his own lights. What it does show is that the initial argument is unsound: maximizing expected utility in a single instance cannot be justified by citing the 'fact' that if the gamble were to be repeated and he were to make the same choice every time, the expected utility maximizer would do better (by his own lights) than a decision maker following any other strategy.

This argument was doomed from the beginning. Discussing what an agent should do in a repeated gamble in order to show that an agent should not care about risk straightforwardly begs the question. Repeating a gamble is a way to minimize risk, because repeating a gamble reduces the variance; in the limit, the variance goes to zero. Thus, it is obvious that someone who cares about variance will not behave in the non-repeated situation as he will in the repeated situation.[61]

[61] This brings up an objection: does this disconnect between preferences in the non-repeated situation and the repeated situation mean that if we offer a risk-weighted expected utility maximizer a gamble that is to

Justifying a single instance of expected utility maximization by reference to what happens in the long run obscures the very thing at issue: namely, whether variance (and related properties) can matter to a rational agent.

5 Individuation of Outcomes

In the first section, I gave several examples in which people tend to violate expected utility theory. Of course, people's preferences only violate EU theory if we assume that the outcomes are as stated in the examples. In this section, I will address a common move that expected utility theorists make to salvage their theory in light of preferences that seem to violate it. This move consists in individuating outcomes so that the preferences in question do obey the axioms of the theory.

To justify this move, the theorist points out that the outcomes are not stated in enough detail to capture everything of value to the agent. In my examples, the theorist might claim that it is relevant to the value of the outcomes that things could have turned out otherwise. Specifically, it is relevant to the value of not receiving any prize that Ralph might have gotten something had he chosen differently, because he would feel regret, which itself has negative value. Thus, the correct description of the outcomes is really more fine-grained than as initially presented in the problem; the options are:

	HH	HT	TH	TT
Deal 1	*Elvis towel*	*Elvis towel and gloves*	***Regret***	*Gloves*
Deal 2	*Elvis towel*	*Elvis towel*	*Gloves*	*Gloves*

The outcome in TH is not simply the status quo, so Ralph's preference for deal 2 over deal 1 no longer violates expected utility theory.[62]

Alternatively, if this does not seem to correctly describe Ralph's preferences, I've said that Ralph prefers deal 2 to deal 1 because he is sure to win a prize. The standard decision theorist could say that this is because the fact of a getting a prize for certain adds value to each outcome in deal 2, perhaps because Ralph enjoys anticipating a prize: e.g. "gloves and

be repeated, he will act in his interests in each instance of the gamble and thereby act against his interests in the gambles taken together as a whole? Here the REU maximizer has a similar reply as the one given by the EU maximizer in the previous paragraph. Both, if they know a gamble is to be repeated, will consider which whole collection of gambles maximizes (risk-weighted) expected utility. And for neither agent will the value of a collection of gambles always be the sum of the values of each individual gamble: for the EU maximizer because values are not always independent (utility is not always additive), and for the REU maximizer because the probability of a final outcome depends on all the gambles in play (relatedly, the r-function is not always additive). For more on the relationship between an individual gamble and repeated instances of the same gamble, see Buchak 2009.

[62] u(deal 1) = ¼u(towel and gloves) + ¼u(towel) + ¼u(gloves) + ¼u(regret), and since the term "u(regret)" does not appear in the equation for the value of u(deal 2), there is no necessary connection between the two.

anticipating some prize" is better than "gloves." We could describe the outcomes to take this into account:

	HH	HT	TH	TT
Deal 1	*Elvis towel*	*Elvis towel and gloves*	*Nothing*	*Gloves*
Deal 2	*Elvis towel* ***& surety***	*Elvis towel* ***& surety***	*Gloves* ***& surety***	*Gloves* ***& surety***

Again, Ralph's preference for deal 2 over deal 1 no longer violates expected utility theory.[63]

The same strategy works in response to the Allais paradox, and indeed is standardly employed. The theorist points out that it is relevant to the value of receiving \$0 in L_3 that you might have received \$1m had you chosen differently, because you feel regret, which itself has negative value. Thus, the correct description of deal L_3 is "\$1,000,000 with probability 0.89, \$5,000,000 with probability 0.1, \$0 *and regret* otherwise." If these are the outcomes in L_3, then the common preferences no longer violate expected utility theory.[64] Again, there are other ways to individuate outcomes to salvage the theory: L_4 might be "\$1m with probability 1, and surety." I do not mean to privilege any particular description; the point is that if some of the states in which the agent receives \$1m (or \$0) are different from some of the other states in which she receives \$1m (\$0), then the classic preferences no longer violate the sure-thing principle, or indeed expected utility theory.[65]

There are other classic examples of re-individuation of outcomes in response to purported violations of the standard axioms, but I will not discuss them all here.[66] However, I will mention one more example because it will become an important contrast to examples involving risk. John Broome uses the re-individuation strategy in response to a common criticism of utilitarianism: that utilitarianism cannot account for our intuition that inequality is a bad thing. Specifically, that a person's good depends not just on her own situation, but also on the relation between her

[63] u(deal 2) = ½u(towel and surety) + ½u(gloves and surety); again, neither of the terms (nor "u(surety)" alone) appear in the equation for the value of u(deal 1).

[64] Expected utility theory can now capture the preference for L_1 over L_2 and L_4 over L_3: there is no contradiction between the following two equations:

$$0.1(u(\$5m)) + 0.9(u(\$0)) > 0.11(u(\$1m)) + 0.89(u(\$0))$$

$$u(\$1m) > 0.89(u(\$1m)) + 0.1(u(\$5m)) + 0.01(u(\$0 \text{ and regret}))$$

On the contrary, it is easy to find utility functions that satisfy both equations.

[65] This strategy can also be employed for Margaret's preferences, although the re-individuation may be more complex: instead of a coin flip between \$0 and \$100, the gamble is a coin flip between \$0 with regret and \$100 – or "\$0 as the result of a gamble that had a 50% probability of \$100" and "\$100 as the result of a gamble that had a 50% probably of \$0." Formulations of the latter type may be necessary if we know a lot of her preferences, e.g. if there is some utility function and r-function (that is not the identity function) that represent her preferences under risk-weighted expected utility theory.

[66] See, e.g. Pettit (2002). Pettit also cites Peter Diamond (1967). I discuss differences between these cases at length in Buchak (2008), "Risk Without Regret," unpublished manuscript.

situation and others' situation – that it is worse for her if she has less good than other people.[67] Broome points out that utilitarianism can take account of this, simply by making an individual's good a function of her "basic" good (apart from the matter of equality) and the good (or bad) of equality (her "complaint"). For example, if my total good is the good I derive from my income minus the bad of how much poorer I am than my neighbor,[68] then utilitarianism captures our intuition that it is better for me to earn $100 when my neighbor earns $50 than it is for me to earn $100 when my neighbor earns $200. Making good a function of both the amount of money an agent receives and the amount other agents receive is a way of finely individuating the outcomes in which an agent receives the same amount of money. The relevant outcomes – the ones that the agent has decision-theoretic preferences over – specify what happens at "locations" other than the agent.

There are two motivations for re-individuation. The first is the view that the outcomes are genuinely under-described in many of these problems. The second comes from a philosophical picture of what the probability function and the utility function represent. On this picture, an agent's degrees of belief and desire are simply those entities which play the correct functional roles in the theory: namely, degrees of belief and (degrees of) desire are the values of a probability function and a utility function that represent the agent as an expected utility maximizer.[69] Since we are not trying to get the utility function to line up with something else "in the head," there is more leeway in assigning utility values. However, on this picture, as well as on a prescriptive picture of the aims of decision theory, there must be some constraints on how outcomes are individuated. Otherwise, when does it all end? If we can always re-describe outcomes whenever an agent's preferences seemingly conflict with decision theory, then the theory will not tell us anything substantive about which preferences are rational, or about how we should interpret agents.

For example, assume we have an agent who chooses chicken over steak at one time, and steak over chicken at another. One way to interpret her might be to say that she prefers steak to chicken when the tide is high in Alaska, and chicken to steak when it is not. This interpretation makes her consistent, but if she does not have any knowledge of the Alaskan tide, and if this

[67] Broome (1991), pp. 180-182.
[68] This is the equation Broome uses. Specifically, if we have two agents, then the individual good of one agent is $g_1 = \underline{g_1}(y_1) – \max\{0, a(y_2 – y_1)\}$, where y_x is the income of agent x, $\underline{g_x}$ is the good she derives from that income, and a is the poorer agent's complaint. Ibid.
[69] See, e.g., Hurley (1989). Hurley (1989), especially chapters 4 and 5. Here, Hurley cites, among others, Broome (1990), Lewis (1983), and Davidson (1986). Of course, adopting this picture by itself does not obviate the need for the expected utility theorist to respond to my counterexamples, since it is still be an open question which theory is the correct one.

event is wholly unconnected with her decision, then this interpretation will not help us get at her real beliefs and desires. Thus, if we are interested in decision theory as a framework for discovering an agent's beliefs and desires, we want to rule out certain interpretations of the agent's beliefs and desires, interpretations that don't make sense of what she is doing. And if we are prescriptive theorists, we also want to rule out some sets of preferences, as does the agent who looks to decision theory as a practical guide to her actions: upon realizing that she prefers chicken to steak when the tide is high, the agent need not be in doubt about her preference when it is low (and when all else is the same) – and we need not be in doubt about what advice to give her about that preference. In other words, we might think that "chicken when the Alaskan tide is high" and "chicken when the Alaskan tide is low" should not count as different options for rational agents. It is clear that we need a restriction on when preferences can be finely individuated and when they cannot be.

Broome has such a restriction.[70] He calls it the Principle of Individuation by Justifiers:[71]

> **(PIJ) Outcomes should be distinguished as different if and only if they differ in a way that makes it rational to have a preference between them.**

So "chicken when the Alaskan tide is high" and "chicken when the Alaskan tide is low" are the same outcome, since it is not rational to have a preference between them.[72] Therefore, they must be interchangeable – if one is preferred to steak, the other must also be – and the agent must be indifferent between them. Actually, strictly speaking, Broome thinks that *any* two outcomes can be individuated, but that rationality requires an agent to be indifferent between some of them. But as he points out, (PIJ) implies a rational requirement of indifference: unless two outcomes differ in a way that makes it rational to have a preference between them, an agent must be indifferent between them.[73] Unless chicken-at-high-tide and chicken-at-low-tide differ in a way that makes it rational to prefer one to the other, the agent must be indifferent between them, and therefore, again, they must be interchangeable. Since the difference between the two principles does not matter for our discussion, we will follow Broome in using (PIJ) rather than an indifference principle, which he does for expository reasons.

With this principle in place, Broome has an ingenious argument that no rational agent can violate the sure-thing principle. Remember, we only need decision theory to accommodate Ralph's preferences if they are rational. Broome's argument, if it succeeds, sidesteps the debate

[70] Philip Pettit (2002) has a slightly different restriction. I discuss this in Buchak (2008).

[71] Broome (1991), pp. 103.

[72] It might sometimes be rational to have a preference between them, in which case they would count as different outcomes. The point is that whenever we want to say that two outcomes are not (relevantly) different, we can point out that it is not rational to have a preference between them.

[73] Broome (1991), pp. 103-104.

about the correct theory of decision-theoretic rationality, since he argues that agents who are rational, *however we spell out this concept*, cannot violate STP. We will see how this argument works in the case of the Allais paradox, and then generalize it. Broome writes:

> "All the [rationalizations of the Allais preferences] work in the same way. They make a distinction between outcomes that are given the same label in [the initial presentation of the options], and treat them as different outcomes that it is rational to have a preference between. And what is the argument that Allais's preferences are inconsistent with the sure-thing principle? It is that all the outcomes given the same label [initially] are in fact the same outcome. If they are not…[the decision problem] will have nothing to do with the sure-thing principle. Plainly, therefore, the case against the sure-thing principle is absurd. It depends on making a distinction on the one hand and denying it on the other."(Broome 107)

Broome points out that in order to use the Allais paradox to show that rational agents violate the sure-thing principle, one needs to show both that the common Allais preferences are rational, and that they violate the sure-thing principle. In order to show that they violate the sure-thing principle, one must show that the outcomes that appear the same in the original choice problem should not be individuated (i.e., that the original choice problem really is an instance of the sure-thing schema). That is, we need to show that there is *no rational difference between the outcomes*. However, if the preferences are rational, it must be true that there is a difference between some of the outcomes that appear the same – a difference that the agent can rationally care about.

I mentioned above that there are several different ways in which the decision theorist can individuate outcomes in the Allais paradox so that the common preferences do not violate standard decision theory. And, Broome presumes, any way of rationalizing the preference for L_1 over L_2 and L_4 over L_3 will make reference to a difference in some of the outcomes: it will be interpretable as one of the ways to individuate outcomes more finely than they are individuated in the original set-up of the problem. If it is rational for the agent to distinguish between those outcomes, then the decision theorist can also distinguish between them, and the agent's preferences will be rational but will not violate STP. On the other hand, if it is not rational for the agent to differentiate those outcomes, then the decision theorist cannot do so either, and the agent's preferences will violate STP but will not be rational; so it will not matter that they violate the theory.

This argument can be extended to any purported violation of the sure-thing principle. If the agent can justify a purported violation of STP, then (by Broome's line of reasoning) it will be by reference to differences among some of the outcomes that initially appear the same; but then the decision theorist will point out that the preferences over options with the newly described outcomes do not actually violate the sure-thing principle.

Broome's argument is clever. However, it makes an assumption that (I argue) is false: *that any way of arguing that the common Allais preferences (or any STP-violating preferences) are rational relies on making a distinction between outcomes that are initially given the same label*. Obviously, if this is not the case, then Broome's argument does not go through: for if there are decision situations in which it is rational to have preferences that violate the sure-thing principle, but in which it is not rational to have a preference between outcomes that are, as stated, the same – and therefore in which the decision theorist must not individuate preferences any more finely then they are already individuated – then it is rational to violate the sure-thing principle, and individuation will not save standard decision theory.[74]

Broome's assumption, I claim, is not true in the case of the Allais preferences: it is not that people have rational preferences between $1m as the result of a gamble and $1m without the worry of gambling, or between $0 with regret and $0 without regret, but rather that $1m and $0 contribute something different to gambles L_3 and L_4 than they do to L_1 and L_2. *My justification for the Allais preferences does not depend on distinguishing between outcomes that appear identical, but on how identical outcomes contribute to the overall gamble.* Again, I do not think, once the gamble has been decided and the agent is left holding $0, $1m, or $5m, that what she could have gotten makes any difference to the value of her actual winnings. And similarly with Ralph: once we know whether he has an Elvis towel, gloves, both, or only what he had before he took the gamble, there is nothing more to know about the value of his holdings. But when Ralph and the Allais agent approach the gambles before their results are known, how outcomes are distributed over the various states of nature may make a difference; identical outcomes may not affect the value of two gambles equally, since the part they play in a gamble depends on other possible outcomes of the gamble.

Broome's argument fails because he assumes that any differences in what outcomes contribute to a gamble must be differences in the outcomes themselves. He makes this explicit: "The value that Allais associates with interactions between states, is really *dispersed* amongst the states themselves. In this, I was faithfully following all the available rationalizations of Allais's preferences; they all depend on feelings of some sort…Nearly all the published counterexamples to the sure-thing principle are like this."[75] Surely there is a way of reading any counterexample to STP as having different outcomes than originally thought; but that is not the only way of reading them, and that is not always the correct way to read them. The properties that globally sensitive

[74] At least, it won't save it if we employ a principle like (PIJ). It could still save a decision theory that is not supposed to have any normative content (e.g. one whose axioms are trivially true and in which re-individuation is always allowed).

[75] Broome (1991), pg. 110.

agents value – like low variance or a high minimum – are properties that attach to gambles before their results have been determined. After an agent wins his prize, he will not care whether it was the result of a risky gamble or was simply given to him. This is why the values of these properties cannot be dispersed among the states; they will not truly be the values of the outcomes by themselves.[76] Since riskiness is a property of a gamble before its result is known, it need not, so to speak, leave a trace in any of the outcomes.

So the main disagreement I have with Broome comes down to this: on his picture, (rational) preferences that are sensitive to the riskiness of options can only shows up as different specifications of what the outcomes are. They might show up as feelings the agent has about receiving one outcome rather than another – for example, as regret – or they might show up as feelings the agent suffered by getting the outcome in a particular way – for example, as anxiety about getting the outcome as the result of a gamble, instead of as a no-fail alternative. There is no room on Broome's picture, as there is on mine, for risk to enter into an agent's feelings about a gamble but *not* about any particular outcome.

I claim that *it is rational to be indifferent between two outcomes but to allow them to make different overall contributions to different gambles*. Ralph might have reasons to be indifferent between receiving nothing and receiving nothing when he might have gotten gloves – that is, to not care about what might have been had he taken a different gamble. He might not care about particular non-actualized possibilities: the fact that he could have gotten certain amounts of money or prizes, had the world turned out differently, does not affect the value of his property in the actual world. It does not affect what he can buy with the money he has, or the amount of pleasure he can get from what he has. In other words, counterfactual money won't pay the bills, or make them harder to pay in the actual world. To put it more concretely, if an agent prefers X to Y, then there is some amount of money he is willing to pay to have X rather than Y (or would be willing to pay if that were possible). But Ralph might not be willing to pay to eliminate possibilities that were never realized. For example, he might not pay any amount of money to trade in "nothing when I might have had gloves" for "nothing when I could not have had gloves" or "$0 when I might have had $1,000,000" for "$0 when I couldn't have had $1,000,000" (if this were possible). We surely cannot blame him for that. And yet, these reasons for not caring about non-actualized possibilities do not undermine his reasons for caring about global properties: the former are all reasons that apply *after* the coin has been flipped. They are

[76] And if we try to incorporate them as such, we will get incorrect answers about Ralph's other preferences, e.g. our representation might entail that he prefers gloves without risk to gloves as the result of risk, when, as his other preferences will indicate, this really doesn't make a difference to the value of gloves for him. See Buchak (2008).

considerations about how the values of the outcomes should be affected by other (non-actualized) outcomes. And they apply precisely because (and when) the outcomes are *non-actualized*. So clearly they cannot apply before the coin has been flipped, when the actualization or non-actualization of the possibilities is not yet known to Ralph. To summarize: caring about risk while all the possibilities are still on the table need not entail experiencing regret in some of these possible occurrences.

Contrast taking risk into account in the outcomes with taking inequality into account in the outcomes, as above in Broome's defense of utilitarianism. Ralph's two considerations against taking risk into account in the outcomes do not hold in the case of taking inequalilty into account in the outcomes. What goods other people have does affect my standing in the actual world (by affecting how far my money goes, or my relative prestige, for example). Furthermore, it might be rational (from a purely self-interested point of view) for me to pay to reduce my neighbor's wellbeing.[77] And anyone who does not care about what other people have, or would not pay to decrease inequality by altering his neighbor's fortune,[78] seems to not really care about inequality. But someone who does not care about non-actualized possibilities might still reasonably care about risk.

To counter Broome's argument, I showed that there could be reasons for having STP-violating preferences that do not entail differences between outcomes. And the failure of the argument highlighted an important difference between Broome's treatment (and the standard treatment) of the counterexamples and mine that I have been pointing to all along: I think riskiness should be treated as a property of a gamble as a whole, not as a property of any particular outcome. Caring about risk – or, more precisely, evaluating risky gambles in a way that is different from simply averaging the values of the outcomes – need not correspond to caring about any features of the outcomes besides their stated features.

If an agent genuinely cares about the riskiness of a gamble, then expected utility theory cannot capture his preferences. And if, as I have argued in the previous sections, it is rationally permissible to take risk into account in a non-standard way (e.g. to maximize *risk-weighted* expected utility), then there are rationally permissible sets of preferences that expected utility theory cannot accommodate.

[77] Broome also mentions that it be rational to reduce my neighbor's wellbeing. Broome (1991), pg 96.

[78] Setting aside moral or politeness considerations against doing this.

6 Conclusion

Decision theory ingeniously connects belief, desire, and preference; however, the theory as it currently stands assumes a particular connection among them that rules out being sensitive to certain global properties of gambles that are intuitively part of their riskiness. There are several sets of preferences – e.g. those of Ralph, Margaret, and the Allais agent – that seem reasonable but that decision theory is unable to capture. Allais, of course, noticed this; and there has been a more recent spate of psychological literature showing how people actually make decisions. However, these deviations from the standard theory have traditionally been thought of as failures of rationality on the part of decision makers, and whereas the theories from psychology are intended to be purely descriptive, I want to put forth a theory of *rational* preferences among risky gambles. I do not think that all of the noted tendencies of people to deviate from the standard theory are rational, but I think that preferences stemming from attitudes towards risk in particular deserve a more sympathetic treatment.

My proposal is not a rejection of the standard theory from the outside, by someone who thinks decision theory has little to offer; on the contrary, I think that by relaxing a certain assumption and adding another parameter, we can arrive at a theory that is better able to represent the full range of agents with rational attitudes towards risky gambles.

MATHEMATICAL APPENDIX

Appendix A: Gambles between multiple options

The way I've set up the risk-weighted expected utility function emphasizes that an agent considers his possible gain above the minimum he is guaranteed (the interval between the low outcome and the high outcome), and discounts that gain by a factor which is a function of the probability of obtaining the gain, a function that depends on how he regards risk. Analogously, when he stands to obtain one of *more than two* possible outcomes, it seems natural that he should consider the possible gain between each neighboring pair of outcomes and his chance of arriving at the higher outcome or better. 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 risk-weighted expected utility of the gamble should be u(\$1) + r(½)[u(\$2) – u(\$1)] + r(¼)[u(\$4) – u(\$2)]. This method of calculating gambles is a "bottom up" approach, in which the agent is treated as if he starts off with the worst option, and at each stage takes a gamble to see if he moves up to being guaranteed the next worst option. That is, it emphasizes the minimum utility. We could instead emphasize the maximum utility: treat the agent as if he starts off with the best option and weights the probability of doing no better than the next best option, and so forth. I find the emphasis on the minimum more intuitive. In any case, it makes no formal difference, because the "top down" and the "bottom up" approach will lead to the same weighed-expected utility values, if we transform the r-function accordingly.

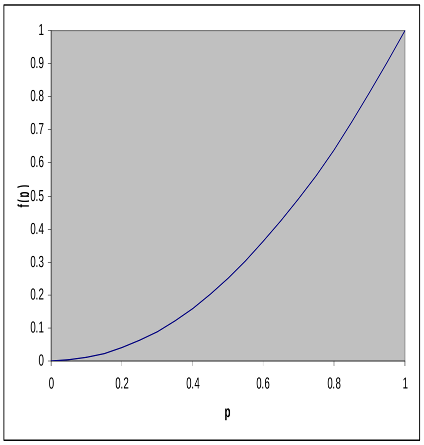
In general, the risk-weighted expected utility of a gamble {O_1, p_1; O_2, p_2; …; O_n, p_n}, where $u(O_1) \leq \ldots \leq u(O_n)$, is:

$$u(O_1) + r(\sum_{i=2}^{n} p_i)(u(O_2) - u(O_1)) + r(\sum_{i=3}^{n} p_i)(u(O_3) - u(O_2)) + \ldots + r(p_n)(u(O_n) - u(O_{n-1})$$

And the risk-weighted expected utility of an act in which the probabilities of events are not given {O_1, E_1; O_2, E_2; … ; O_n, E_n}, $u(O_1) \leq \ldots \leq u(O_n)$ is:

$$u(O_1) + r(\sum_{i=2}^{n} p(E_i))(u(O_2) - u(O_1)) + r(\sum_{i=3}^{n} p(E_i))(u(O_3) - u(O_2)) + \ldots + r(p(E_n))(u(O_n) - u(O_{n-1}))$$

Appendix B: Example of r-function



Here is an example of a possible r-function. For this agent, the following gambles and sure-thing monetary amounts are equivalent (assuming he values amounts of money under \$10 linearly):

Gamble {prize, probability; prize, probability}	Equivalent sure-thing amount
{\$0, 0.5; \$10, 0.5}	\$2.50
{\$0, 0.5; \$5, 0.5}	\$1.25
{\$1, 0.5; \$9, 0.5}	\$3.00
{\$0, 0.3; \$10, 0.7}	\$4.90
{\$0, 0.5; \$2.50, 0.5}	\$0.63
{\$2.50, 0.5; \$10, 0.5}	\$4.38
{\$0.63, 0.5; \$4.38, 0.5}	\$1.57

Appendix C: Ralph's, Margaret's, and Allais Preferences are all acceptable on REU

Ralph

	HH	HT	TH	TT
Deal 1	*Elvis towel*	*Elvis towel and gloves*	*Nothing*	*Gloves*
Deal 2	*Elvis towel*	*Elvis towel*	*Gloves*	*Gloves*

Assuming without loss of generality that the towel is weakly preferred to the gloves and u(nothing) = 0:

$\mathrm{REU}(\text{deal } 1) = u(\text{nothing}) + r(0.75)[u(\text{gloves}) - u(\text{nothing})] + r(0.5)[u(\text{towel}) - u(\text{gloves})] + r(0.25)[u(\text{towel and gloves}) - u(\text{towel})]$

$= u(\text{nothing})[1 - r(0.75)] + u(\text{gloves})[r(0.75) - r(0.5)] + u(\text{towel})[r(0.5) - r(0.25)] + u(\text{both})[r(0.25)]$

$= u(\text{nothing})[1 - r(0.75)] + u(\text{gloves})[r(0.75) - r(0.5) + r(0.25)] + u(\text{towel})(r(0.5))$, invoking Independence

$= u(\text{gloves})[r(0.75) - r(0.5) + r(0.25)] + u(\text{towel})(r(0.5))$, invoking $u(\text{nothing}) = 0$.

$\quad = u(\text{gloves})[r(0.75) - r(0.5) + r(0.25)] + u(\text{towel})(r(0.5))$

$\mathrm{REU}(\text{deal } 2) = u(\text{gloves}) + r(0.5)[u(\text{towel}) - u(\text{gloves})]$

$\quad = u(\text{gloves})[1 - r(0.5)] + u(\text{towel})(r(0.5))$

Ralph can prefer deal 2 to deal 1 without contradiction.

Margaret

$u(\$x) = x$; $r(p) < p$ for $p \neq 0,1$ will accommodate Margaret's preferences.

Allais

Ticket	1	2-11	12 – 100
L_1	*\$0*	*\$5m*	*\$0*
L_2	*\$1m*	*\$1m*	*\$0*
L_3	*\$0*	*\$5m*	*\$1m*
L_4	*\$1m*	*\$1m*	*\$1m*

$L_1 > L_2 \Leftrightarrow u(\$0) + r(.09)[u(\$5m) - u(\$0)] > u(\$0) + r(.1)[u(\$1m) - u(\$0)]$

$\quad \Leftrightarrow r(.09)[u(\$5m) - u(\$0)] > r(.1)[u(\$1m) - u(\$0)]$.

$L_4 > L_3 \Leftrightarrow u(\$1m) > u(\$0) + r(.99)[u(\$1m) - u(\$0)] + r(.09)[u(\$5m) - u(\$1m)]$.

These two equations do not contradict, so an REU maximizer can have the standard Allais preferences.

Appendix D: r-functions that are equivalent to various decision rules

Maximin, the agent picks the gamble with the highest minimum value: $r(p) = \left\{ \begin{matrix} 0, p \neq 1 \\ 1, p = 1 \end{matrix} \right\}$

Maximax, the agent picks the gamble with the highest maximum value: $r(p) = \left\{ \begin{matrix} 0, p = 0 \\ 1, p \neq 0 \end{matrix} \right\}$

Hurwicz criteria,[79] the agent sets a coefficient of optimism α, weights the best possible outcome by α, weights the worst possible outcome by 1 - α, and sums them:

$$r(p) = \left\{ \begin{matrix} 0, p = 0 \\ \alpha, p \neq 0,1 \\ 1, p = 1 \end{matrix} \right\}$$, where α is the coefficient of optimism.

Expected utility maximization: r(p) = p

[79] Cited in Ellsberg (1962).

Works Cited

Allais, Maurice (1953). "Criticisms of the postulates and axioms of the American School." In Rationality in Action: Contemporary Approaches, Paul K. Moser, ed. Cambridge University Press, 1990. (Reprint of 1953 original).

Broome, John (1990). "Rationality and the Sure-Thing Principle." In Rationality, Self-Interest, and Benevolence, ed. Gay Meeks. Cambridge: Cambridge University Press.

Broome, John (1991). Weighing Goods: Equality, Uncertainty and Time. Blackwell Publishers Ltd.

Davidson, Donald (1986). "A Coherence Theory of Truth and Knowledge." In Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson, ed. Ernest Lepore. Blackwell Publishers.

Diamond, Peter (1967). "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: A Comment." *Journal of Political Economy* 75.

Ellsberg, Daniel (1962). Risk, Ambiguity and Decision. Routledge, 2001. (Reprint of 1962 original).

Hurley, Susan (1989). Natural Reasons, Personality, and Polity. Oxford University Press.

Hong, Chew Soo and Peter Wakker (1996). "The Comonotonic Sure-Thing Principle." *Journal of Risk and Uncertainty* 12, pg. 5-27.

Jeffrey, Richard (1965). The Logic of Decision. McGraw Hill.

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

Kobberling, Veronika and Peter Wakker (2003). "Preference Foundations for Non-expected Utility: A Generalized and Simplified Technique." *Mathematics of Operations Research* 28, pg. 395-423.

Lewis, David (1983). "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61:4.

MacCrimmon, Kenneth R. (1968). "Descriptive and Normative Implications of Decision Theory." In Risk and Uncertainty, eds. Karl Borch and Jan Mossin. New York: St. Martin's Press.

MacCrimmon, Kenneth R. and Stig Larsson (1979). "Utility Theory: Axioms versus 'Paradoxes." In Expected Utility Hypotheses and the Allais Paradox, eds. Maurice Allais and Ole Hagen. Dordrecht: D. Reidel.

Machina, Mark (1991). "Dynamic Consistency and Non-expected Utility." In Foundations of Decision Theory, Michael Bacharach and Susan Hurley, eds. Basil Blackwell.

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

McClennen, Edward (1983). "Sure-thing doubts." Reprinted in Decision, Probability, and Utility, eds. Peter Gärdenfors and Nils-Eric Sahlin. Cambridge University Press, 1988.

Pettit, Philip (2002). "Folk Psychology and Decision Theory," reprinted in Pettit, Rules, Reasons and Norms, Oxford University Press.

Putnam, Hilary (1986). "Rationality in Decision Theory and Ethics." *Critica* 54.

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

Savage, Leonard (1954). The Foundations of Statistics. John Wiley & Sons, Inc.

Slovic, Paul and Amos Tversky (1974). "Who Accepts Savage's Axiom?" *Behavioral Science* 19, pg. 368-73.

Wakker, Peter and Amos Tversky (1993). "An Axiomatization of Cumulative Prospect Theory." *Journal of Risk and Uncertainty* 7:7, 147-176.