Gandalf's Solution to the Newcomb Problem

Ralph Wedgwood

> 'I wish it need not have happened in my time', said Frodo.
>
> 'So do I,' said Gandalf, 'and so do all who live to see such times. But that is not for them to decide. All we have to decide is what to do with the time that is given us.'
>
> Tolkien (1954, 60)

In a recent article, Andy Egan (2007) has presented some compelling counterexamples to causal decision theory (CDT). In this article, I shall outline a new theory of rational decision. This theory is designed to stay as close as possible to CDT, while accommodating Egan's intuitive counterexamples in a principled way.

In recent philosophical debates, the most prominent rival to causal decision theory (CDT) has been *evidential* decision theory (EDT).[1] Both CDT and EDT are versions of expected utility theory. We do not need to worry about which of the many possible interpretations of "utility" we should adopt here. (For example, on some interpretations, the relevant "utility function" is in a sense constructed out of the relevant agent's preferences, while on other interpretations, it is a measure of how objectively good or beneficial the available courses of action will be.) For our present purposes, this question does not matter. The crucial point for our purposes is that both of these theories define a rational choice as a choice that in some sense *maximizes expected utility* – where the "expected utility" of a choice is defined as the weighted sum of the choice's utility according to each member of a relevant set of hypotheses about one's situation, when each of these utilities is weighted by the relevant *probability* of the hypothesis.

CDT and EDT differ from each other in two ways. First, they differ in their view of what I have just called “the relevant set of hypotheses about one’s situation”; secondly, they differ in their view of the relevant sort of “probabilities”, which are to be used in defining the expected utility of the choice. On the first point, both CDT and EDT agree that the “relevant set of hypotheses about one’s situation” must form a *partition* – that is, a set of propositions about one’s situation such that one is rationally certain that *exactly one* of these propositions is true. However, the two theories differ on which sort of partition is relevant here. According to EDT, the relevant set of hypotheses can be *any* partition of propositions about one’s situation whatsoever.[2] According to CDT, on the other hand, the relevant partition must be a partition of *states of nature* – that is, one must be rationally certain that it is completely beyond one’s control which of the propositions in this partition is true.[3] (According to many versions of CDT, these “states of nature” – which are sometimes called “causal dependency hypotheses” – consist of conjunctions of “non-back-tracking” counterfactuals, where each of these counterfactuals has the form ‘If I did act A_n, outcome O_m would result’.[4] But we need not assume that states of nature must always take this form. The important point for our purposes is just that one is rationally certain that it is utterly beyond one’s control which of these states of nature one is actually in.)

Secondly, according to CDT, the relevant probabilities are simply the *unconditional* probabilities of these states of nature – that is, the probabilities that correspond to the degree of credence that it is rational for one to have in each of these states of nature. According to EDT, on the other hand, the probabilities that are relevant to evaluating a possible course of action *A* are *conditional* probabilities, reflecting the conditional degree of credence that it is rational for one to have in the relevant propositions about one’s situation, *on the assumption that* one chooses

A – whereas the probabilities relevant to evaluating a different course of action *B* are the conditional probabilities that one rationally assigns to each of these propositions on the assumption that one chooses *B*. (So, according to EDT, the probabilities that it is rational to use for evaluating one course of action *A* may in principle be *different* from the probabilities that it is rational to use for evaluating another course of action *B*.)

In the first section of this paper, I shall briefly present some of the cases that are discussed in Egan (2007), and I shall explain why neither CDT nor EDT can be reconciled with my intuitive judgments about these cases. I shall also argue against variants of EDT and CDT that impose an additional condition on rational decision – namely, the condition that a rational decision must be "ratifiable". It seems then that we need a new theory of rational decision.

In the rest of this paper, I shall make a first stab at articulating a new theory. The fundamental idea of this new theory is that although it agrees with CDT in focusing on a partition of *states of nature*, and agrees with EDT in using *conditional* rather than unconditional probabilities, it disagrees with *both* theories in that it is not a version of expected *utility* theory. Instead of using the concept of "utility" as the relevant measure of the value of each of the available options' possible outcomes, it uses a different measure of value instead – specifically, what I shall call the options' "comparative value" in each of the relevant states of nature.

In Section 2, I shall briefly present some reasons for thinking that the relevant probabilities that should guide rational choice are *conditional* probabilities, of roughly the sort that are employed by EDT. In Section 3, I shall present some reasons for thinking that the relevant measure of value is a purely *comparative* measure of how the available options compare with each other *within* each state of nature – a measure that completely ignores any comparisons of *absolute* levels of value or desirability that can be made *across* distinct states of nature. Then, in Section 4, I shall explain how this approach handles the troublesome cases that were discussed in Section 1. Unfortunately, as we shall see in Section 5, this approach raises a pressing question, which I am not certain how it is best to answer. In Section 6, I shall canvas two possible answers to this question. Finally, in Section 7, I shall defend this approach against the main objection that many decision theorists will be inclined to raise against it.

1. Counterexamples to EDT and CDT

Let us start with Nozick's (1969) presentation of the original Newcomb problem:

> Suppose a being [call her 'Alice'] in whose power to predict your choices you have enormous confidence. (One might tell a science-fiction story about a being from another planet, with an advanced technology and science, who you know to be friendly, and so on.) You know that this being has often correctly predicted your choices in the past (and has never, so far as you know, made an incorrect prediction about your choices), and furthermore you know that this being has often correctly predicted the choices of other people, many of whom are similar to you, in the particular situation to be described below. One might tell a longer story, but all this leads you to believe that almost certainly this being's prediction about your choice in the situation to be discussed will be correct. There are two boxes …

Suppose that one box is opaque, and the other is transparent. You can see that there is $1,000 in the transparent box, but you do not know what is in the opaque box. You have the choice of either taking both boxes, or taking just the opaque box. You know that Alice has made a prediction about which choice you will make; if she predicted that you would take only the opaque box, she put $1,000,000 in the opaque box – whereas if she predicted that you would take both boxes, she put *nothing* in the opaque box.

Since you are (to all intents and purposes) *certain* that Alice's prediction was correct, you are also (to all intents and purposes) *conditionally* certain that, given the assumption that you choose both boxes, there is nothing in the opaque box, and also conditionally certain that, given the assumption that you choose only the opaque box, there is $1,000,000 in the opaque box. On the other hand, let us suppose that the *unconditional* probability of there being $1,000,000 in the opaque box is p, and the unconditional probability of there being nothing in the opaque box is $1 - p$.

To simplify the case, let us also suppose that your utilities are exactly proportional to the monetary payoffs. Then the situation is given by the following tables:

Payoffs

	$1 million in opaque box; $1,000 in transparent box	$0 in opaque box; $1,000 in transparent box
You choose opaque box	$ 1,000,000	$ 0
You choose both boxes	$ 1,001,000	$ 1,000

Conditional Probabilities

	$1 million in opaque box; $1,000 in transparent box	$0 in opaque box; $1,000 in transparent box
You choose opaque box	1	0
You choose both boxes	0	1

Evidentially Expected Utilities

	$1 million in opaque box; $1,000 in transparent box	$0 in opaque box; $1,000 in transparent box	**Total**
You choose opaque box	$ 1,000,000	$ 0	$ 1,000,000
You choose both boxes	$ 0	$ 1,000	$ 1,000

Unconditional Probabilities

$1 million in opaque box; $1,000 in transparent box	$0 in opaque box; $1,000 in transparent box
p	$1 - p$

Causally Expected Utilities

	$1 million in opaque box; $1,000 in transparent box	$0 in opaque box; $1,000 in transparent box	**Total**
You choose opaque box	$\$ 1,000,000 \times p$	$\$ 0 \times (1 - p)$	$\$ 1,000,000 \times p$
You choose both boxes	$\$ 1,001,000 \times p$	$\$ 1,000 \times (1 - p)$	$\$ 1,001,000 \times p + \$ 1,000 \times (1 - p)$

In this case, the evidentially expected utility of one-boxing is clearly higher than that of two-boxing. So, in this case, EDT implies that one-boxing is the unique rational choice to make. On the other hand, whatever the unconditional probabilities of the two relevant states of nature may be, the causally expected utility of two-boxing is bound to be higher than that of one-boxing. So CDT is bound to say that two-boxing is the unique rational choice to make.

Like many (although admittedly not all) philosophers, it seems to me that in this case, CDT is right and EDT is wrong. Two-boxing is rational and one-boxing is irrational. Leaving money on the table is just foolish. As David Lewis (1981, 5) put it, EDT endorses “an irrational policy of massaging the news”. That is, EDT endorses the perverse choice to act in such a way as to give yourself good news about which state of nature you are in, even though there is absolutely nothing that you can do to determine which of these states of nature you are actually in, and even though the alternative course of action is bound to be preferable whichever state of nature you are in. So this case shows that without some further refinement, EDT cannot be the correct theory of rational choice.

Unfortunately, there are counterexamples to CDT as well. Here is the *Psychopath Button* case, which is one of the central examples of Andy Egan (2007). Suppose that what you want is that – unless it turns out that you are a psychopath yourself – all psychopaths should be exterminated. On the other hand, if it *is* the case that you are a psychopath yourself, your preference for staying alive strongly outweighs your preference for exterminating psychopaths. As it happens, you can press a button that will kill all psychopaths. But you are virtually certain that only a psychopath would press the button. You are also for some reason convinced that if you *don’t* press the button, that will prove beyond reasonable doubt that you are not a psychopath.

Utilities

	You are a psychopath	You are not a psychopath
You press the button and kill all psychopaths	– 90	+ 10
You do not press the button	0	0

Unconditional Probabilities

You are a psychopath	You are not a psychopath
p	$1 - p$

Causally Expected Utilities

	You are a psychopath	You are not a psychopath	**Total**
You press the button and kill all psychopaths	$- 90 \times p$	$+ 10 \times (1 - p)$	$- 90 \times p +$ $10 \times (1 - p)$
You do not press the button	$0 \times p$	$0 \times (1 - p)$	0

In this case, if p is smaller than 0.1, CDT will imply that the unique rational choice is to choose to press the button. But intuitively, it seems that the unique rational choice here is to choose *not* to press the button. After all, in pressing the button, you should be certain that you are psychopath, and that by pressing the button you will kill yourself (along with all the other psychopaths). So it seems that it would be irrational for you to choose to press the button in this case.

Egan (2007) gives an extensive defence of this intuitive judgment on the Psychopath Button case; I shall not repeat at length what he says there.[5] As he points out, it is not plausible to argue that cases of this kind are impossible: for you to be in this case, all that is required is that for some reason it is *rational* for you to have a high *conditional credence* that if you choose to press the button, you are almost certainly a psychopath. It is hard to see how one could show that it cannot possibly to be rational for you to have conditional credences of this kind.

As Egan also points out, the same case also seems to show that we cannot solve this problem by insisting that a rational choice must be "ratifiable" in the sense that was defined by

Richard Jeffrey (1983, 15–20). In general, the choice of a course of action A_i is "ratifiable" if and only if there is no other course of action A_j such that the conditionally expected utility of A_j, on the assumption that one chooses A_i, is greater than the conditionally expected utility of A_i on the assumption that one chooses A_i.

In the *Psychopath Button* case, *neither* course of action is ratifiable. So according to the ratifiability approach, this situation would be a paradoxical situation in which there is no possibility of making a rational choice. This can be illustrated by means of the following tables.

Conditional Probabilities

	You are a psychopath	You are not a psychopath
You press the button and kill all psychopaths	1	0
You do not press the button	0	1

Conditionally Expected Utilities (on the assumption that you choose to press)

	You are a psychopath	You are not a psychopath	**Total**
You press the button and kill all psychopaths	− 90 × 1	+ 10 × 0	− 90
You do not press the button	0 × 1	0 × 0	0

Conditionally Expected Utilities (on the assumption that you choose *not* to press)

	You are a psychopath	You are not a psychopath	**Total**
You press the button and kill all psychopaths	− 90 × 0	+ 10 × 1	+ 10
You do not press the button	0 × 0	0 × 1	0

In this case, given the assumption that you choose *not* to press the button, pressing has a higher expected utility; and given the assumption that you choose to press, *not* pressing has a higher expected utility. So neither option is ratifiable. Yet it is rational to choose not to press. This is not a paradoxical situation in which there is no rational choice that you can make.[6]

In this way, then, none of the best-known approaches gives a satisfactory verdict on all these cases. We need a new approach.

2. Why conditional probabilities?

It is clear that one crucial factor in Egan's Psychopath Button case is the difference between the *conditional probability* of your being a psychopath (i) conditionally on your choosing to press the button and (ii) conditionally on your choosing not to press the button. So any decision theory that can accommodate Egan's counterexamples will have to involve the idea that the relevant probabilities to use when deciding what to do are *conditional* probabilities, of roughly the sort that are employed by EDT. In this section, I shall raise a couple of considerations in support of this idea.

I suggest that this is in fact a general feature of rational reasoning – including both practical reasoning and theoretical reasoning. Practical reasoning is the kind of reasoning that involves forming or revising one's *choices* or *intentions* about what to do, while theoretical reasoning is the kind of reasoning that involves forming or revising one's *beliefs* about what is the case. As I shall argue, it is not a special feature of practical reasoning that it needs to use conditional probabilities of the sort that are employed by EDT; essentially the same point is true of theoretical reasoning as well.

In general, one needs to rely on probabilities because one is ignorant and uncertain, and so does not know which beliefs are, in the most objective sense of the term, the “right” beliefs to hold, or which choices are the “right” choices to make. But to see how one should take account of such probabilities in forming and revising one’s beliefs and choices, it may be useful to see what beliefs or choices would be (in this objective sense) the “right” beliefs or choices to form – that is, which beliefs or choices count as completely achieving what we might call the “ultimate purpose” of the sort of reasoning that one was engaged in.

As many epistemologists have claimed,[7] it seems plausible that the ultimate purpose that one is pursuing when engaged in theoretical reasoning is – to put it roughly – to believe the truth and nothing but the truth about the question that is at issue. (This rough statement needs to refined in many ways, but we may ignore all the necessary refinements here.) So, it might seem tempting to think that whenever a proposition is true, and one consciously considers this proposition, then the belief that is (in this objective sense) the right belief for one to form is simply a belief in this very proposition.

In fact, however, this tempting thought would be mistaken. It might well be that the following proposition is true:

> It is raining and I don’t believe that it is raining.

However, even if this proposition is in fact true, it will *not* be true if I believe it (at least on the plausible assumption that belief distributes over conjunction, so that one cannot believe ‘*p* & *q*’ without believing *p*).[8] So the fact that this troublesome proposition is true, and that I consider this proposition, does not seem to be enough to make it the case that believing this proposition is the “right” belief for me to hold. In general, for a proposition to be such that believing it will achieve the ultimate goal of theoretical reasoning, it is not enough that the proposition should simply be true: the proposition must have the *conditional* property of being such that it will be true *if one believes it*.

The point that I have just made concerns when a belief is the “right” belief to hold – or to put it in other terms, when believing a proposition achieves the “ultimate purpose of theoretical reasoning”. But it seems to me that an analogous point holds of the notion of *rationally* believing a proposition, given one’s evidence. My evidence might make it *virtually certain* that this troublesome proposition is true. That is, it might be virtually certain given my evidence that it is raining and I don’t believe that it is raining. But it seems to me that it is not rational for me to believe this proposition, even for a split second. Why is it irrational to believe this proposition? I suggest that the explanation is that this proposition has a very low *conditional* probability, *given the assumption that I believe it*. If that is right, then presumably it is true quite generally that the probability that should guide me in forming and revising my belief in a proposition *p* is the conditional probability of *p on the assumption that I believe it*.[9]

In short, because the goal of theoretical reasoning is not just the truth, but *believing* the truth, it seems plausible that the probabilities that should guide one in forming and revising one’s beliefs in whatever propositions are in question are the conditional probabilities of each of those propositions on the assumption that one believes it.

A similar point, it seems to me, holds of practical reasoning. There may be some ways in which you might act that would have utterly wonderful outcomes; but it does not follow that you can achieve these wonderful outcomes through *choosing* to act in any of those ways. One problem is that it might be that *choosing* to act in one of these ways will not result in your *actually* acting in that way. (Perhaps you will succumb to temptation and abandon your choice at the last moment; or perhaps your hand will "tremble", resulting in your failing to act as you have chosen to.) It is not clear exactly how a theory of practical reasoning should cope with this problem. I shall simply sidestep this problem by restricting my attention to cases in which each of the options (or courses of action) that you are choosing between is such that it is rational for you to have a very high conditional credence – virtually amounting to conditional certainty – that given the assumption that you choose that option (or course of action), you will in fact carry out your choice (and so will indeed act in the way that you have chosen).[10]

A second problem that needs to be considered, however, is that in some cases, it could be that even if acting in a certain way will have an utterly wonderful outcome if you act in that way *without* having chosen to do so, acting in that way as a result of *choosing* to do so will prevent the wonderful outcome from coming about. In some of these cases, it will not be objectively right for you to not choose to act in this way. In general, what it is objectively right to choose is some way of acting which is such that *if* you choose it, you will carry out your choice, and thereby bring about a suitably valuable outcome.

Now it seems to me that a broadly similar point applies to the question of what it is *rational* for you to choose, in cases where you do not know for certain what the outcome of your choice will be. The probabilities that should guide you in choosing between two courses of action *A* and *B* are conditional probabilities: in the case of *A*, they are the conditional probabilities, *given the assumption that you choose A*, of various hypotheses to the effect that doing *A* will have a certain sort of value to a certain degree; and in the case of *B*, the relevant probabilities are the conditional probabilities, *given the assumption that you choose B*, of various hypotheses to the effect that doing *B* will have the relevant sort of value to a certain degree.

This then, roughly, is why it seems *prima facie* plausible that the relevant sorts of probabilities that one should rely on in deciding what to do are conditional probabilities, of broadly the same kind as those that are relied on by EDT.

3. Frodo and Gandalf

Consider the dialogue that I have chosen as the epigraph to this paper, from the second chapter of Tolkien's *Lord of the Rings*, in which Gandalf responds to a remark of Frodo's. (Of course, Gandalf's response is little more than a truism; but in philosophy, few things are more important than achieving a correct understanding of such truisms.)

In this dialogue, Frodo expresses the wish that Sauron had not returned in his lifetime, and that Bilbo had never found the ring and given it to Frodo. Gandalf acknowledges that such wishes are universal (and so presumably natural). But he goes on to draw a sharp distinction between *wishes* and *decisions*. Wishes of the sort that Frodo has just expressed are simply irrelevant for the business of making decisions. There is a certain "time", as Gandalf puts it, that is simply "given" to us; our task in making decisions is not to contemplate possible alternatives to this "time" (as we do when we wish that things were different from how they are), but merely to decide what to do with this time that is given to us.

What exactly is this "time" that is "given" to us, which Gandalf is speaking of here? I suggest that this "time" consists of all the *facts* about the situation in which we must act which it is *utterly beyond our power* to change or affect in any way. In other words, the "time" that is "given" to us is precisely what advocates of CDT would refer to as a "state of nature". Specifically, it is not merely a possible state of nature; it is the state of nature that *actually* *obtains*.

Now, of course, some of these "times" or "states of nature" are a great deal nicer than others. Indeed, some of these states of nature (such as those that Frodo has to face in Tolkien's tale) are downright nasty. We naturally wish to be in a nice state of nature, and not a nasty one; and we naturally wish that we were in a nicer state of nature than the state of nature that we actually are in. But Gandalf insists that for the purposes of decision making, it is irrelevant whether we are in a nice state of nature or in a nasty state of nature. There is simply nothing that we can do to affect which state of nature we are in. Indeed, for the purposes of deciding what to do, we *do not even need to know* whether we are in a nice state of nature or a nasty one. All that we need to know is which of available courses of action are better, and which are worse, than the available alternatives, in the state of nature that actually obtains. The best that we can possibly do is to make a decision that is at least as good, in the state of nature that actually obtains, as any other decision that we could have made in that state of nature.

Of course, different theories of practical reason will differ quite profoundly about the sort of "goodness" that is relevant to the question of which choice you should make. Many versions of expected utility theory identify the relevant sort of goodness with *utility*, and interpret utility with a measure of how strongly one *prefers* various outcomes (at least on the assumption that one's preferences meet certain basic conditions of rational coherence). Other theories of practical reason would interpret the relevant sort of goodness in some quite different way. For present purposes, I do not need to take a stand on the question of which of these theories is correct.[11] All that matters for my purposes is that there is *some* such notion of what it is for one choice to be better or worse than (or at least as good as) another, which plays this sort of crucial role in the theory of practical reasoning.

To revert to the terminology that I used in the previous section, my suggestion is that the objectively "right" choices – the choices that achieve the ultimate "goal" of practical reasoning – would be the choices that are in this way *optimal in the state of nature that actually obtains*. This is determined purely by the way in which one's chosen option compares with the available alternatives in the state of nature that actually obtains. It is quite irrelevant how one's chosen option compares with the available options in *other* states of nature. The only comparisons that are relevant for determining which are the objectively "right" options to choose are comparisons of how good the available options are *within* the state of nature that actually obtains.

So far, I have focused on the question of which options are the objectively "right" ones to choose. But it seems plausible that a similar point applies to the question of which choice it is *rational* to make, when one is not certain which state of nature one is actually in. As I suggested above, to determine which options are the objectively right ones to choose, one simply *does not need to know* whether one is in a nice state of nature or a nasty state of nature (given that it is completely beyond one's control which of these states of nature one is actually in). In a similar way, I suggest, to make a rational choice when one is not certain which state of nature one is in, one *does not need to consider* whether one is in is a nice state of nature or a nasty one. All that one needs to consider are the *degrees* to which each of the available options is better (or worse) than the available alternatives *within* each of the relevant states of nature. Admittedly, when one is uncertain which state of nature that one is in, one must make certain comparisons across the relevant states of nature. But perhaps the only relevant comparisons are comparisons of the *differences* in levels of goodness between the various options *within* each of these states of nature with the *differences* between those options within each of the other states of nature – not any comparisons of the *absolute* levels of goodness that those options have in one state of nature with the absolute levels of goodness that they have in any other state of nature.

If this suggestion is correct, then it provides a new diagnosis of what exactly is irrational about one-boxing in the original Newcomb problem. The classic version of EDT responds to this case in the following way: "If you choose only the opaque box, almost certainly Alice will have predicted that, and so almost certainly there is $1,000,000 in the opaque box, and so your gain will almost certainly be $1,000,000. On the other hand, if you choose both boxes, then almost certainly Alice will have predicted that, and so almost certainly there is nothing in the opaque box, and so your gain will almost certainly be just $1,000. Clearly, $1,000,000 is better than $1,000; so you should choose only the opaque box."

In comparing the outcome of getting $1,000,000 with the outcome of getting $1,000, one is comparing the *absolute* level of goodness or desirability of one-boxing in *one* state of nature with the *absolute* level of goodness of two-boxing in *another* state of nature. But according to our new suggested diagnosis, this comparison is in fact irrelevant. The only relevant comparison across the two states of nature is a comparison of the *differences* between the two options *within* each state of nature – that is, the comparison of the difference between $1,000,000 and $1,001,000 in the first state of nature with the difference between $0 and $1,000 in the other state of nature. In this case, the result of this comparison is that in *both* of these two states of nature, the payoff of two-boxing is $1,000 more than that of one-boxing. According to my Gandalf-inspired approach, this is the *only* relevant comparison across these states of nature.

If the agent is swayed by the essentially irrelevant comparison between the absolute payoff of one option in one state of nature with the absolute payoff of an alternative option in another state of nature, then she will in effect be acting in a way that will give her *good news*. That is, she will be acting in a way that indicates that she is in a *nice* state of nature (such as a state of nature where there is $1,000,000 in front of her that is hers for the taking) rather than a *nasty* one (such as a state of nature where there is no $1,000,000 that is hers for the taking). But by the very definition of states of nature, there is nothing that she can do to affect which state of nature she is actually in. So being swayed by this irrelevant comparison is indeed, as Lewis (1981, 5) complained, “an irrational policy of massaging the news”.

It is a striking fact that CDT already implicitly respects this point. Comparisons between the *absolute* levels of utility across *different* states of nature in fact have no effect within CDT on which options it is rational to choose. The only comparisons across states of nature that have any effect on which option it is rational to choose according to CDT are comparisons of the *differences* between the available options *within* each state of nature.[12] It is this feature of CDT (and *not* its focus on unconditional as opposed to conditional probabilities) that I shall incorporate into the approach that I shall articulate here.

4. Evidentially expected comparative value

If the foregoing suggestions are along the right lines, then we need a theory of rational choice according to which a rational choice is one that maximizes the *evidential expectation* of some purely *comparative* measure of how the chosen option compares to other options *within* each state of nature. This purely comparative measure must be insensitive to any comparisons of the *absolute* values of options *across* different states of nature: it must reflect nothing but the options’ *comparative values* – that is, the *differences* in value between these options – *within* each of the relevant states of nature. An appropriate measure of this kind, if it can be found, can be called a measure of the options' "comparative value". We could then say that a rational choice then is one that *maximizes evidentially expected comparative value*.

There is a fairly natural measure of "comparative value" to use in the case where there are just *two* options that one is choosing between. In this section, I shall articulate a way of dealing with the two-option case. In Section 5, I shall explain why generalizing this approach to the many-option case is rather less straightforward than one might have hoped. In Section 6, I shall canvas a couple of ways of generalizing this approach to the many-option case.

Let <A_1, A_2> be a pair of acts each of which the agent rationally believes to be available to her. Let S_1, … S_n be an (exhaustive and exclusive) partition of "states of nature". As I explained above, "states of nature" are states of affairs that the agent rationally believes to be entirely beyond her control: that is, the agent is convinced that it lies altogether beyond her power to have any effect on which of these states of nature she is actually in.

Let us assume that for each of these acts A_i, and each of these states of nature S_j, S_j has a certain "probability" conditional on the agent's choosing A_i; let us abbreviate this conditional probability as '$pr(S_j \mid A_i)$'. The agent may not be certain which of these states of nature she is actually in, but for every act A_i and every state of nature S_j, the agent *is* certain what "value" or "degree of goodness" A_i has in S_j.

To fix ideas, let us assume that the "probability" of S_j conditional on the agent's doing A_i is just the conditional credence that it is rational for the agent to place in S_j on the assumption that she will choose A_i. Let us assume that the "value" or "degree of goodness" of A_i in S_j is the degree to which A_i is a good thing for the agent to do in S_j. This is simply determined by the content of S_j. In effect, S_j must somehow entail a conjunctive proposition of the form 'A_1 is good to degree d_1 & A_2 is good to degree d_2', for real numbers d_1, d_2.

We can define each of these acts' "comparative value" (in relation to the other act) in S_j in the following way. Suppose that the value of A_1 in S_j is d_1, and the value of A_2 in S_j is d_2. Take the *average* of d_1 and d_2, which I shall call the "benchmark", *b*. (As we shall see in the next two sections, we may sometimes wish to take a different "benchmark" from this. But it seems natural to start with the "neutral" setting – exactly at the midpoint between the two acts A_1 and A_2.) Then the comparative value of A_1 (in relation to A_2) in S_j is $d_1 - b$, and the comparative value of A_2 (in relation to A_1) in S_j is $d_2 - b$. (So, if $d_1 > d_2$, then the comparative value of A_1 in relation to A_2 in S_j will be a positive number, and the comparative value of A_2 in relation to A_1 in S_j will be a negative number.) In general, let us abbreviate the comparative value of A_i (in relation to its alternative A_k) in S_j as '$cv\ (A_i, A_k, S_j)$'.[13]

Then, for any act A_i that belongs to this pair of acts $<A_1, A_2>$, the "evidentially expected comparative value" of act A_i (in relation to the other act A_k) is the sum of A_i's "comparative value" (in relation to the other act) in each of these states of nature S_j, weighted by the probability of S_j conditional on the agent's choosing A_i. That is, the evidentially expected comparative value of A_i in relation to its alternative A_k can be expressed as follows:

$$\sum_j cv\ (A_i, A_k, S_j)\, pr(S_j \mid A_i)$$

Then we can say that it is rational to choose an act A_i, in preference to its alternative A_k, just in case the evidentially expected comparative value of A_i (in relation to A_k) is *no less* than the evidentially expected comparative value of A_k (in relation to A_i).

The following tables show how this approach deals with Newcomb's problem (for simplicity, we shall again assume that each option's degree of goodness in each state of nature is exactly proportional to its monetary payoff).

Payoffs

	$1 million in opaque box; $1,000 in transparent box	$0 in opaque box; $1,000 in transparent box
You choose opaque box	$ 1,000,000	$ 0
You choose both boxes	$ 1,001,000	$ 1,000

Comparative Values in the Possible States of Nature

	$1 million in opaque box; $1,000 in transparent box	$0 in opaque box; $1,000 in transparent box
You choose opaque box	– $ 500	– $ 500
You choose both boxes	+ $ 500	+ $ 500

Evidential Probabilities

	$1 million in opaque box; $1,000 in transparent box	$0 in opaque box; $1,000 in transparent box
You choose opaque box	1	0
You choose both boxes	0	1

Evidentially Expected Comparative Values

	$1 million in opaque box; $1,000 in transparent box	$0 in opaque box; $1,000 in transparent box	**Total**
You choose opaque box	– $ 500	$ 0	– $ 500
You choose both boxes	$ 0	+ $ 500	+ $ 500

So, according to this approach, you should choose both boxes. In the original Newcomb problem, then, this approach agrees with CDT. In general, it is easy to see that in any case in which there is a *dominant* option – in the sense that this option is optimal in *every* state of nature – then the approach that I am suggesting implies that such dominant options will be the only options that it is rational to choose. Moreover, it is also easy to see that this approach will also agree with CDT in all the "normal" cases, where all the various relevant states of nature are probabilistically independent of one's choice. This approach will only differ from CDT in cases where there is no dominant option, and the relevant states of nature are *not* probabilistically independent of one's choice. One such case is Andy Egan's Psychopath Button case, which we discussed in Section 1. The way in which my approach deals with this case can be illustrated by means of the following tables.

Payoffs

	You are a psychopath	You are not a psychopath
You press the button and kill all psychopaths	– 100	+ 10
You do not kill all psychopaths	0	0

Comparative Values in the Possible States of Nature

	You are a psychopath	You are not a psychopath
You press the button and kill all psychopaths	– 50	+ 5
You do not kill all psychopaths	+ 50	– 5

Evidential Probabilities

	You are a psychopath	You are not a psychopath
You press the button and kill all psychopaths	1	0
You do not kill all psychopaths	0	1

Evidentially Expected Comparative Values

	You are a psychopath	You are not a psychopath	**Total**
You press the button and kill all psychopaths	– 50	0	– 50
You do not kill all psychopaths	0	– 5	– 5

So, in this case, the approach that I am suggesting here implies that the rational choice is to choose *not* to press the button – which again seems intuitively the right result.

5. The problem of the many-option cases

Unfortunately, it is not as straightforward as one might have hoped to generalize this approach to the *many*-option case. The natural way to extend this approach to the many-option case would be by taking the "benchmark" b in each state of nature S_j as the average degree of goodness in S_j of *all* the available options; then we can use this benchmark b to define the "overall comparative value" of *all* the available options in S_j, by simply identifying the "overall comparative value" of any option A_i in S_j with $d_i - b$. Then we could say that it is rational to choose any option that has maximal evidentially expected "overall comparative value" of this kind. But if we extend the approach to the many-option case in this way, then the resulting theory of rational choice turns out to have a troubling defect.

The problem is that, in fact, in every situation in which we have to make a choice, there is an enormous number of perfectly dreadful courses of action that are at least physically (if not psychologically) “available”. For example, you could commit a serious crime in a way that is certain to result in your being arrested; you could give every penny that you have to the politician whom you most despise; you could commit suicide in various agonizing ways, using various items that you keep in the cupboards and drawers of your kitchen; and so on. Of course, we hardly ever actually think of doing such insane things, but in principle they are available. It does not seem that the bare fact that these courses of action are available should make a difference to what it is rational to choose.

If we include all these insane courses of action among the set of options that you are choosing between, then the “benchmark” within each state of nature will be a lot lower than it would otherwise be. These lower benchmarks would often have the effect of leading to the opposite decision between a pair of options *A* and *B* from the decision that one would make if the benchmarks for deciding between them were the average degree of goodness of *A* and *B* in each of the relevant states of nature. We can illustrate this point with the following variant of Newcomb’s problem.

Suppose that Alice the predictor confronts you with two boxes, Left and Right. You must choose exactly *one* of these two boxes. Alice has predicted which you will choose. If she predicted that you would choose the Left box, she put $100 in the Left box, and $120 in the Right box. If she predicted that you would choose the Right box, she put $500 in the Right box, and $1,000 in the Left box. Then the situation is as follows:

Payoffs

	$100 in Left box; $120 in Right box	$1,000 in Left box; $500 in Right box
You choose Left box	$ 100	$ 1,000
You choose Right box	$ 120	$500

Comparative Values in the Possible States of Nature

	$100 in Left box; $120 in Right box	$1,000 in Left box; $500 in Right box
You choose Left box	– $ 10	+ $ 250
You choose Right box	+ $ 10	–$ 250

Evidential Probabilities

	$100 in Left box; $120 in Right box	$1,000 in Left box; $500 in Right box
You choose Left box	1	0
You choose Right box	0	1

Evidentially Expected Comparative Values

	$100 in Left box; $120 in Right box	$1,000 in Left box; $500 in Right box	**Total**
You choose Left box	– $ 10	$ 0	– $ 10
You choose Right box	$ 0	–$ 250	–$ 250

For this case, the rule proposed in the previous section implies that the rational choice for you to make is to choose the Left box, because then it is conditionally certain, on the assumption that you make this choice, that this choice will only get you $20 less than the maximum that you could have won – whereas on the assumption that you choose the Right box, it is conditionally certain that choosing the Right box is *atrociously* suboptimal, giving you a full $500 less than the maximum you could have won. (I find this an intuitively acceptable verdict about this case.

Indeed, the choice to pick the Right box would strike me as an irrational choice to "massage the news" in precisely the way that Lewis stigmatizes. However, I should report that many of the other philosophers whose intuitions I have canvassed disagree with me about this case.)

However, now suppose that we switch to the measure of "overall comparative value" of the sort that I outlined at the beginning of this section. Since there are so many insane options available, it will be a long and tedious business to give an exact calculation of the "overall" comparative values of these options in each of the possible states of nature. So instead I shall just stipulate that the "overall" comparative values and the evidential probabilities are as follows:

"Overall" Comparative Values in the Possible States of Nature

	$100 in Left box; $120 in Right box	$1,000 in Left box; $500 in Right box
You choose Left box	+ $ 100,099	+ $ 101,000
You choose Right box	+ $ 100,119	+ $ 100,500
Numerous utterly insane options …	– $ 100,000	– $ 100,000

Evidential probabilities

	$100 in Left box; $120 in Right box	$1,000 in Left box; $500 in Right box
You choose Left box	1	0
You choose Right box	0	1
Numerous utterly insane options …	0.5	0.5

Evidentially Expected "Overall" Comparative Values

	$100 in Left box; $120 in Right box	$1,000 in Left box; $500 in Right box	**Total**
You choose Left box	+ $ 100,099	$ 0	+ $ 100,099
You choose Right box	$ 0	+ $ 100,500	+ $ 100,500
Numerous utterly insane options …	– $ 50,000	– $ 50,000	– $ 100,000

Here, however, the rule of maximizing evidentially expected "overall" comparative value implies that the rational choice is to choose the Right box – reversing the decision between Left and Right that was recommended by the rule that I outlined in Section 4.

I suggested above that the bare fact that these utterly insane courses of action are available should not make a difference to what it is rational to choose. Any theory according to which it does make a difference suffers from a significant defect. But as we have just seen, the natural way of extending the approach that I proposed in the previous section to the many-option case suffers from this defect.

So we need to find a different way of extending the approach that I proposed for the two-option case in Section 4 so that it can handle the many-option case in a satisfactory way. I shall canvas some of these ways of extending this approach in the next section.

6. A "reasonable" benchmark

As we have seen, my approach needs to identify a "benchmark" for each state of nature in order to measure the comparative value of the available options within that state of nature. There are admittedly *some* situations where it will not matter exactly where this "benchmark" lies. For example, in all the "normal" situations, where the states of nature are probabilistically quite independent of one's choice, my approach will simply coincide with CDT irrespective of where exactly the benchmark lies. Moreover, in any two-option case in which there is a "dominant" option (as in the original Newcomb problem), it will also make no difference where the benchmark is: wherever this benchmark may be, in such situations, all and only the dominant options will be rational. However, there will also be other situations – such as Egan's (2007) Psychopath Button case – where it does make a difference where this benchmark lies.

One fairly extreme benchmark is given by the *optimal* value in that state of nature. Measuring the comparative value of the options in terms of how far they *fall short* from the optimal value in the state of nature gives us the measure that has become known as "*regret*".[14] If we took this benchmark, we could say that a rational choice is one that *minimizes evidentially expected regret*. Another extreme benchmark is given by the *lowest* value in the relevant state of nature. Measuring the comparative value of the options in terms of how far they exceed this lowest value would give us a measure that we could call "*relief*". Taking this as our benchmark, we could say that a rational choice is a choice that *maximizes evidentially expected relief*.

Both of these two views of rational choice would be rather extreme. According to the rule of maximizing evidentially expected relief, in every state of nature, the options that are better than the lowest-ranked options are "marked up" for the extent to which they are better than the lowest-ranked options, but the lowest-ranked options are not in any way "marked down" for the extent to which they are worse than the other options. So consider any case that – like Egan's Psychopath Button case – has the following features. First, there are only two options, *A* and *B* (for example, suppose that *A* is not pressing the button and *B* is pressing the button). Second, on the assumption that you choose *A*, it is virtually certain that *A* is the lowest-ranked option and that option *B* is *slightly* better than *A* – whereas, on the assumption that you choose *B*, it is virtually certain that *B* is the lowest-ranked option but that option *A* is *dramatically* better than *B*. Then the rule of maximizing evidentially expected relief would rate these two choices as equally rational – thereby overlooking an intuitively obvious reason for favouring option *A* over *B*.

On the other hand, according to the rule of minimizing evidentially expected *regret*, the suboptimal options in every state of nature are "marked down" for the extent to which they are inferior to the optimal options, but the optimal options are in any way not "marked up" for the extent to which they are better than those suboptimal alternatives. For example, consider the following case: on the assumption that one chooses *A*, it is to all intents and purposes conditionally certain that *A* will be optimal, and that *B* will be *very slightly* less good than *A* – whereas on the assumption that one chooses *B*, it is conditionally certain that *B* will be optimal, and that *A* will be *dramatically* less good than *B*. The rule of minimizing evidentially expected regret would treat these two choices as equally rational; but this seems to overlook an intuitively obvious reason for favouring *B* over *A*.

This was why in giving my preferred measure of "comparative value" in Section 4, I avoided both of these two extreme benchmarks, and fixed on the benchmark that was *exactly half-way* between the values that the two options in question have in the relevant state of nature. This suggests that the best way to extend this approach to the many-option case is to try to find what we could call the "reasonable benchmark". For the rest of this section, I shall briefly canvas two suggestions about what exactly this "reasonable benchmark" might be.

The problem that we considered in Section 5 crucially involved focusing on all of the perfectly insane and terrible options that are in fact available at all times, but which we usually never even consider. One natural way to avoid this problem is to insist on excluding from consideration all the options that one is virtually certain that one will never choose. There is no point in seriously considering an option if one already knows perfectly well that there is no chance that one will ever choose that option. We should also insist on excluding from consideration all those options that are (at least weakly) dominated by some other option. If one option *A* is dominated by another option *B* – in the sense that *B* is better than *A* in some state of nature of nature and at least as good as *A* in *every* state of nature – then the dominated option *A* should not be taken seriously in practical reasoning, and so should be simply excluded from consideration altogether.[15]

This point suggests one relatively simple way of extending the conception of rational decision that I proposed in Section 4 from the two-option case to the many-option case. First, we should simply exclude all the options that should not be taken seriously in practical reasoning; then we could simply say that the "reasonable benchmark" for each state of nature is the average value in that state of nature of the options that have not been excluded in this way.

We might also consider a second slightly more complicated way of extending this conception of rational decision from the two-option case to the many-option case. This more complicated approach would involve a more *dynamic* conception of the "reasonable" benchmark. First, one would start by excluding from consideration all the options that are not worth taking seriously for the reasons that I have already sketched. Then one would consider the *range of admissible values* that the benchmark could take for each state of nature – where an admissible value for the benchmark is the average value of any (proper or improper) subset of the options that are still being taken seriously. If there is any option that has suboptimal evidentially expected comparative value *whichever* of these admissible benchmarks is used, that option can also be excluded from consideration; then the whole process can be run again on the yet more restricted set of options. This process can be repeated any number of times until no more options can be excluded in this way. Then we can say that the remaining options are all options that it is rational to choose.

Clearly there could be yet other variants of these two ways of extending this approach from the two-option case to the many-option case. Here I have simply canvassed these two ways in order to make it plausible that the problem may not be insoluble. However, it will certainly require much further investigation to determine which is the best version of this approach.[16]

7. “Independence of Irrelevant Alternatives”

There is a fundamental objection that many decision theorists will raise against this approach – namely, that this approach conflicts with the principle that is sometimes called as the “independence of irrelevant alternatives”.[17] In this final section, I shall explain why it really is inevitable that this approach will collide with the most popular way of understanding this principle. Then I shall argue that there is no reason at all for a proponent of my approach to accept this principle. At most, there is a reason to accept a superficially similar though in fact quite different principle. Fortunately, as I shall explain, my approach is quite consistent with this second principle.

According to one common understanding of it, the principle of the “independence of irrelevant alternatives” rules out the following possibility. Suppose that there are two choice situations CS_1 and CS_2 such that in CS_1 there are three available options *A*, *B*, and *C*, while CS_2

is simply a “contracted” version of CS_1 – that is, in CS_2 there are only two available options, *A* and *B*, but otherwise CS_1 and CS_2 are as similar to each other in every respect as it is possible for them to be. Then according to the “independence of irrelevant alternatives”, it is impossible for it to be simultaneously the case that the rational choice in CS_1 is *A* and the rational choice in CS_2 is *B*. The ranking of *A* and *B* relative to each other must be the same in both CS_1 and CS_2 because it is “independent” of the third “irrelevant” alternative *C*.

It is easy to show that my approach is bound to conflict with this principle. This is because we can always find a choice situation involving three options *A*, *B*, and *C*, that has the following feature: in the most similar “contracted” choice situation involving just *A* and *B*, my approach will require choosing *A*; in the most similar “contracted” situation involving just *B* and *C*, my approach will require choosing *B*; and in the most similar “contracted” situation involving just *A* and *C*, my approach will require choosing *C*. So, wherever we put the benchmark in this three-option choice situation that involves all three options *A*, *B*, and *C*, one of the rankings of these options in the “contracted” choice situations (*A* over *B*, *B* over *C*, and *C* over *A*) will be “overturned” in the three-option situation – producing a violation of this version of the “independence of irrelevant alternatives”.

Let me give an example of the sort of case that I have mind. Suppose that you can pick one of three Boxes – Box A, Box B, and Box C. There are two possible states of nature: in the first state of nature S_1, there is nothing in Boxes A and C, and \$2,000 in Box B; in the second state of nature S_2, there is \$900 in Box A, nothing in Box B, and \$1800 in Box C. Your choosing either Box A or Box B is strong (but not overwhelming) evidence that you are in state of nature S_2, while your choosing Box C is strong (but not overwhelming) evidence that you are in state of nature S_1.

The following tables show what follows about this case if the benchmark in all two-option choice situations is simply the "neutral" setting, exactly at the mid-point between the two options in each state of nature. (If the benchmark in two-option situations is not always at the mid-point, we might have to fiddle with the numbers, but it would still be possible to construct a case of this sort.)

Evidential Probabilities

	Box A – $0 Box B – $2,000 Box C – $0	Box A – $900 Box B – $0 Box C – $1,800
Box A	0.1	0.9
Box B	0.1	0.9
Box C	0.9	0.1

The "Contracted" Choice Situations: Comparative Values in the States of Nature

Box A	− $1,000	+ $450
Box B	+ $1,000	− $450

Box B	+ $1,000	− $900
Box C	− $1,000	+ $900

Box A	$0	− $450
Box C	$0	+ $450

The "Contracted" Choice Situations: Evidentially Expected Comparative Values

			Total
Box A	– $100	+ $405	+ $305
Box B	+ $100	– $405	– $305

Box B	+ $100	– $810	– $710
Box C	– $900	+ $90	– $810

Box A	$0	– $405	– $405
Box C	$0	+ $45	+ $45

As these tables show, in the "contracted" choice situation involving just *A* and *B*, my approach requires choosing *A*, while in the "contracted" situation involving just *B* and *C*, it requires choosing *B*, and in the "contracted" situation involving just *A* and *C*, it requires choosing *C*. So whatever my approach says about the "expanded" choice situation involving *A*, *B*, and *C*, it will end up violating the independence of irrelevant alternatives.

I submit that this result casts no doubt whatsoever on my approach. The "contracted" choice situation is simply a *different* situation from the "expanded" choice situation. It is logically impossible for any agent to be in both situations at the same time. It cannot be the case both that the only options available to you are *A* and *B*, and also that your available options include a third option *C* as well. The central idea of my approach is precisely that the crucial factor in rational decision making is the way in which all the available options compare with each other within each state of nature, and so my approach will obviously accept that two choice situations in which *different* options are available will be crucially different from each other – at least so long as all of those options are ones that deserve to be taken seriously by a rational deliberator.[18] For this reason, the ranking of the options in terms of their evidentially expected comparative value was always meant to be situation-relative – that is, relative to the choice situation of the relevant agent at the relevant time. For the purposes of a theory of rational choice, all that we need is a ranking of options that is situation-relative in this way. In every choice situation that one might be in, one can be guided by a ranking of options that is relative to that choice-situation. Rankings of options that are not relative to particular choice-situations in this way are not necessary for guiding the rational choices of deliberating agents.

Now there is a second principle that may sound somewhat similar to this version of the independence of irrelevant alternatives, although it is in fact a fundamentally different principle. This second principle does not try to link different choice situations with each other: it links pieces of "partial" practical reasoning that focus on different *subsets* of the available options within the *same* choice situation. According to this second principle, a rational piece of "partial" reasoning that focuses on a proper subset of the available options should rank these options in the same way as a rational piece of "total" reasoning that considers all of the available options.

However, it is clear that my approach is consistent with this second principle. The only reason why my approach collided with the first version of the "independence of irrelevant alternatives" was because the benchmark for each state of nature can shift between one choice situation and another. But according to my approach, the benchmark for each state of nature is fixed by the relevant choice situation; *within* each choice situation, the benchmarks are fixed. If the benchmarks are fixed, then the comparative values of each option within each state of nature are also fixed; and so long as the conditional probabilities also do not shift, my approach will generate the same ranking of a set of options – regardless of whether or not they are all of the available options or only a proper subset of them. So my approach is quite compatible with this second principle.

In conclusion: It is clear that much more investigation will be required to determined whether or not the approach sketched here is really correct. Above all, we should find out whether we can develop anything that plays the role in this approach that the “representation theorems” play in the traditional theories of rational choice such as CDT and EDT.[19] All that I have aimed to do here is to articulate a new conception of rational choice, and to explore it they will lead. In particular, as I have argued, this conception seem to have the advantage that while it is relatively conservative, and agrees with CDT both about all cases where the states of nature of probabilistically independent of one’s choice, and about all cases in which there is a dominant option, it can also accommodate the intuitively compelling counterexamples to CDT that have recently been articulated by Egan (2007).

Notes

1. The leading exponent of EDT is Jeffrey (1983); the most distinguished recent exponent of CDT is Joyce (1999).

2. This feature of EDT is rightly stressed by Joyce (1999, 121–2).

3. I take the term ‘state of nature’ from Broome (1991, 22).

4. For an argument that all these ways of presenting causal decision theory are essentially equivalent, see Lewis (1981).

5. Other attempts might be made to undermine the force of our intuitive judgment about

this case. For example, someone might base a case against this intuitive judgment on what Eells (1982, 170–4) says about the "tickle defence". I believe that these other attempts to defend CDT against this counterexample would also fail, although unfortunately I cannot argue for this point in detail here.

6. There may be such paradoxical situations, such as the famous *Death in Damascus* case discussed by Gibbard and Harper (1978). My point is just that the *Psychopath Button* case does not seem to be a paradoxical case of this kind.

7. For a defence of this idea, see Wedgwood (2002).

8. I owe this example to Krister Bykvist and Anandi Hattiangadi (2008).

9. A rival explanation would appeal, not to the *conditional probability* of *p*, but to the *probability of a conditional* – such as the counterfactual conditional 'If I were to believe *p*, then *p* would be true'. This rival explanation would be hard to reconcile with the *probabilistic* assumption that the degrees of credence that it is rational for you to have at any one time must be capable of being modelled by a probability function. The explanation that I proposed can be adapted so that it says that at any one time you should have a set of credences such that your credence in any proposition *p* is the conditional probability of *p* on the assumption that you adopt this set of credences. But we cannot adapt the rival explanation to accommodate probabilism in the same way: as Lewis (1976) famously showed, the probabilities of a set of conditionals 'If *S* were the case, then φ' are not guaranteed to form a probability distribution over the *consequents* of these conditionals. So this seems to give us a reason to favour the explanation according to which the formation and revision of one beliefs in *p* should be guided by the conditional probability of *p* (rather than by the probability of some conditional that has *p* as its consequent).

10. It may even be reasonable for us to regard this as a condition on the options' being genuinely "available" in the relevant sense. (In other words, it may be reasonable to assume that agents must always choose between courses of action that are *available to choice* in this way.) However, it seems clear that for some other theoretical purposes it will be important that for many decisions (indeed, perhaps for *all* decisions), there is a non-zero probability that (as a result of one's "hand trembling" or the like) one will somehow fail to execute one's decision.

11. For my views on this question, see Wedgwood (2003).

12. One way to see that CDT has this feature is by reflecting on the following two points. First, there is a well-known structural analogy between CDT and *utilitarianism* in cases where the total population is fixed: the different states of nature correspond to subgroups of the population; and the probability of each state of nature corresponds to the size of each subgroup as a proportion of the total population. (EDT, by contrast, is more like utilitarianism in cases where one's actions can affect the total population, since in EDT the states of nature may have different probabilities depending on which option is being evaluated.) Secondly, as Amartya Sen (1973) pointed out, in cases where population is fixed, utilitarianism does not need to make any interpersonal comparisons of *absolute* levels of welfare, but only interpersonal comparisons of welfare *differences* for individuals between the different outcomes. As Sen (1973, 44) put it: "the utilitarian formula declares *x* to be socially preferred to *y* if and only if the welfare differences for all the individuals between *x* and *y* summed together turn out to be positive, and in defining differences the question of the 'origin' of the utility function is irrelevant." In the same way, CDT implies that *x* is to be preferred to *y* if and only if the probability-weighted sum of the utility differences between *x* and *y* in each of the relevant states of nature is positive. So the position of the utility function's "origin" or zero-point is irrelevant for CDT (it can even be a

different zero-point for each of the different states of nature).

13. In effect, this comes out as equivalent to the “regret / rejoice” measure that was proposed by Loomes and Sugden (1982), according to which the “regret / rejoice” measure of A_1 (in relation to A_2) in the state of nature S_j is $d_1 - d_2$, while the “regret / rejoice” measure of A_2 (in relation to A_1) in S_j is $d_2 - d_1$. But in view of the discussion that follows in Sections 5 and 6, it is useful to introduction the idea of a “benchmark” at this point.

14. This notion of “regret” was first introduced by Loomes and Sugden (1982). What I am here calling “relief” has been called “rejoicing” or “jubilation” by others, but the term ‘relief’ seems more appropriate to me.

15. I am indebted to Rachael Briggs and Brian Hedden for pointing out to me that the options that are to be taken seriously must exclude all (strongly or weakly) dominated options. But it does not seem to me *ad hoc* to exclude dominated options in this way. On my approach, if option *A* is dominated by option *B*, *A* will lose any pair-wise comparison with *B* in a particularly decisive way – since *A* will lose any pair-wise comparison with *B* irrespective of where the benchmark lies, and irrespective of the probabilities involved (at least so long as these probabilities are all non-zero). So on my approach it is natural to insist that all dominated options should simply be excluded from consideration.

16. In a fascinating unpublished paper, Rachael Briggs has developed an objection to my approach, based on an analogy between the theory of individual rational choice and social choice theory. If the relevant probabilities are conditional probabilities, of the kind that I have appealed to here, then we can view each option as analogous both to a *voter* and to a *candidate* in a voting system. (For example, if on the assumption that one chooses *A*, *B* has higher expected utility that *A*, we can say that *A* “votes for” *B*.) Then an analogue the famous impossibility theorem of

Arrow (1963) will apply. However, I do not think that we should be troubled by this. In this context, there is simply no rationale for the analogue of the Pareto principle. On my approach, all that is relevant is how strongly and favourably each option "votes for" *itself*. To put it picturesquely, my approach is not modelled on a procedural democracy, in which the winning candidate is the one that gains the most votes, but on a deliberative process, in which the winning candidates are those who argue most persuasively on their own behalf.

17. For a discussion of this principle, see especially Sen (1993) – although Sen calls this principle "basic contraction consistency (Property α)". I am fundamentally in complete agreement with Sen's argument that there is no reason to believe that this principle is an *a priori* requirement of "internal consistency of choice".

18. Of course, I have argued in Section 5 that the availability of "insane" options that should not be taken seriously by a rational deliberator should *not* make a difference to which options it is rational to choose. But this is compatible with saying that the availability of "sane" options, of the sort that should be taken seriously by a rational deliberator, *should* make a difference to which options it is rational to choose.

19. For a discussion of these representation theorems, see especially Joyce (1999, 78–82). As a matter of fact, I am inclined to favour a quite different approach from the approach that is grounded on these representation theorems; but I cannot take the time to explain why I have this inclination here.

References

Arrow, Kenneth (1963). *Social Choice and Individual Values*, 2nd edition (New Haven: Yale University Press).

Broome, John (1991). *Weighing Goods* (Oxford: Blackwell).

Bykvist, Krister, and Hattiangadi, Anandi (2007). “Does Thought Imply Ought?”, *Analysis* 67, no. 4: 277–85.

Eells, Ellery (1982). *Rational Decision and Causality* (Cambridge: Cambridge University Press).

Egan, Andy (2007). “Some Counterexamples to Causal Decision Theory”, *Philosophical Review* 116: 93–114.

Gibbard, Allan, and Harper, William (1978). “Counterfactuals and Two Kinds of Expected Utility”, in C. A. Hooker et al., eds., *Foundations and Applications of Decision Theory* (Dordrecht: Reidel), vol. 1: 125–62.

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

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

Lewis, David (1976). “Probabilities of conditionals and conditional probabilities”, *Philosophical Review* 85: 297–315. Reprinted in Lewis (1985, 133–152).

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

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

Loomes, Graham, and Sugden, Robert (1982). “Regret Theory: An Alternative Theory of Rational Choice under Uncertainty”, *Economic Theory* 92: 805–824.

Nozick, Robert (1969). “Newcomb’s Problem and Two Principles of Choice”, in *Essays in Honor of Carl G. Hempel*, ed. Nicholas Rescher (Dordrecht: Reidel): 107–33.

Sen, Amartya (1973).*On Economic Inequality* (Oxford: Oxford University Press).

———— (1993). “Internal consistency of choice”, *Econometrica* 61, no. 3: 495–521.

Tolkien, J. R. R. (1954). *The Fellowship of the Ring: being the first part of The Lord of the Rings*, original edition (London: Allen & Unwin).

Wedgwood, Ralph (2002). “The Aim of Belief”, *Philosophical Perspectives* 16 (2002): 267–297.

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