The Evidentialist's Wager[1]

William MacAskill, Aron Vallinder, Caspar Oesterheld, Carl Shulman, Johannes Treutlein

Abstract

Suppose that an altruistic agent who is uncertain between evidential and causal decision theory finds herself in a situation where these theories give conflicting verdicts. We argue that even if she has significantly higher credence in CDT, she should nevertheless act in accordance with EDT. First, we claim that the appropriate response to normative uncertainty is to hedge one's bets. That is, if the stakes are much higher on one theory than another, and the credences you assign to each of these theories aren't very different, then it's appropriate to choose the option which performs best on the high-stakes theory. Second, we show that, given the assumption of altruism, the existence of correlated decision-makers will increase the stakes for EDT but leave the stakes for CDT unaffected. Together these two claims imply that whenever there are sufficiently many correlated agents, the appropriate response is to act in accordance with EDT.

1 Introduction

Suppose you find yourself in the following decision situation:

[1] For discussion and feedback, we are grateful to Christian Tarsney, Hayden Wilkinson, and an anonymous referee.

Moral Newcomb

> In front of you are two boxes, A and B. You can choose either B only, or both A and B. Box A is guaranteed to contain one dose of a cure for a fatal disease, whereas box B may or may not contain ten such doses. A highly reliable predictor has made a prediction about your decision. If she predicted that you would take both boxes, she left box B empty. If she predicted that you would take box B only, she put ten doses in that box. What should you do?

Cases like this give rise to the well-known debate between causal decision theory (CDT) and evidential decision theory (EDT). Roughly speaking, according to CDT, you should perform the action which is likely to *cause* a good outcome, whereas according to EDT, you should perform the action which *provides strong evidence* that a good outcome will occur.[2]

And let's suppose—as would seem to be natural—that you represent the decision situation as follows:

[2] More precisely, CDT tells you to perform the action with the highest *causal expected value* (CEV), where the CEV of an action A is the sum product of the value of each outcome and the probability that A causes that outcome. By contrast, EDT tells you to perform the action with the highest *evidential expected value* (EEV), where the EEV of an action A is the sum product of the value of each outcome and the probability of that outcome on the assumption that A is performed.

	Cure in both	Cure in one only
Take one box	Ten lives	Nothing
Take both boxes	Eleven lives	One life

Given that the prediction has already been made, your decision has no influence on whether or not box B contains anything. Therefore, CDT recommends that you choose both boxes, because that way you are guaranteed to obtain one more dose than you otherwise would, regardless of whether or not B is empty. By contrast, EDT recommends that you choose only choose box B, because you would thereby receive strong evidence that you will obtain ten doses of the cure. If you instead choose both boxes, you would thereby receive strong evidence that you will merely obtain one dose. CDT has garnered more adherents (Bourget and Chalmers, 2014), and hence we assume most decision theorists believe that you should two-box in the *Moral Newcomb* problem. But here's an argument for one-boxing that, to our knowledge, has not appeared in the literature. The argument relies on three premises.

First, even though one might have higher credence in CDT than EDT, there is still an ongoing debate, and a number of intelligent and well-informed decision theorists endorse EDT. In the face of such expert disagreement, one shouldn't be anywhere near certain that CDT is correct. Instead, one should assign at least some credence to EDT. Second, once we take into account our background knowledge of the world, the stakes become significantly higher for EDT than CDT in the *Moral Newcomb*

problem. The universe is probably very big indeed, and there are very many individuals very similar to you who will face similar decision problems. As a result of this similarity, your decision in the *Moral Newcomb* problem is correlated with decisions made by many other agents elsewhere in time and space. This means that the simple state-consequence matrix above does not in fact capture everything that is relevant to the decision problem: we have to refine the state space so that it also describes whether or not correlated agents face boxes with cures in both. By taking one box, you gain evidence not only that you will obtain more doses of the cure, but also that these other agents will achieve good outcomes too. Therefore, the existence of correlated agents has the effect of increasing the stakes for EDT. By contrast, there is no causal connection between your decision and the decision of these other agents, and hence taking them into account does not affect the stakes for CDT. Third, in the face of uncertainty, the rational thing to do is to hedge one's bets. If the stakes are much higher on one hypothesis than another, and the credences you assign to each of these hypotheses aren't very different, then it's rational to choose the option which performs best on the high-stakes hypothesis.

If these three premises are true, then one-boxing is the rational thing to do in the *Moral Newcomb* problem. But, for an altruistic and morally motivated agent, there's nothing special about *Moral Newcomb* compared to other decision problems where EDT and CDT disagree. So we can generalise our conclusion as follows:

> In general, and across a wide variety of decision contexts, if you are an altruistic and morally motivated agent, and are uncertain between EDT and CDT, you should typically act in line with EDT even if you have significantly higher credence in CDT.

Call this argument *The Evidentialist's Wager*.

Should we accept the Wager? There are several steps of the argument that can be questioned. First, one could debate whether you ought to hedge in the face of decision-theoretic uncertainty. While this is not the primary focus of this article, we will give some motivation for hedging in the next section. Second, in order for the claim that the stakes are higher for EDT than for CDT to be meaningful, we need some way of making an intertheoretic value comparison — that is, some way of comparing the magnitude of an action's evidential expected value with the magnitude of its causal expected value. We present a proposal for doing this in section 3. Third, one might wonder whether we should really expect the world to contain sufficiently many correlated decision-makers facing similar decision problems. We will address this question in section 4. One way for there to be many such agents is if the universe is infinite, as is implied by several leading cosmological theories. But the infinite case gives rise to a range of further issues. These are addressed in section 5.

Although we have presented the Wager as applying to agents who are uncertain between EDT and CDT, its scope is in fact wider than this. It applies equally well to agents who are uncertain between CDT and other non-causal decision theories, such as the recently developed *functional decision theory*, or FDT.[3] For our purposes, the relevant features of EDT are (i) that it sometimes disagrees with CDT,

[3] For formulations of functional decision theory, see Soares and Levinstein (forthcoming), Yudkowsky and Soares (2017), and Greene and Levinstein (manuscript). Further arguments in favour of FDT and some related theories are given in Greene.

and (ii) that in some of those cases it is sensitive to non-causal correlations between decision makers. FDT shares both of these features, and our argument therefore provides reason for altruistic and morally motivated agents who are uncertain between CDT and FDT to follow the latter as well.

2 Arguments for Hedging

The Wager relies crucially on the idea that, in the face of decision-theoretic uncertainty, one ought to hedge. But is that correct? Perhaps what one ought to do is just what one ought to do according to the correct first-order decision theory? We're not going to be able to resolve this question in this article. But we'll show that although the view has its issues to resolve, there are some significant arguments in its favour and it should therefore at least be taken seriously. We present three such arguments.

2.1 Dominance

Consider the following variant on Newcomb's problem.

> *Almost Equal Amounts*
>
> In front of you are two boxes, A and B. You can choose either B only, or both A and B. Box A is guaranteed to contain $0.98M, whereas box B may or may not contain $1M. A 99% reliable predictor has made a prediction about your decision. If she predicted that you would take both boxes, she left box B empty. If she instead predicted that you would take box B only, she put $1M in that box.

	Money in both	Money in one only
Take one box	$1M	$0
Take both boxes	$1.98M	$0.98M

In this case, the evidential expected value of taking both boxes is the same as that of taking one box only, whereas the causal expected value of taking both boxes is $0.98M greater than that of taking one box only.[4] Suppose now that you have 90% credence in EDT and only 10% credence in CDT. From the perspective of EDT, you have nothing to lose by two-boxing, whereas from the perspective of CDT, you have $0.98M to lose by one-boxing. In short, two-boxing state-wise (or 'theory-wise') dominates one-boxing. Thus it seems clear that you should two-box, in at least some sense of 'should'.[5]

In *Almost Equal Amounts*, EDT is indifferent between the two actions and CDT prefers one over the other. But we can also construct a case where CDT is indifferent between the two actions and EDT prefers one over the other. Consider:

Empty Box

[4] EEV(A&B) = 0.01 × $1.98M + 0.99 × $0.98M = $0.99M and EEV(B) = 0.99 × $1M + 0.01 × $0 = $0.99M.

[5] For discussions of different senses of 'should' in the context of moral uncertainty, see for example MacAskill, Bykvist and Ord (2019, chapter 1) and Sepielli (2010, chapter 1).

> In front of you are two boxes, A and B. You can choose either B only, or both A and B. Box A is guaranteed to be empty, whereas box B may or may not contain $1M. A 99% reliable predictor has made a prediction about your decision. If she predicted that you would take both boxes, she left box B empty. If she instead predicted that you would take box B only, she put $1M in that box.

	Money in B	Both empty
Take one box	$1M	$0
Take both boxes	$1M	$0

In this case, the situation is reversed: the causal expected value of taking both boxes is the same as that of taking one box only, whereas the evidential expected value of taking one box only is $0.98M greater than that of taking both boxes.[6] Suppose now that you have 90% credence in CDT and only 10% credence in EDT. From the perspective of CDT, you have nothing to lose by one-boxing, whereas from the perspective of EDT, you have $0.98M to lose by two-boxing. In short, one-boxing dominates two- boxing. So it seems clear that you should one-box.

But suppose now that the right thing to do in the face of decision-theoretic uncertainty is simply to follow the recommendation of the correct (first-order) decision theory. If EDT is correct, you should be indifferent between the two actions in *Almost Equal Amounts*, and one-box in *Empty Box*. If CDT

[6] EEV(A&B) = 0.01 × $1M + 0.99 × $0 = $0.01M and EEV(B) = 0.99 × $1M + 0.01 × $0 = $0.99M.

is correct, you should two-box in *Almost Equal Amounts*, and be indifferent in *Empty Box*. Hence neither of the two theories can make sense of the intuition that one should two-box in *Almost Equal Amounts* and one-box in *Empty Box*. By contrast, if there is some notion of ought which is sensitive to all decision theories one has positive credence in, these suggestions make perfectly good sense.

2.2 Stakes Sensitivity

MacAskill (2016) argues that our intuitions in Newcomb cases are sensitive to the relative stakes for EDT and CDT, and that this stakes sensitivity can be explained by the fact that one should hedge under decision-theoretic uncertainty.[7,8] He presents the following two cases (slightly altered here) in support of this conclusion:

> *High-Stakes Predictor I (HSP-I)*
>
> Box A is opaque; Box B, transparent. If the Predictor predicted that you choose Box A only, then she put a million doses of the cure for a fatal disease in Box A. If she

[7] More specifically, MacAskill argues that this stakes sensitivity is explained by the fact that, in the face of decision-theoretic uncertainty, one should follow *Meta Decision Theory* (MDT), and choose the option with the highest *meta expected value*, where the meta expected value of an option is given as the sum, for each first-order decision theory, of the option's value according to the decision theory, weighted by one's credence in that decision theory. MDT entails, but is not entailed by, the claim that one should hedge. Nozick (1993:45) also notes this stakes sensitivity but goes on to propose that both EDT and CDT have genuine normative force, and that the weights we assign them should measure their relative normative force, rather than quantify our credences in them.

[8] The dominance argument above can be seen as a special case of such stakes sensitivity, where there's something at stake for one theory but not for the other.

predicted that you choose both Box A and Box B, then she put nothing into Box A. Box B—transparent to you—contains a stick of gum. You have two options only: Choose Box A, or choose both Box A and Box B.

	Cure in both	Cure in one only
Take one box	1,000,000 lives	Nothing
Take both boxes	1,000,000 lives + gum	Gum

In HSP-I, intuitively you should one-box, because the stakes seem much greater for EDT. For CDT, the decision between one-boxing and two-boxing is merely a choice of whether or not to get a stick of gum. By contrast, for EDT the decision is about whether or not to save a million lives. But consider now:

High-Stakes Predictor II (HSP-II)

Box C is opaque; Box D, transparent. If the Predictor predicted that you choose Box C only, then she put a million doses of the cure for a fatal disease into Box C and also a stick of gum. If she predicted that you choose both Box C and Box D, then she put nothing into Box C. Box D—transparent to you—also contains a million doses of the cure for a fatal disease. However, Box D contains no gum. One has two options only: Choose Box C only, or both Box C and Box D.

	Cure in both	Cure in one only
Take one box	1,000,000 lives + gum	Nothing
Take both boxes	2,000,000 lives + gum	1,000,000 lives

In HSP-II, intuitively you should two-box, because the stakes seem much greater for CDT. Here the situation is reversed: for EDT, the decision between one-boxing and two-boxing is merely a choice of whether or not to get a stick of gum. By contrast, for CDT the decision is about whether or not to save a million lives.

However, in both HSP-I and HSP-II, the evidential expected value of one-boxing is greater than that of two-boxing, and the causal expected value of two-boxing is greater than that of one-boxing. Again, if there is no sense of 'ought' which is relative to one's decision-theoretic uncertainty, we cannot account for these verdicts.

What's more, as MacAskill argues, the standard intuitions about the existing most influential decision-theoretic thought experiments—*Standard Newcomb, The Smoking Lesion*, and *The Psychopath Button*—can also be explained in terms of the idea that we are intuitively hedging in the face of decision-theoretic uncertainty. And it seems that, empirically, people's intuitions about Newcomb problems are stakes sensitive: in a study that we and some coauthors ran, subjects were more likely to one-box in Newcomb's problem if the stakes were higher for EDT than for CDT.

Insofar as accordance with people's intuitions in these cases has been taken to be a desideratum of a decision theory, this provides some support for the idea that, in some sense of 'ought', one ought to hedge in the face of decision-theoretic uncertainty.

2.3 Paying for Information

Consider again the *Moral Newcomb* problem with which we started. Suppose that before making up your mind, you learn that, via a time machine, a book has been delivered from the year 10,000 CE. On the blurb it claims that after long reflection our descendants have discovered the correct decision theory, with arguments so clear and compelling that anyone can understand and everyone will be convinced. It even has various applications of the view, including to the precise problem you're facing. However, you have to pay $5 in order to read the book.

Let's bracket, for the purpose of this example, any intrinsic value from knowledge of the truth about rationality, any fame that the reader might receive for explaining its arguments to others, and so on: we can imagine that you know you will immediately forget the contents of the book after making your decision. If there is no sense of 'ought' which is relative to one's decision-theoretic uncertainty, you should not buy and read the book. From EDT's perspective, paying $5 to read the book and then one-boxing is strictly worse than just one-boxing right away. Similarly, from CDT's perspective, paying $5 to read the book and then two-boxing is strictly worse than just two-boxing right away. So, if you ought to do simply what you ought to do on the correct decision theory, you ought not to pay the $5 to read the book.

This seems deeply counterintuitive: given the stakes—ten lives saved for EDT and one life saved for CDT—it would seem reckless not to incur such a trivial cost before taking action. But if so, there is some notion of 'ought' which is relative to one's decision-theoretic uncertainty.

2.4 Outstanding Issues

The argument from stakes-sensitivity supports the claim that one should hedge in the face of decision-theoretic uncertainty, whereas the other two arguments only support the weaker claim that there is some notion of 'ought' that takes decision-theoretic uncertainty into account. However, it seems clear to us that if one accepts that there is such a notion of 'ought', one should plausibly accept hedging as well. In the context of moral uncertainty, most of the resistance to hedging from those who accept the relevant notion of 'ought' is driven by skepticism about the possibility of intertheoretic value comparisons (e.g. Gustafsson and Torpman 2014:165). But as we will argue in the next section, such comparisons are significantly easier to make in the case of decision-theoretic uncertainty.

Nevertheless, the view that one should hedge under decision-theoretic uncertainty still faces some outstanding issues. In order to know precisely how to take (first-order) decision-theoretic uncertainty into account, one needs a second-order decision theory. Although we have not presented a specific second-order decision theory, we have argued that any plausible account should endorse hedging; candidate second-order decision theories would include 'meta causal decision theory' and 'meta evidential decision theory' (MacAskill 2016). However, if one should be uncertain among first-order decision theories, one should presumably also be uncertain among second-order decision theories. Moreover, if uncertainty among first-order decision theories can make a difference for what one

should do, then presumably so can uncertainty among second-order decision theories. In order to take this uncertainty into account we need a third-order decision theory, and thus we are seemingly lead into an infinite regress of uncertainty at higher and higher orders. This infinite regress represents a significant challenge to the claim that one should hedge in the face of first-order decision-theoretic uncertainty. If we could somehow be certain that the correct second-order decision theory—whatever it is—endorses hedging, then we would be in the clear. But we should not assign zero credence to *My Favourite Theory*, according to which one should simply act in accordance with the first-order theory one has highest credence in, and which therefore does not endorse hedging.[9] We recognise that this is a major challenge, but for the purposes of this paper we will assume that some solution can be obtained.[10]

A second challenge concerns the question of what notion of 'ought' is at play in the claim that one ought to hedge under decision-theoretic uncertainty.[11] In the corresponding literature on moral uncertainty, one common strategy is to say that although the notion of 'ought' that is provided by the correct moral theory is a moral ought, the notion of 'ought' that is relative to one's moral uncertainty is a rational ought.[12] However, on the face of it, this strategy appears to be unavailable in the present context. The ought which is relative to the correct (first-order) decision theory is already a rational

[9] See Gustafsson and Torpman (2014) for a discussion and defence of *My Favourite Theory* in the context of moral uncertainty.

[10] For detailed discussion of the regress problem, see Trammell (2019), MacAskill, Ord, and Bykvist (2019, chapter 1.3), and Tarsney (manuscript).

[11] See Weatherson (2014), Harman (2015) and Hedden (2016) for this objection in the context of moral uncertainty.

[12] See Sepielli (2013) and MacAskill, Ord, and Bykvist (forthcoming) for further discussion.

ought, and therefore it seems we cannot appeal to that distinction here.[13] Again, we shall not try to resolve this issue here.

For these two reasons, we think it's an open question whether you ought to hedge in the face of decision-theoretic uncertainty: whichever view you take, you have some hard problems to grapple with. However, we believe the arguments we've given in favour of hedging make the idea that one should hedge sufficiently plausible that it's interesting and important to work out what one should do *if* one should hedge. In what follows, we will therefore assess the conditional claim that if you ought to hedge, then you should generally take the action that EDT recommends.

3 Intertheoretic Comparisons

In order for the claim that one ought to hedge in the face of decision-theoretic uncertainty to be meaningful, we must have some way of comparing evidential expected value with causal expected value. Otherwise we will not be able to make sense of the claim that the stakes are higher for EDT than they are for CDT. In the literature on moral uncertainty, this is known as the problem of intertheoretic value comparisons. To get a sense of the difficulty of this problem, consider the

[13] However, if we think of supplementing a given axiology with either an evidential or a causal decision principle as giving rise to two distinct moral theories, we can still appeal to this distinction. As an analogy, consider the fact that a utilitarian axiology can be combined with both a risk-neutral and a risk-averse principle for evaluating actions under empirical uncertainty. If we think of risk-neutral and risk-averse utilitarianism as two distinct moral theories, then we should plausibly say the same thing about evidential and causal utilitarianism. Of course, this maneuver will not work for all cases of uncertainty between EDT and CDT.

following case. Suppose that you're facing the *Footbridge* trolley problem, but your credence is evenly split between a consequentialist theory and a deontological theory. If you should hedge under moral uncertainty then you should sacrifice the one in order to save the five if the stakes are higher for the consequentialist theory, and refrain from doing so if the stakes are higher for the deontological theory. But how are you to put the two theories on a common scale in order to know whether the stakes are higher for the consequentialist or the deontological theory? The theories themselves do not seem to come equipped with an answer to this question, and it's hard to see what other information one could appeal to in order to resolve it. Multiple proposals have been made in the literature, but it's clear that there is not yet any consensus solution.[14]

Luckily, the problem of intertheoretic comparisons is significantly easier in the case of decision-theoretic uncertainty. In order to see this, let us first state EDT and CDT more precisely. According to EDT, you should choose the action with the highest evidential expected value (EEV) , calculated as follows:

$$EEV(A) = \sum_{i=1}^{n} P(O_i \mid A) V_{EDT}(O_i),$$

where O_1, ..., O_n are the possible outcomes, P is the agent's credence function, and V_{EDT} is her value function. The EEV of an action A is the sum product of the value of each outcome and the probability of that outcome on the assumption that A is performed. According to CDT, by contrast, you should choose the action with the highest causal expected value (CEV), calculated as follows:

[14] For discussion of intertheoretic comparisons in the case of moral uncertainty, see Sepielli (2009, 2010), MacAskill (2014, chapter 4), Tarsney (2018), and MacAskill, Ord and Bykvist (2019, chapter 5).

$$CEV(A) = \sum_{i=1}^{n} P(A \Rightarrow O_i) V_{CDT}(O_i),$$

where $A \Rightarrow O_i$ denotes the counterfactual conditional "if I were to do A, then O_i would occur." The CEV of an action A is the sum product of the value of each outcome and the probability that A causes that outcome. Thus EDT and CDT are both versions of the claim that one should maximise expected value. They only disagree on how this expectation is to be calculated: according to EDT, one should use the conditional probability, whereas according to CDT, one should use the probability of the counterfactual conditional.

This observation suggests a natural procedure for intertheoretic value comparisons. If the disagreement between EDT and CDT only concerns which type of probability should be used to calculate expected value, then we should not construe them as having different value functions. Intuitively, a decision theory does not prescribe that you value outcomes in some particular way. Instead, it tells you how to act in light of the values (and beliefs) you in fact hold. Therefore, we will argue that the problem of intertheoretic value comparisons is solved by requiring that $V_{EDT} = V_{CDT}$. By imposing this condition we place the two theories on a common scale, thereby rendering intertheoretic value comparisons straightforward. To determine whether the stakes are higher for EDT or CDT in the *Moral Newcomb* problem, we simply calculate all expected values with respect to the shared value function and then compare the two quantities EEV(*One-box*) – EEV(*Two-box*) and CEV(*Two-box*) – CEV(*One-box*). Moreover, this proposal has the desirable implication that the evidential and causal expected value of an action A will only differ if there is some outcome O such that $P(O \mid A) \neq P(A \Rightarrow$

O). That is, they only differ in those cases where one would intuitively expect EDT and CDT to come apart.

This proposal might seem obvious to some but to argue for it we must first say a bit more about the nature of value functions in decision theory. For our purposes, there are two salient ways of thinking about these value functions. First, we can think of them as externally given, and therefore independent of the choice of a particular decision theory. Second, we can think of them as constructed, via a representation theorem, from the agent's relational attitude. Let us begin with the former. Given that we presented the Wager as an argument for why *altruistic and morally motivated* agents should generally favour EDT in the face of decision-theoretic uncertainty, it is natural to think of the value functions as being externally provided by some moral theory. In particular, suppose that the agent facing the *Moral Newcomb* problem is certain of some axiology, and that she does not take there to be any relevant side constraints: in deciding whether to one-box or two-box, she simply wants to perform the action that in expectation leads to the best outcome according to her favoured axiology. The only thing she's uncertain about is how she should take into account information about causation and correlation when making her decision. In this case it is clear that she does not face a problem of intertheoretic comparisons, because both V_{CDT} and V_{EDT} are given by the relevant axiology.

The same point holds if she is instead acting under axiological uncertainty. If she has settled on some value function V which represents the value she assigns to outcomes in light of her axiological uncertainty (for example, V might be the expected value of the outcome under axiological uncertainty), then this is the value function that should be used to calculate both causal and evidential

expected value. Again, if it were the case that $V_{CDT} \neq V_{EDT}$, then at least one of the two would not accurately represent how she values states of affairs under axiological uncertainty. In general, if there is some externally given value function over outcomes, there will not be any problem of intertheoretic comparisons to begin with.

But suppose that there isn't any externally given value function. A second approach is to say that the value functions are constructed (along with the corresponding probability functions) via a representation theorem from the agent's relational attitudes, typically her preferences over options. A representation theorem shows that if an agent's preferences over options satisfy certain conditions, then she can be represented as maximising expected value with respect to some probability function P and some value function V, in the sense that for any pair of options *A* and *B*, she prefers *A* to *B* just in case the expected value of *A* is greater than the expected value of *B* (with the expected value calculated relative to P and V).

Suppose therefore that both V_{EDT} and V_{CDT}, along with their respective probability functions P_{EDT} and P_{CDT}, are constructed via representation theorems. In order to compare them, we need a framework which is broad enough to be able to express both EDT and CDT. This means that, for our purposes, the most relevant result is Joyce's (1999:239) very general representation theorem, which can underwrite both evidential and causal decision theory. Joyce takes as his starting point both a preference relation and a comparative belief relation. Both of these relations are *conditional* (or *suppositional*) in the sense that they are defined over options of the form '*A* on the condition that *B*.' For example, I might prefer going for a walk on the supposition that its sunny to staying inside reading

on the supposition that it rains, or I might find rain on the supposition that it's cloudy more likely than a breeze on the supposition that it's sunny.

Joyce imposes requirements on both the preference relation and the comparative belief relation which together ensure the existence of a unique probability function and a unique (up to positive linear transformation) value function. By imposing further conditions on the comparative belief relation so as to make the supposition behave either like evidential or causal supposition, he is able to derive both EDT and CDT as special cases. Now, if we take the relevant preferences and comparative beliefs to be the ones that the agent in fact has, then we cannot construct both V_{CDT} and V_{EDT} simultaneously. After all, if she satisfies the conditions needed for EDT she will fail to satisfy the conditions needed for CDT. But we can proceed as follows. In order to construct a cardinal value function, it is sufficient to consider the agent's preferences and comparative beliefs on just a single supposition. So all we have to do is find a proposition such that the preferences and comparative beliefs of EDT are in agreement with those of CDT on the supposition that this proposition is true. And here a natural candidate suggests itself: the tautology. Evidential and causal supposition will not yield different constraints on preferences and comparative beliefs when the proposition being supposed is the tautology.[15] This allows us to say that EDT and CDT share the same value function.[16]

[15] The only case in which this would not work is if there is some proposition A such that $P(A \mid \top) = P(\top \Rightarrow A) = 0$, yet some other proposition B such that $P(B \Rightarrow A) > 0$.

[16] More generally, Joyce (1999:178) concedes that Jeffrey's (1983) evidential decision theory has the correct account of value ("desirability") and argues that causal expected utility is simply desirability from the perspective of the supposition that an act will be performed. The same sentiment is expressed by Bradley (2017:170). Again, these claims indicate that an agent's value function should not depend on which decision theory she takes to be correct.

In summary, either there is an externally given value function, for example one provided by an axiology, or the value functions of EDT and CDT are constructed via representation theorems. In the former case, we can straightforwardly conclude that $V_{EDT} = V_{CDT}$. In the latter case, we can use Joyce's general representation theorem to construct both value functions, thereby ensuring that they are in agreement with one another.

You might object that our proposal about how to make intertheoretic value comparisons 'stacks the deck' in favour of EDT because, as the evidentialist's wager shows, in general under decision-theoretic uncertainty, with this choice of intertheoretic comparisons, one will act in accordance with EDT's recommendation. As an analogy, consider the question of how to do intertheoretic value comparisons between average and total utilitarianism. Suppose we fix such comparisons by stipulating that in a world with only one individual, the difference in value between any two actions is the same for average and total utilitarianism. This implies that in a world with *n* individuals, giving each of them one additional unit of welfare would be *n* times more valuable according to total utilitarianism than according to average utilitarianism. In effect, this means that total utilitarianism will swamp average utilitarianism: if one should hedge under moral uncertainty, then one will generally go with total rather than average utilitarianism whenever they are in conflict, because the stakes are much higher for the former theory. It has been argued that this is a good reason for thinking that we should not make the intertheoretic comparisons between average and total utilitarianism in this way.[17] Perhaps using $V_{EDT} = V_{CDT}$ to settle comparisons between EDT and CDT is analogous to using the one-person

[17] See for example Hedden (2016) and Cotton-Barratt, MacAskill and Ord (forthcoming).

world to settle comparisons between total and average utilitarianism, and should be rejected for the same reason.

However, the analogy is flawed. When normalising average and total utilitarianism at the one-person world, there are no scenarios in which the stakes are higher for average utilitarianism than for total utilitarianism. By contrast, using the case of no non-causal correlations to normalise does allow for scenarios in which the stakes are higher for CDT than for EDT, as the following example shows.

Evil Twin

In front of you are two boxes, A and B. You can choose either B only, or both A and B. Box A is guaranteed to contain ten thousand dollars to be donated to an effective charity, whereas box B may or may not contain twenty thousand dollars to be donated to an effective charity. A 99% reliable predictor has made a prediction about your decision. If she predicted that you would take both boxes, she left box B empty. If she predicted that you would take box B only, she put twenty thousand in that box. However, you also know that your evil twin is facing the same decision problem. Being evil, he will donate the money to an anti-charity. One dollar to the anti-charity precisely counterbalances one dollar to the charity. However, all of the monetary amounts in his decision problem are half the size of yours. Being your twin, his decision is perfectly correlated with your own.

For CDT, the fact that your evil twin also faces a similar decision problem makes no difference to the stakes: CEV(Two-box) – CEV(One-box) = $10,000. But for EDT, it means that the stakes are half as big as they would otherwise have been: EEV(One-box) – EEV(Two-box) = $4,800. Hence our proposal for how to do intertheoretic comparisons, when combined with a claim about correlations, now yields the result that the stakes are lower for EDT than for CDT, rather than the other way around. Therefore, it is not true that our account makes EDT swamp CDT in general. Whether or not it does will depend on which correlations the agent takes to hold.

4 Finite Case

We have argued (or at least attempted to make plausible) that it is rational to hedge in the face of decision-theoretic uncertainty. That is, we have argued that even if your credence in CDT is substantially higher, you should nevertheless follow EDT if the stakes are much higher for EDT. We have also presented an account of intertheoretic comparisons that allows us to say when the stakes are higher for one decision theory than another. Together, these imply that if the expected number of correlated decision-makers is large enough, you should one-box in the *Moral Newcomb* problem even if you have significantly higher credence in CDT than in EDT. Recall that the possible payoffs are as follows:

	Cure in both	Cure in one only
Take one box	Ten lives	Nothing

Take both boxes	Eleven lives	One life

Suppose that the predictor is 99% reliable. This means that, before taking any correlated agents into account, the stakes are 8.8 times higher for evidential than for causal decision theory. Suppose now that you have merely 1% credence in EDT, and 99% credence in CDT. If we were then trying to maximise expected value over decision-theoretic uncertainty, the existence of correlated agents would have to increase the relative stakes for EDT by a factor of 11.25 in order for one-boxing to be the rational option.

Why should one expect there to be all these correlated agents facing *Moral Newcomb* problems? Our argument doesn't require that the correlation in question be perfect. Consider now the vast number of humans and similarly reasoning aliens that could exist in the future. For example, Bostrom (2013:18) estimates that over the course of the future, Earth could sustain 10^{16} lives of normal duration. If we instead assume that we will at some point spread beyond Earth, or that life will eventually be primarily digital, he gives the corresponding estimates of 10^{34} and 10^{54} life years respectively. Probably, some of these future people will be similar to you in terms of how they approach Newcomblike problems. Therefore, your decision to one-box constitutes evidence that they will also one-box. Of course, the evidence isn't perfect, but if the number of correlated agents is sufficiently large, the impact will be the same as that of many identical copies. For example, if your decision to one-box only increased the probability that the correlated agents one-box by 1%, there would only have to be 1025 agents who are facing problems similar to *Moral Newcomb* and who are correlated at this level in order for hedging to be rational.

Now, establishing what credence distribution one should have over the number of correlated agents facing decision problems of this kind, and their degree of correlation, is a thorny empirical matter over which there will be reasonable disagreement. Yet we believe that the numbers provided here are conservative, and that it would certainly not be unreasonable for someone to have such beliefs in light of the vast number of people that could exist in the future.[18] [19]

This establishes that if you are a morally motivated and altruistic agent, you should one-box in *Moral Newcomb* even if you have significantly higher credence in causal than evidential decision theory. But earlier, we claimed that there is nothing special about the *Moral Newcomb* case, and that our reasoning therefore supports the following more general conclusion:

---

[18] Moreover, if you object to the argument on empirical grounds alone, you still have to accept the surprising implication that beliefs about distant future agents are relevant for how you should act under decision-theoretic uncertainty.

[19] You might object that perhaps there are also anti-correlated decision makers out there, who tend to one-box when you two box. If there are as many anti-correlated decision-makers as there correlated ones, this would effectively cancel things out for EDT. If there are sufficiently many more anti-correlated than correlated ones, EDT would recommend two-boxing rather than one-boxing. Our argument assumes that there are many more correlated decision-makers than there are anti-correlated ones. But unless you have some particular reason to think that you are special, this seems like a standard inductive inference. Finally, if you remain unconvinced of the claim that there are more correlated than anti-correlated decision makers, you shouldn't believe that there are equally many decision makers of both types, but rather that there are many more anti-correlated ones. After all, it would be a rather striking coincidence if the two types of decision makers were evenly balanced. But if there are more anti-correlated than correlated decision-makers, then our argument would support the different but similarly startling conclusion that even if you're virtually certain of EDT, you should nevertheless *two-box* in the *Moral Newcomb* case.

> In general, and across a wide variety of decision contexts, if you are a morally motivated and altruistic agent, and are uncertain between EDT and CDT, you should typically act in line with the former even if you have significantly higher credence in the latter.

Is this more general claim correct? Or could it be that, although the Wager works in the *Moral Newcomb* case, there are other cases in which EDT and CDT come apart, where hedging under decision-theoretic uncertainty doesn't lead an altruistic agent to choose the option recommended by EDT? Recall that what is driving our argument is the claim that your decision is correlated with the decisions of other agents. This means that your decision to perform an action provides evidence that correlated agents will also perform that action. Moreover, if a correlated agent's decision to perform that action provides evidence that some desirable outcome will obtain, then your decision also provides evidence that that outcome will obtain.

In general, the existence of correlated agents will affect the stakes for EDT, but not for CDT. Now, as we saw in the *Evil Twin* case, it is possible to construct cases in which such correlations have the effect of *decreasing* the stakes for EDT. If the decrease is sufficiently large, the wager will run in the opposite direction: even if one has significantly higher credence in EDT than in CDT, one should nevertheless act in accordance with the latter. However, we contend that such cases are few and far between. In the *Evil Twin* case, you are only correlated with one other agent who faces a problem with lower stakes, and whose decision will partially cancel out (by donating to an anti-charity) the good you achieve

through your own decision (by donating to a charity). But once we take into account the vast number of people who may exist in the future, it's overwhelmingly likely that whenever your decision is correlated with one person's decision, it will also be correlated with very many other people's decisions. It would therefore be a striking coincidence if the effect of these correlations would be to lower the stakes for EDT. If you have sufficiently many evil twins (or evil *n*-tuplets), the stakes will be higher for EDT than CDT (although now EDT will also recommend two-boxing). If you have just one good twin (and no evil twins), the stakes will also be higher for EDT. In general, once we take into account that there will be very many correlated agents, the only case in which the stakes will be lower for EDT is when (i) either the degrees of correlation or the "isolated" stakes faced by correlated agents are sufficiently low, and (ii) performing the action that EDT would have recommended in the absence of correlated agents provides evidence that these agents will partially cancel out the good you achieve through your own decision (in the same sense that your evil twin partially cancels out the good you achieve through one-boxing).

What if one believes that anti-correlated agents vastly outnumber correlated ones? If one believes this in the *Moral Newcomb* case, the conclusion of our argument is that one should two-box rather than one-box. However, this does not constitute a counterexample to our claim that a morally motivated and altruistic agent should generally act in line with EDT even if she has significantly higher credence in CDT. To see this, note that even if the agent were certain of EDT, she would still two-box rather than one-box. Rather than being a counterexample, this is simply a case in which the agent's belief that there are many more anti-correlated than correlated agents (together with the assumption that she is altruistic and morally motivated) imply that EDT recommends a different option than what one

might naïvely expect. What's more, the stakes are again higher for EDT than they are for CDT, so in this sense it's still EDT that is driving the decision.[20]

In summary, the existence of correlated decision-makers will affect the stakes for EDT but not for CDT. We have argued on empirical grounds that it's reasonable to believe that there are sufficiently many such correlated decision-makers so as to make hedging by following the recommendation of EDT rational. The empirical argument was based on the assumption that the universe is finite. So let's now consider what happens if the universe is instead infinite.

5 Infinite Case

If, as many of our best cosmological theories imply, the universe is infinite, and every physically possible configuration of matter is realised in infinitely many regions of spacetime, there will clearly be sufficiently many correlated agents.[21] Indeed, there will be infinitely many identical copies of you who are facing the same decision situation. So you might think that, if anything, the argument for the Wager becomes stronger if the universe is infinite. But the infinite case gives rise to a host of further complications, and we need to introduce additional principles to be able to compare worlds that contain infinite amounts of value, or to be able to compare actions in such worlds. We will not be able

[20] If one believes that there are equally many correlated and anti-correlated agents (counting oneself among the correlated ones), then EDT will be indifferent between the two options, and one's decision will therefore be driven by CDT. But such symmetry should be regarded as extremely unlikely.

[21] See Garriga and Vilenkin (2001) and Knobe, Olum and Vilenkin (2006) for discussion of these cosmological theories.

to undertake a complete survey of proposed principles so, instead, we will merely note that the argument goes through on at least one leading account of infinite ethics.

Our first problem: In infinite worlds, can we even say that EDT recommends one-boxing over two-boxing in *Moral Newcomb*? If you one-box, you obtain strong evidence that infinitely many other agents will receive ten doses of the cure. But if you two-box, you obtain strong evidence that infinitely many other agents will receive one dose of the cure. In standard cardinal arithmetic, the total number of doses will be the same in both cases, thereby seemingly implying that EDT will be indifferent between the two actions. But that seems like the wrong conclusion. So we should invoke additional principles that allow us to discriminate between different infinite worlds.

To keep things simple, let us for the moment only consider agents who are perfect duplicates of you, and let us compare two worlds: in w_1, you all one-box and save ten lives each, and in w_2, you all two-box and and save one life each. We will assume that these two worlds are perfectly alike in all other respects. That is, these two worlds share all of the same locations, and for all locations except those corresponding to lives that are saved in one world but not in the other, the value at that location is the same in both worlds. To solve the first problem, we need an account which allows us to say that w_1 is better than w_2. One such account is that of Vallentyne and Kagan (1997). Let R_1, R_2, R_3, ... be a sequence of larger and larger spacetime regions that grow without bound, centred on the location in which you find yourself. The proposal is as follows:

Catching Up

> For any worlds w_1 and w_2, if there is some n such that for any $m > n$, the value in region R_m of w_1 is greater than the value in region R_m of w_2, then the value of w_1 is greater than that of w_2.[22]

If we assume that there is some n such that for any $m > n$, the region R_m contains more correlated than anti-correlated agents, it follows that w_1 is better than w_2. This means that although anti-correlated agents may outnumber correlated ones in some finite regions, we can always find larger regions that contain them where the reverse is true. If we are happy to assume that there are more correlated than anti-correlated agents in a finite universe, we should also be happy to make this further assumption in the infinite case. So the *Catching Up* principle allows us to make sense of the idea that w_1 is better than w_2, and therefore that EDT recommends one-boxing in *Moral Newcomb*.

The second problem is to make sense of the claim that the decision in *Moral Newcomb* is much higher stakes for EDT than it is for CDT, and that you should therefore hedge under decision-theoretic uncertainty. That is, we need to be able to say that even if your credence in CDT is significantly higher, you should nevertheless follow the recommendation of EDT. On its own, *Catching Up* doesn't allow us to say this, because it only tells us how to compare *worlds* that contain potentially infinite amounts of value. But what we need is a way of comparing *actions* in such worlds. In light of this,

[22] This is roughly equivalent to the SBI2 proposal of Vallentyne and Kagan (1997:14), skipping over some technical details. This proposal assumes that it makes sense to speak of the same location in the two worlds being compared. In the present case, this assumption will be satisfied.

Arntzenius (2014) suggests a modification of the view. His proposal is simple: just replace talk of value in a region with talk of *expected value* in a region.

> *Expected Catching Up*
>
> For any actions A and B, if there is some n such that for any $m > n$, the expected value of A in region R_m is greater than the expected value of B in region R_m, then A is better than B.

In order to apply this to the case at hand, let's suppose for the moment that our approach to decision-theoretic uncertainty is to maximise meta expected value, where the meta expected value of an action A is given as $MEV(A) = EEV(A)P(EDT) + CEV(A)P(CDT)$. For EDT, the stakes will keep growing without bound as we consider larger and larger regions. By contrast, for CDT the stakes will remain the same in any finite region. Therefore, we will be able to find an n such that for any $m > n$, the meta expected value in R_m of one-boxing is greater than that of two-boxing, and thereby conclude that one-boxing is rational. But importantly, this conclusion does not depend on the assumption that maximising meta expected value is the appropriate way to behave under decision-theoretic uncertainty. For any non-zero credence in EDT and any view about how large the difference in stakes must be in order for hedging to be rational given that credence, we will always be able to find a region R_n such that for any $m > n$, the difference in stakes between EDT and CDT is sufficiently large to justify one-boxing. Note, in addition, that this also means that we don't have to rely on any possibly contentious claims about how to do intertheoretic comparisons between EDT

and CDT. As long as $V_{CDT} = kV_{EDT} + m$ for some finite k and m, it follows that we will be able to find an appropriate region R_n.

So if *Expected Catching Up* is correct, the Wager goes through in infinite worlds as well. And some other accounts, such as Bostrom's (2011) "hyperreal" approach, would also endorse the Wager.[23] Roughly speaking, by representing the value of infinite worlds using hyperreal numbers, we are able to say that the world in which all correlated agents receive ten doses of the cure is better than the world in which they only receive one dose of the cure, even though both worlds contain infinite amounts of value. And given that hyperreal numbers can be straightforwardly multiplied with probabilities, this approach allows us to make sense of the claim that the stakes are higher for EDT than for CDT.[24]

Of course, however, infinite ethics is a fiendish topic, and there is no consensus on what the right view is, because any view faces grave problems. *Expected Catching Up*, for example, conflicts with the very plausible *Pareto* principle:

> *Pareto*
>
> For any worlds w_1 and w_2 that contain exactly the same individuals, if every individual has at least as much welfare in w_1 as she does in w_2, then w_1 is at least as good as w_2. If in

[23] Another class of theories that will also endorse the Wager are those that discount value by distance, at least provided that the discount rate is not steep enough to lower the stakes for EDT to the point where hedging no longer favours one-boxing, although it's fair to say that such theories are not regarded as leading contenders.

[24] However, see Arntzenius (2014:49–52) for discussion of some difficulties for the hyperreal approach.

addition at least one individual has more welfare in w_1 than in w_2, then w_1 is better than w_2.

To see this, assume for simplicity that either all correlated agents save ten lives each and all anti-correlated agents save one life each, or vice versa. You can either one-box or two-box. This gives us four worlds: one in which you and all correlated agents one-box, one in which you one-box but your correlated agents two-box, and so forth. We can now consider permutations of these worlds that contain exactly the same individuals at exactly the same welfare levels, except that all people saved by anti-correlated agents are closer to you in spacetime than any of the people saved by correlated agents. With respect to these permuted worlds, *Expected Catching Up* implies that two-boxing has greater expected value than one-boxing. In both cases, we calculate expected value by assigning the same probability to a world and its permutation, so if one-boxing has greater expected value in one case but not the other, it follows that not all permuted worlds are equally as good as the worlds they permute. But this is precisely what *Pareto* rules out.

However, as Askell (2018) shows, if one endorses *Pareto* and some other very plausible assumptions, one will have to accept that there is widespread incomparability between infinite worlds. So we can state our conclusion, with respect to the infinite case, in a restricted form: *if* one wishes to avoid widespread incomparability between infinite worlds, then one will probably endorse a view that supports the Wager.

6 Conclusion

We have argued that an altruistic and morally motivated agent who is uncertain between CDT and EDT (or some other non-causal decision theory like FDT) should in general act in accordance with the latter, even if she has greater credence in the former. To arrive at this conclusion, we first argued that it is rational to hedge in the face of decision-theoretic uncertainty. That is, if the stakes are much higher for one theory than another, and the credences assigned to the two theories aren't very different, then one should act in accordance with the higher-stakes theory. In order to say whether the stakes are higher for one theory rather than another, we need an account of intertheoretic value comparisons. We argued that such comparisons should be made by letting EDT and CDT have the same value function over outcomes. We noted that, given the assumption of altruism, the existence of correlated decision-makers will affect the stakes for EDT but not for CDT. Finally, we argued that for reasonable credence distributions over EDT and CDT, there will be sufficiently many correlated decision-makers so as to make it rational in general to follow the recommendation of EDT. In the finite case, we appealed to estimates about how many people will exist in the future. In the infinite case, there is guaranteed to be sufficiently many correlated agents, but further complications arise. We showed that on one natural way of solving these, the argument still goes through. If we are altruistic and morally motivated agents, we should mostly follow EDT.[25]

7 Bibliography

[25] MacAskill came up with the central argument. Vallinder further developed this argument and took the lead in writing the paper. Oesterheld, Shulman, and Treutlein (listed in alphabetical order) had all independently discovered the same argument. All coauthors contributed extensive comments and suggestions in the process of writing the paper.

Arntzenius, Frank (2014). Utilitarianism, Decision Theory, and Eternity. *Philosophical Perspectives* **28**:31–58.

Askell, Amanda (2018). *Pareto Principles in Infinite Ethics*. PhD thesis, New York University.

Bostrom, Nick (2011). Infinite Ethics. *Analysis and Metaphysics* **10**:9–59.

Bostrom, Nick (2013). Existential Risk Prevention as Global Priority. *Global Policy* **4**(1):15–31.

Bourget, David and Chalmers, David (2014). What do philosophers believe? *Philosophical Studies* **170**(3):465–500.

Bradley, Richard (2017). *Decision Theory with a Human Face*. Cambridge: Cambridge University Press.

Cotton-Barratt, Owen, MacAskill, William, and Ord, Toby (forthcoming). Normative Uncertainty, Intertheoretic Comparisons, and Variance Normalisation.

Garriga, Jaume and Vilenkin, Alexander (2001). Many worlds in one. *Physical Review D* **64**.

Greene, Preston (2018). Success-First Decision Theories. In Arif Ahmed (ed.), *Newcomb's Problem*. Cambridge: Cambridge University Press.

Greene, Preston and Levinstein, Ben (manuscript). Act Consequentialism without Free Rides.

Gustafsson, Johan E. and Torpman, Olle (2014). In Defence of My Favourite Theory. *Pacific Philosophical Quarterly* **95**(2):159–174.

Harman, Elizabeth (2015). The Irrelevance of Moral Uncertainty. *Oxford Studies in Metaethics* **10**.

Hedden, Brian (2016). Does MITE Make Right?: Decision-Making Under Normative Uncertainty. *Oxford Studies in Metaethics* **11**:102–128.

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

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

Knobe, Joshua, Olum, Ken D., and Vilenkin, Alexander (2006). Philosophical Implications of Inflationary Cosmology. *British Journal for the Philosophy of Science* **57**(1):47–67.

MacAskill, William (2014). *Normative Uncertainty*. DPhil thesis, Oxford University.

MacAskill, William (2016). Smokers, Psychos, and Decision-Theoretic Uncertainty. *Journal of Philosophy* **113**(9):425–445.

MacAskill, William, Ord, Toby, and Bykvist, Krister (2019). *Moral Uncertainty*. Oxford: Oxford University Press.

Nozick, Robert (1993). *The Nature of Rationality*. Princeton: Princeton University Press.

Sepielli, Andrew (2009). What To Do When You Don't Know What To Do. *Oxford Studies in Metaethics* **4**:5–28.

Sepielli, Andrew (2010). *'Along an imperfectly-lighted path': practical rationality and normative uncertainty*. PhD thesis, Rutgers University.

Sepielli, Andrew (2013). What To Do When You Don't Know What To Do When You Don't Know What To Do... *Noûs* **47**(1):521–544.

Soares, Nate and Levinstein, Ben (forthcoming). Cheating Death in Damascusi. *The Journal of Philosophy*.

Tarsney, Christian (2018). Intertheoretic Value Comparison: A Modest Proposal. *Journal of Moral Philosophy* **15**(3):324–344.

Tarsney, Christian (manuscript). Metanormative Regress: An Escape Plan.

Trammell, Philip (2019). Fixed-point solutions to the regress problem in normative uncertainty. *Synthese*, Online First.

Vallentyne, Peter and Kagan, Shelly (1997). Infinite Value and Finitely Additive Value Theory. *Journal of Philosophy* **94**(1):5–26.

Weatherson, Brian (2014). Running Risks Morally. *Philosophical Studies* **167** (1):141–163.

Yudkowsky, Eliezer and Soares, Nate (2017). Functional Decision Theory: A New Theory of Instrumental Rationality. arXiv: 1710.05060 [cs.AI].