SMOKERS, PSYCHOS, AND DECISION-THEORETIC UNCERTAINTY*

Recently, there has been significant debate about the nature of decision theory: whether the correct theory is evidential, causal, or something else. The principal problem in this debate is that powerful counterexamples seem to have been raised to all the major views. *The Smoking Lesion* seems to be a decisive counterexample to evidential decision theory, but *The Psychopath Button* seems to be a decisive counterexample to causal decision theory.

In response to this problem, some philosophers have expressed pessimism. Rachael Briggs argues that “no decision rule can do everything that we want”;¹ Andy Egan regretfully asserts that he “do[es] not have . . . a theory to offer”² that is able to get the intuitions right in the cases that have been given in the literature. Others instead have tried to develop new decision theories that satisfy the intuitions,³ but these generally come at the cost of considerable theoretical inelegance or suffer from counterexamples of their own.

In this article, I propose a new way of making rational sense of our seemingly divergent intuitions across cases. The key is the idea, briefly

---

* For very helpful comments on earlier drafts of this article, I would like to thank Arif Ahmed, Frank Arntzenius, Amanda Askell, John Broome, Krister Bykvist, Andy Egan, Hilary Greaves, Toby Ord, Mike Otsuka, Stefan Riedener, Pablo Stafforini, Bastian Stern, Ralph Wedgwood, and audiences at CRNAP, Oxford, and LSE.


suggested by Robert Nozick,⁴ that there is an important sense of ‘ought’ (though not the only sense of ‘ought’) according to which a decision maker ought to take their decision-theoretic uncertainty into account when making decisions. I call the general idea that there are norms that take into account normative uncertainty metanormativism, and any decision theory that takes decision-theoretic uncertainty into account a meta decision theory.

Metanormativism has principally been explored in relation to moral uncertainty (although the project currently lacks a name, hence my introduction of ‘metanormativism’). The standard view in that literature is that, under moral uncertainty, one should maximize expected choice-worthiness, or, equivalently, minimize expected wrongness.⁵ Proponents of metanormativism about moral uncertainty do not propose maximize expected choice-worthiness as a rival view to first-order moral theories. Rather, they think that there are different senses of ‘ought’: a first-order moral sense of ‘ought’, which is not sensitive to a decision maker’s moral uncertainty, and a different (more subjective or less idealized) sense of ‘ought’ that takes moral uncertainty into account. In the same way, in this article I do not propose meta decision theory as a rival to causal decision theory or evidential decision theory. Rather, a different sort of ‘ought’ is at play—one that is relevant to less idealized agents than the ‘ought’ of first-order decision theory.

My intention in this article is not to defend metanormativism about decision theory, although I briefly offer some motivation for it in section II. Rather, my intention is to show that, if metanormativism about decision theory is true, then it has two important implications for the causal versus evidential decision-theory debate. First, it allows us to make rational sense of our seemingly divergent intuitions across The Smoking Lesion and The Psychopath Button. Second, it generates strong new arguments for preferring the causal approach to decision theory over the evidential approach.

The structure of my argument is as follows. After quickly describing Newcomb’s problem and the causal versus evidential distinction in section I, in section II, I introduce and briefly motivate metanormativism


about decision theory and discuss Nozick’s views. In section iii, I give the most natural formulation of meta decision theory and show how it gets the right intuitive results in both The Smoking Lesion and The Psychopath Button. I show how meta decision theory can convincingly explain why we get the intuitions we do in a way that is far more theoretically elegant than other accounts that have been proposed in the literature.

I then argue that if metanormativism about decision theory is true, then we have strong grounds for preferring the causal approach to decision theory over the evidential approach. In section iv, I show that decision-theoretic uncertainty undermines the intuitive case for evidential decision theory over causal decision theory. In section v, I show that decision-theoretic uncertainty gives us the resources to construct a counterexample to the “Why Ain’cha Rich?” argument in favor of evidential decision theory.

1. NEWCOMB’S PROBLEM

Newcomb’s problem is typically introduced through the following case:

*Standard Predictor*

You have two boxes in front of you, Box A and Box B. Box A is opaque; Box B, transparent. You have the option to take either Box A only, or both B and A. You can see that Box B contains $1,000. Box A either contains $1 million or $0. Moreover, someone (“The Predictor”) with an amazing ability to predict other people’s actions had control over the boxes. If the Predictor predicted that you would choose Box A only, then he put $1 million into Box A. If the Predictor predicted that you would choose both boxes, then the Predictor put nothing into Box A. What should you do?

Representing the decision problem in a table, we have:

<table>
<thead>
<tr>
<th></th>
<th>Money in both boxes</th>
<th>Money in one box only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take one box only</td>
<td>$1,000,000</td>
<td>$0</td>
</tr>
<tr>
<td>Take both boxes</td>
<td>$1,001,000</td>
<td>$1,000</td>
</tr>
</tbody>
</table>

There are two distinct but each seemingly compelling available lines of reasoning. First, I could reason that if I take Box A only, then I am almost certainly going to get $1 million. In contrast, if I take both boxes, I am almost certainly going to get only $1,000. So I should take Box A only. Such reasoning motivates *evidential decision theory* (EDT). According to EDT, one should choose the option with the maximal evidential expected value.
where the evidential expected value of an action is defined as the sum, over all possible outcomes, of the value of the outcome, given that you perform that action, multiplied by the probability of the outcome conditional on you performing that action. According to this account, in Standard Predictor, you should one-box. The precise formalization of this view will not matter for the purposes of this paper, but one simple way to formalize the view is as follows:

\[ EEV(A) = \sum_{i=1}^{n} C(O_i | A) V(O_i & A) \]

In the above equation, \( C \) is the decision maker’s credence function, and \( A, B, C, \) and so on are actions that are available to the decision maker. \( O_1, O_2, \ldots, O_n \) are propositions that describe a way the world might be. \( V \) is the decision maker’s value function. For any outcome \( O_i \), \( V(O_i) \), takes a real number that measures how valuable \( O_i \) is to the decision maker.

The above gave a line of reasoning that favored evidential decision theory. But there is an alternative line of reasoning. I could reason that the Predictor has already put the $1 million in Box A, or that the Predictor decided against doing so. My choosing both boxes cannot change that. Additionally, no matter what amount of money is in Box A, I will get an additional $1,000 if I take both boxes. So I should take both boxes. Such reasoning motivates causal decision theory (CDT). According to CDT, one should choose the option with the highest causal expected value, where the causal expected value (CEV) of an action is defined as the sum, over all outcomes, of the value of that outcome multiplied by the probability of the outcome counterfactually conditional on one’s action. There are many ways to formalize CDT, but these will not matter for my purposes, so I will use the following simple formulation:

\[ CEV(A) = \sum_{i=1}^{n} C(A \Rightarrow O_i) V(O_i & A) \]

In this equation, ‘\( \Rightarrow \)’ denotes the counterfactual conditional—that is, a conditional of the form, “If I were to perform \( A \), \( O_i \) would happen.” According to this account, in Standard Predictor you should two-box.

Different people’s intuitions vary strongly in response to the Standard Predictor. So in attempting to adjudicate between causal and evidential decision theory, other cases are normally used. But before moving on to them, I will introduce and motivate the idea of meta decision theory.

II. META DECISION THEORY

Given the trenchant disagreement between intelligent and well-informed philosophers, it seems highly plausible that one should not be certain in either causal or evidential decision theory. In light of this fact, Robert Nozick

briefly raised an interesting idea: that perhaps one should take decision-theoretic uncertainty into account in one’s decision making. He noticed that our intuitions in Newcomb problems seem to be stakes-sensitive. That is, it seems that we can generate clear counterexamples to both EDT and CDT simply by playing around with the Standard Predictor case. By altering the stakes, we can alter our intuitions. Consider, first, the following case:

*High-Stakes Predictor I (HSP-I)*

Box A is opaque; Box B, transparent. If the Predictor predicts that you choose Box A only, then he puts one wish into Box A. With that wish, you would save the lives of one million terminally ill children. If he predicts that you choose both Box A and Box B, then he will 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.

Representing this in a table:

<table>
<thead>
<tr>
<th></th>
<th>Wishes in both boxes</th>
<th>Wishes in one box only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take one box only</td>
<td>1,000,000 lives</td>
<td>Nothing</td>
</tr>
<tr>
<td>Take both boxes</td>
<td>1,000,000 lives + gum</td>
<td>Gum</td>
</tr>
</tbody>
</table>

In this case, intuitively, should you one-box or two-box? Although it can be difficult not to let theory cloud one’s judgment, my intuitive view is clearly that if someone two-boxes in that case, they have made the wrong decision. So do we have a slam-dunk argument in favor of EDT? Unfortunately not. Consider the following case:

*High-Stakes Predictor II (HSP-II)*

Box C is opaque; Box D, transparent. If the Predictor predicts that you choose Box C only, then he puts one wish into Box C and also a stick of gum. With that wish, you save the lives of one million terminally ill children. If he predicts that you choose both Box C and Box D, then he will put nothing into Box C. Box D—transparent to you—contains an identical wish, also with the power to save the lives of one million children, so if one had both wishes, one would save two million children in total. However, Box D contains no gum. One has two options only: Choose Box C only, or both Box C and Box D.

---

9 Nozick, *The Nature of Rationality*, op. cit., pp. 43–50, although Toby Ord and I, in conversation, independently came up with this idea before discovering that Nozick had suggested it.

10 This example and the next are structurally the same as examples given in Nozick, *The Nature of Rationality*, op. cit. I have just altered them a little bit to make the case even stronger.
Representing this in a table:

<table>
<thead>
<tr>
<th>Take one box only</th>
<th>Wishes in both boxes</th>
<th>Wishes in one box only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take both boxes</td>
<td>2,000,000 lives + gum</td>
<td>1,000,000 lives</td>
</tr>
</tbody>
</table>

In this case, intuitively, should you one-box or two-box? My intuitive view is clear: If someone one-boxes in the above case, they have made the wrong decision.

What is going on in these two cases? From one perspective, they are structurally identical. In both cases, EDT recommends one-boxing, because one-boxing has the higher evidential expected value. In both cases, CDT recommends two-boxing, because two-boxing has the higher causal expected value (and, indeed, dominates one-boxing). From another perspective, however, they are very different. In HSP-I, one’s decision is of huge consequence, according to EDT. From its perspective, the difference in value between one-boxing and two-boxing is the difference in value between saving a million innocent lives and getting a free stick of gum. For CDT, however, one’s decision in HSP-I is fairly trivial. The decision about whether to one-box or two-box is merely the decision about whether to get a free stick of gum or not. In contrast, in HSP-II, the decision is of huge consequence for CDT. The decision between one-boxing and two-boxing is the decision about whether to save a million innocent lives. For EDT, however, the decision in HSP-II is fairly trivial: It merely concerns whether to get a free stick of gum or not.

As Nozick noticed, this sort of stakes-sensitivity is suggestive of the idea that our intuitions are governed at least in part by uncertainty over both CDT and EDT. We feel the force of both sorts of decision theory, and so we have credence in both of them. Then, when making decisions, we hedge our bets, going with CDT when the relative stakes are sufficiently high for CDT, and going with EDT when the relative stakes are sufficiently high for EDT.11 I call this idea Meta Decision Theory (MDT).12 According to MDT, one should maximize meta expected value, where the meta expected value (MEV) of an action is defined as the

---

11 Of course, this is not the only possible explanation for why our intuitions switch in the two cases. In sections ii and iii, I consider and ultimately reject alternative explanations of this phenomenon.

12 A terminological clarification: I will use Meta Decision Theory (capital letters) or MDT to refer to the specific view that one ought to maximize expected choice-worthiness over decision theories. I will use meta decision theory to refer to any decision theory that claims that what one ought to do (in the relevant sense) is determined in part by one’s credences in first-order decision theories.
sum, over all decision theories, of the probability of that decision theory multiplied by the value of that action on that decision theory. Or, formally (and again with the caveat that there are many possible ways to formalize this idea):

\[ MEV(A) = \sum_{i=1}^{n} C(D_i)D_i(A) \]

In this formula, \( D_1, D_2, \ldots, D_n \) each refers to a decision theory, and \( D_i(A) \) is the value that \( D_i \) assigns to \( A \). In section V, I distinguish between the causal version of MDT and the evidential version of MDT. Until that point, however, the distinction will not matter for my purposes, so I state MDT simply in terms of unconditional probabilities.\(^{13}\)

On the reasonable assumption that we have at least small positive credence in each of EDT and CDT, MDT would make sense of the stakes-sensitivity suggested above. Because \( \text{HSP-I} \) has such higher stakes according to EDT than according to CDT, even very small credence in EDT would make one-boxing have the higher MEV. The same is true \textit{vice versa} for \( \text{HDP-II} \).

These high-stakes predictor cases make me think that some version of meta decision theory is true. Oddly, however, Nozick himself ultimately \textit{rejects} that idea in favor of a subtly different one. He says:

I suggest that we go further and say not merely that we are uncertain about which \textit{one} of these two principles, [CDT] and [EDT], is (all by itself) correct, but that both of these principles are legitimate and each must be given its respective due. The weights, then, are not measures of uncertainty but measures of the legitimate force of each principle. We thus have a \textit{normative} theory that directs a person to choose an act with maximal decision-value.\(^{14}\)

And also:

Theorists of rationality have been intent upon formulating the one correct and complete set of principles to be applied unreservedly in all decision situations. But they have not yet reached this—at any rate, we do not have complete confidence that they have. In this situation, won’t a

\(^{13}\) A couple of other notes on this. First, we should of course have non-zero credence in decision theories other than CDT and EDT, such as Benchmark Theory as described in Wedgwood, “Gandalf’s Solution to the Newcomb Problem,” \textit{op. cit.}, and so uncertainty about these other theories will also have to be taken into account. In order to keep things simple, however, I will leave these alternative decision theories to the side. Second, one might worry whether meta decision theory suffers from the problem of intertheoretic comparisons. However, the problem of intertheoretic comparisons is substantially easier in the case of EDT and CDT than it is between different moral theories. EDT and CDT both agree on what a decision maker should do in all of the many cases where \( cr(A \Rightarrow q) = cr(Alq) \). We can use this agreement to normalize the two theories.

\(^{14}\) Nozick, \textit{The Nature of Rationality}, \textit{op. cit.}, p. 45.
prudent and rational individual hedge her bets? I want to say more, namely, that no one of the principles alone is wholly adequate—it’s not simply that we have yet to find the knockdown argument for the one that is correct.\footnote{Ibid., pp. 46–47.}

That is, as I understand him, Nozick rejects what I call meta decision theory in favor of what might be called decision-theoretic pluralism (DTP).\footnote{The analogy is with pluralist moral theories. Someone who maximizes expected choice-worthiness under uncertainty about whether only well-being, or both knowledge and well-being, are of value looks a lot like someone who is conforming with a first-order moral theory that assigns value to both well-being and knowledge. In the same way, someone who follows MDT looks a lot like someone who is conforming with a first-order decision theory that gives weight to both casual expected value and evidential expected value.} Whereas MDT is not a rival to CDT or EDT, DTP is a rival first-order theory.

What is odd about Nozick’s suggestion is that, even though MDT seems to be the natural explanation of our stakes-sensitive intuitions, he gives no argument for preferring DTP to MDT (apart, perhaps, from the cryptic suggestion that MDT would not be “normative”). We already know that we are decision-theoretically uncertain and that expected utility theory is in general the best way to handle uncertainty. This is enough to make MDT plausible, and MDT is enough to explain our stakes-sensitive intuitions. There therefore seems to be nothing to gain by suggesting that DTP is true, so DTP seems unmotivated.

Moreover, DTP is not merely unmotivated: It also has two major problems that MDT lacks. First, DTP has multiple explanatory gaps. Why weigh EDT against CDT in one way rather than another? MDT has a principled answer to this—namely, that the weights are one’s credences—whereas DTP does not. And why should the values that EDT and CDT assign to acts be additively separable? Again, MDT offers an explanation for this—that taking an expectation requires values across states to be additively separable—whereas DTP does not. And, finally, why even think that there would be different sorts of “decision-theoretic value”? Decision-theoretic pluralism is very different from other sorts of pluralism about value: Typical pluralist theories make sense of different values because different values supervene on different sorts of stuff. In contrast, the different decision-theoretic values that Nozick suggests arise merely out of how uncertainty is taken into account. So Nozick’s account does not gain plausibility from the plausibility of pluralism about value in general.
Second, DTP misrepresents what is going on in the stakes-adjusted Newcomb cases. To see this, consider a variation on his cases.\textsuperscript{17}

\textit{Four-Box Predictor}

Box A and Box C are opaque; Box B and Box D, transparent. The Predictor has a 100\% success rate at predicting which box or boxes you will choose. You have the following four options:

1. Take A and C only
2. Take A, B, and C
3. Take A, C, and D
4. Take A, B, C, and D

If the Predictor predicts that you will take Box B, he will put nothing in Box A. If he predicts that you will not take Box B, he will put into Box A a wish with the value of one million children’s lives.

If the Predictor predicts you will take Box D, he will put nothing in Box C. If he predicts that you will not take Box D, he will put into Box C a wish with the value of one million children’s lives.

Box B—transparent to you—contains a stick of gum. Box D—also transparent to you—contains a wish with the value of one million children’s lives and also a stick of gum.

Representing this in a table:

<table>
<thead>
<tr>
<th></th>
<th>Wish in neither A nor C</th>
<th>Wish in A, but not C</th>
<th>Wish in C, but not A</th>
<th>Wish in both A and C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take A and C</td>
<td>Nothing</td>
<td>1,000,000 lives</td>
<td>1,000,000 lives</td>
<td>2,000,000 lives</td>
</tr>
<tr>
<td>Take A, B, and C</td>
<td>1 stick of gum</td>
<td>1,000,000 lives + 1 stick of gum</td>
<td>1,000,000 lives + 1 stick of gum</td>
<td>2,000,000 lives + 1 stick of gum</td>
</tr>
<tr>
<td>Take A, C, and D</td>
<td>1,000,000 lives + 1 stick of gum</td>
<td>2,000,000 lives + 1 stick of gum</td>
<td>2,000,000 lives + 1 stick of gum</td>
<td>3,000,000 lives + 1 stick of gum</td>
</tr>
<tr>
<td>Take A, B, C, and D</td>
<td>1,000,000 lives + 2 sticks of gum</td>
<td>2,000,000 lives + 2 sticks of gum</td>
<td>2,000,000 lives + 2 sticks of gum</td>
<td>3,000,000 lives + 2 sticks of gum</td>
</tr>
</tbody>
</table>

The astute reader might have noticed that someone in a \textit{Four-Box Predictor} situation is just someone who faces both \textit{HSP-I} and \textit{HSP-II} at the same time. The very astute reader might have noticed that this therefore constitutes a “Jackson case” under decision-theoretic uncertainty: a case in which one ought (in some sense) to do something that one knows one ought (in some other sense) not to do.\textsuperscript{18}

\textsuperscript{17} I thank Toby Ord for this suggestion.

According to CDT, one ought to perform act (4). According to EDT, one ought to perform act (2). But we should think that, in at least some sense of ‘ought’, what the decision maker ought to do is act (3). By performing (4) rather than (3), one risks losing the opportunity to save one million children for the sake of a stick of gum. (This was the motivation for one-boxing in HSP-I.) By performing (1) rather than (3), again one risks losing the opportunity to save one million children for the sake of a stick of gum. (This was the motivation for two-boxing in HSP-II.) And if one performs act (2) rather than (3), one takes both risks at the same time. So one should perform act (3) and take Boxes A, C, and D: That is the only safe bet. And it is the only choice that seems consistent with our intuitions in both HSP-I and HSP-II.

In the above situation, the correct thing to say, I think, is that, in some sense of ‘ought’ (the sense that first-order decision theories are talking about), one ought to perform either act (2) or act (4), but that, in another sense of ‘ought’ (the sense that is relative to decision-theoretic uncertainty), one ought to perform act (3). That is the appraisal that MDT gives of the situation. But that is not the appraisal that Nozick’s view gives. According to Nozick’s view, all there is to say is that one ought to perform (3), because that is what the true decision theory (that is, DTP) claims: It is simply false, in any sense, that one ought to choose either (2) or (4). And that seems to misrepresent what is really going on in our appraisal of the Four-Box Predictor.

For these reasons, for the rest of the paper I will set Nozick’s view to one side, and instead assume that MDT is the most plausible rational explanation of the stakes-sensitivity of our intuitions. Nozick quickly moved on from the suggestion, and as far as I know it has not been pursued elsewhere. But that is unfortunate because MDT has important implications that have not been noticed.

III. THE SMOKING LESION AND THE PSYCHOPATH BUTTON

First, MDT allows us to resolve an apparent conflict in our intuitions. My suggestion is that the divergence in our intuitions across cases in

---

19 Note, however, that Sepielli (“What to Do When You Don’t Know What to Do. . . .,” op. cit.) considers the view in the context of discussing the regress problem.

20 That is not to say that it can resolve the divergence in all of our intuitions. In particular, there is a class of cases that seem to me not to be about whether evidential or causal decision theory is true but rather whether the usual formulation of causal decision theory accurately captures the idea of “causing the best consequences.” I place Andy Egan’s time-travel and oracle cases (“Some Counterexamples to Causal Decision Theory,” op. cit.) in this category, as well as Arif Ahmed’s Nomological Gamble example (“Causal Decision Theory: A Counterexample,” op. cit.). I take these examples to violate the letter of CDT but not its spirit (although Ahmed goes on to argue that there is no way of formalizing CDT so that it gets the right answer in his case and is compatible with free choice; that is an interesting, but very different, argument).
the literature can be understood as hedging between EDT and CDT, in a way that is mandated by MDT.

Consider *The Smoking Lesion*:21

*The Smoking Lesion*

Susan is debating whether or not to smoke. She believes that smoking is strongly correlated with lung cancer, but only because there is a common cause—a lesion that tends to cause both smoking and cancer. Once we fix the presence or absence of this condition, there is no additional correlation between smoking and cancer. Susan prefers smoking without cancer to not smoking without cancer, and she prefers smoking with cancer to not smoking with cancer.

In this case, intuitively, Susan should smoke. But, problematically, EDT recommends against smoking. *The Smoking Lesion* has been taken to be a fatal counterexample to EDT. However, if our intuitions are explained in part by MDT, then our intuition regarding *The Smoking Lesion* should change simply if we alter the stakes. And it seems that it does. Consider the following case:

*Stakes-Adjusted Smoking Lesion*

The lesion is not correlated with mere lung cancer. Rather, the lesion causes people both to smoke before they are 35 and to burst into flames on their 35th birthday, enduring several hours of agony before dying (even though smoking does not cause the spontaneous self-combustion). Moreover, let us suppose that Susan is not really that fussed about smoking. She has not been inclined to smoke previously, but she is feeling whimsical today and so has a slight preference for smoking that cigarette. It is the day before her 35th birthday. Should she smoke?

In this case, it seems very clear to me, intuitively speaking, that Susan should not smoke, even though CDT would recommend smoking.22 Simply by altering the stakes, we have transformed an apparent counterexample to EDT into an apparent counterexample to CDT. This stakes-sensitivity is exactly what MDT would predict.

Next, consider *The Psychopath Button*:23

---

21 Here I use the formulation given in Egan, “Some Counterexamples to Causal Decision Theory,” *op. cit.*, p. 94.

22 Some causal decision theorists I have spoken with have bitten the bullet in this case. But I have a very hard time believing that such a response is genuinely a basic intuition, rather than a judgment that has been tainted by one’s theoretical commitments.

23 Again I just modify slightly the formulation given in Egan, “Some Counterexamples to Causal Decision Theory,” *op. cit.*, p. 97. This case was initially presented in Egan’s paper, but it was suggested to him by David Braddon-Mitchell.
The Psychopath Button

Paul is debating whether to press the “kill all psychopaths” button. It would, he thinks, be much better to live in a world with no psychopaths. And Paul is almost certain that he is not a psychopath. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world with psychopaths to dying. Should Paul press the button?

In this case, intuitively, Paul should not press the button. But, problematically, according to CDT, Paul should press the button. Again, however, if MDT explains our intuitions, then our intuitions about The Psychopath Button should be stakes-sensitive. And it seems that they are. Consider the following modification of the case:

Stakes-Adjusted Psychopath Button

It is 1890, and Paul knows that Baby Hitler is a psychopath, knows that only one other person (who may be him) is a psychopath, and knows of the atrocities that will happen in the following 60 years if Hitler survives. Moreover, let us suppose that Paul has a terminal illness. He will surely die within a few hours. He wants to have those last few hours alive and, being a selfish sort of person, mildly prefers having those hours to killing Hitler. However, because it is so morally important to kill Hitler, he only has a very mild preference for living those few hours at the cost of Hitler’s survival. Now, what should Paul do?

It seems very plausible to me that he should push the button. But, if so, then, again, our intuitions about this case have switched merely by altering the stakes involved. Again, this is exactly what MDT predicts.

We can make this explanation more precise. Let us define the relative stakes ratio, in two option cases, as the ratio of (EEV(right action according to EDT)—EEV(wrong action according to EDT)) to (CEV(right action according to CDT)—CEV(wrong action according to CDT)). If MDT is correct, then it is how this ratio changes that should affect our intuitions.

Now, I would personally be roughly indifferent between a guarantee of $1,000 and 1% chance of $1 million, so, given my preferences, the relative stakes ratio in Standard Predictor is approximately 99:1. The Smoking Lesion is taken to be more favorable to CDT than Standard Predictor is. So, if our intuitions roughly track MDT’s recommendations, then we should expect the relative stakes ratio to be less than 99:1. And that is what we find. If I ask myself, for example, whether Susan would be willing to take up smoking even at the cost of causing a 1% chance of moving from the low-risk group for lung
cancer (to which non-smokers belong) to the high-risk group for lung cancer (to which regular smokers belong), I imagine her being willing to take that cost. I imagine her only becoming indifferent at around 10%, suggesting that, when presented with the case, I intuitively assess the relative stakes ratio as only being about 10:1. This is in line with MDT’s prediction.

*The Psychopath Button* is taken to be more favorable to EDT than *Standard Predictor* is. So, if our intuitions roughly track MDT’s recommendations, then we should expect the relative stakes ratio in *The Psychopath Button* to be greater than 99:1. And that is what we find. If I ask myself, for example, whether Paul would be willing to kill all psychopaths even at the cost of a 1% chance of causing his own death (perhaps he has a gun with 99 empty chambers but one loaded chamber, and pointing the gun at his head and pulling the trigger is the only way to kill all psychopaths), I imagine him not being willing to take that risk. I certainly would not do it. Even if I thought it was okay to murder innocents for a greater good (!), and even if I thought that killing all psychopaths would be a net good, I still value my own life too much to make that sort of sacrifice. But if that is correct, then I have intuitively judged the relative stakes ratio to be greater than 99:1, which is what MDT predicts.24

24 Egan gives another case, called *The Murder Lesion*. It seems to me that, again, the reason our intuitions in this case favor EDT is because of the relative stakes. However, the relative stakes do not seem to be quite as biased toward EDT as they are in *The Psychopath Button*. This might explain why *The Murder Lesion* is not as convincing a counterexample to CDT as *The Psychopath Button* is. MDT allows us to explain a couple of other puzzles as well. First, James Joyce (“Regret and Instability in Causal Decision Theory,” *Synthese*, clxxxvii, 1 (2012): 123–45, at p. 125) says that *The Psychopath Button* “is not original with Egan” because structurally similar cases were given in Paul Weirich, “Decision Instability,” *Australasian Journal of Philosophy*, lxxiii, 4 (1985): 465–72; Allan Gibbard, “Weakly Self-Ratifying Strategies: Comments on McClennen,” *Philosophical Studies*, lxv, 1–2 (1992): 217–25; and Judea Pearl, “The Curse of Free-Will and the Paradox of Inevitable Regret,” *Journal of Causal Inference*, 1, 2 (2013): 255–57. Joyce takes the shared structural similarity to be that all are cases where every option (prima facie) is unratifiable according to CDT. But, if this is right, then why weren’t these earlier cases taken to be grave counterexamples to CDT, in the way that some at least have taken *The Psychopath Button* to be? The answer lies with the stakes. *The Psychopath Button* is not similar to the earlier cases with respect to the relative stakes ratio. And it is the relative stakes ratio that gives *The Psychopath Button* its bite. Second, EDT allows us to explain why our intuitions in *Standard Predictor* seem to favor EDT significantly more if the Predictor is infallible, rather than merely highly accurate (as Jordan Howard Sobel discusses in “Infallible Predictors,” *The Philosophical Review*, xcvi, 1 (1988): 3–24). The answer is twofold. First, increasing the probability of the $1 million further biases the stakes in EDT’s favor. Second, as a matter of psychology, we tend to overvalue a “sure thing” (which is why, for example, the Allais paradox arises), so the move from 99% certainty to 100% certainty biases the stakes in favor of EDT by considerably more than merely the value of an additional 1% chance of $1 million.
One might object that, in the “high-stakes” and “stakes-adjusted” cases given above, we can explain the divergence in our intuitions by appealing to empirical uncertainty. According to this explanation, in HSP-I, we get the one-boxing intuition because we cannot really imagine ourselves to be certain that the Predictor will get it right purely through prediction. In any situation we can imagine, so the objection goes, there will remain some residual uncertainty that choosing the one box causes there to be a wish in the one box, and that is how the Predictor pulls off his trick. Similarly, in Stakes-Adjusted Smoking Lesion, perhaps we simply cannot imagine ourselves not to have some credence that smoking causes bursting into flames. In either case, if we have even small credence in that empirical hypothesis, then both CDT and EDT will recommend one-boxing in HSP-I and not smoking in the Stakes-Adjusted Smoking Lesion. One can attempt an analogous explanation with respect to HSP-II and The Stakes-Adjusted Psychopath Button.

Speaking personally, my intuitions are sufficiently robust that we could replace the stick of gum with the lives of ten thousand children and I would have the same view that one should one-box in HSP-I and two-box in HSP-II. (The same is true for adjustments to the stakes in Stakes-Adjusted Smoking Lesion and Stakes-Adjusted Psychopath Button.) Given this, empirical uncertainty does not seem to be a very good explanation of my intuitions. However, there is a stronger response, which is that these extreme cases are not strictly necessary to the use of MDT as an explanatory hypothesis for why our intuitions favor CDT in The Smoking Lesion and EDT in The Psychopath Button. All we need to show is that the relative stakes are more heavily biased toward CDT in The Smoking Lesion than they are in the Standard Predictor and are more heavily biased toward EDT in The Psychopath Button than they are in the Standard Predictor. And that is exactly what my discussion of the relative stakes ratio accomplished.

So MDT seems to do well in terms of giving a rational grounding for our seemingly conflicting intuitions. In fact, I think that it is the best account of our intuitions in these cases that I know of.

Consider, in contrast, the response to The Psychopath Button suggested by James Joyce, as part of a defense of CDT.\(^{25}\) The idea is that, as Paul decides to perform one action rather than another, he

\(^{25}\) Joyce, “Regret and Instability in Causal Decision Theory,” *op. cit.* His view is very similar to that of Arntzenius (“No Regrets,” *op. cit.*), and both draw heavily on work by Brian Skyrms, *The Dynamics of Rational Deliberation* (Cambridge, MA: Harvard University Press, 1990). What I will say in response to Joyce applies fairly straightforwardly to Arntzenius’s view.
immediately gains evidence about whether he is a psychopath. Given his initial credences, pushing the button has the higher causal expected value. But as soon as he begins to decide to push the button, he gains evidence that he is a psychopath, and his credences should change. And with those new credences, not pushing the button has the highest causal expected value. But as soon as he begins to decide to not push the button, he gains evidence that he is not a psychopath, and suddenly pushing the button has the higher causal expected value again. Eventually, his credences over whether he is a psychopath or not end up in equilibrium, with the expected causal value of both pushing the button and of not pushing the button being the same.

This response is interesting. However, it seems to me that my explanation of our intuitions is significantly better than Joyce’s. This is for three reasons, presented in order of increasing importance.

First, Joyce’s account cannot explain our intuitions in HSP-I and the Stakes-Adjusted Smoking Lesion. In order to explain our intuitions in those cases, he would have to appeal to some other explanation. In contrast, my account can both take the intuitions at face value, rather than having to offer a speculative debunking argument, and offer one unified explanation for our varying intuitions, rather than having to offer two distinct explanations.

Second, his account does not get the intuitions right in very similar cases. Suppose, for example, that the button is wired up to Paul’s brain, so that as soon as he begins to intend to push the button, all psychopaths are killed. His beliefs therefore are not able to achieve equilibrium. In this case, CDT really would recommend that he intend to push the button. But it seems that this minor alteration to the case does not affect our intuitive appraisal of what Paul should intend to do.

Third and finally, Joyce’s account does not capture the intuition even in the original case. Once deliberational equilibrium is reached, pushing the button has the same expected value as not pushing the button. But that is not capturing the intuition, which is clearly in favor of it being a mistake to push the button, rather than it being permissible to push the button. Joyce makes some very brief suggestions, based on the heuristics and biases literature, concerning why we might think that the intuition is not reliable in this case, and, in general, I am perfectly happy to sacrifice fit with the intuitive data for the sake of theoretical elegance. But if we have an independently motivated explanation of why those intuitions are rational, then we should prefer

---

26 Of course, my explanation is not inconsistent with Joyce and Arntzenius’s explanation. What I am questioning is not whether their account is true but whether it is a satisfactory explanation of our divergent intuitions across these cases.
that explanation to the debunking explanation, unless the debunking explanation is on very strong ground indeed. So we should prefer MDT’s explanation to Joyce’s.

Joyce’s is not the only alternative explanation in town. Ralph Wedgwood has introduced “Benchmark Theory,” which gets the right answer in both *The Smoking Lesion* and *The Psychopath Button*. But it suffers from intuitive counterexamples, too. Johan Gustafsson proposes a decision theory that captures the intuitions in both *The Smoking Lesion* and *The Psychopath Button* cases. But that proposal comes at a cost of considerable theoretical inelegance: It imports an idea of *iterated general ratifiability* that does not seem independently motivated. Another potential explanation comes from Huw Price, who suggests that we understand causality in subjectivist terms, so that evidential probability and causal probability are, despite appearances, the same. Again, however, this comes at major theoretical cost, depending on the truth of particular positions in the metaphysics of both causation and free will. And if an alternative account explains our divergent intuitions without using such heavy philosophical machinery, as the MDT account does, then we should prefer that alternative explanation.

In general, we already know (i) that we are decision-theoretically uncertain and (ii) that expected utility theory is in general the best way to accommodate uncertainty. So even independent of its ability to explain our divergent intuitions, we should think that there is an argument for thinking that MDT is true. It explains our divergent intuitions without using any ad hoc philosophical machinery. There is therefore a strong argument via Occam’s razor for preferring the MDT explanation to any other explanation that does not have the same independent plausibility.

So I think that MDT provides the best explanation of our apparently inconsistent intuitions. Now let us turn to further implications of this

---

27 Wedgwood, “Gandalf’s Solution to the Newcomb Problem,” *op. cit.*
28 See, for example, Briggs, “Decision-Theoretic Paradoxes as Voting Paradoxes,” *op. cit.*
29 Gustafsson, “A Note in Defence of Ratificationism,” *op. cit.*
31 Price himself acknowledges this when he says the following: “As we have seen, the EviCausalist relies heavily on the idea that the epistemic viewpoint of an agent is distinctive in certain ways. Roughly, it requires that agents see their own actions as ‘uncaused,’ at least in the midst of deliberation about those actions. This not only binds the fate of the EviCausalist, at least in some sense, to that of free will; it also means, potentially even more uncomfortably, that EviCausalism becomes a rope that binds causation to the fate of free will—no problem, perhaps, if these notions turn out to share the same fate, but a problem if they do not.” *Ibid.*, p. 536 (italics in the original).
view and see how it gives grounds to undermine the two best arguments in favor of EDT.

IV. UNDERMINING THE INTUITIVE ARGUMENT FOR EDT

One way to argue in favor of EDT is via appeal to cases. EDT looks appealing for people like me, who think that you should one-box in the standard Newcomb problem, and for those who are particularly concerned by *The Psychopath Button*. So it looks like EDT is at least fairly well supported by the intuitive data.

Considerations of decision-theoretic uncertainty undermine this argument. The relative stakes ratio in the standard Newcomb problem depends on one’s level of risk-aversion with respect to money, but for any normal agent, it is heavily biased in favor of EDT. For me, the relative stakes ratio is approximately 99:1, so if I had only 1.1% or higher credence in EDT, then, by MDT’s lights, I should one-box in the standard Newcomb problem. So, far from providing an argument for thinking that the evidential approach is the best approach, the intuitions merely show that we should have at least a small credence in EDT. Indeed, because intuitions in the standard Newcomb case are unclear, and because *The Smoking Lesion* favors CDT even though in that case the stakes are still biased toward EDT, it seems that the credence in EDT that is warranted by appeal to intuitions about particular cases is not very large at all.

The fair way to adjudicate, on intuitive grounds, between EDT and CDT would be to consider cases where the stakes are evenly balanced. Such a case would look as follows:

<table>
<thead>
<tr>
<th></th>
<th>Money in both boxes</th>
<th>Money in one box only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Take one box only</td>
<td>$20</td>
<td>$0</td>
</tr>
<tr>
<td>Take both boxes</td>
<td>$30</td>
<td>$10</td>
</tr>
</tbody>
</table>

Even I—who used to self-identify as a stark-raving one-boxer—get almost no intuition in favor of one-boxing in this case. So EDT is not the intuitive view. In fact, I only start to get one-boxing intuitions once the amount that might be in the opaque box is 20 times as great as the amount that is certainly in the transparent box. So, as far as the argument from intuition

---

32 I use small amounts of money so that we can safely assume that utility is approximately linear with respect to money in this case.

33 Anecdote: When I was first presented with *The Smoking Lesion* case, I thought it was supposed to be an argument in favor of EDT.
goes, I should have no more than a small credence in EDT. So the intuitive argument for EDT is far weaker than it first seemed.

Appeal to intuitions has been used as one major argument in favor of EDT. The other argument is the “Why Ain’cha Rich?” argument. Let us consider that now.

V. A COUNTEREXAMPLE TO “WHY AIN’CHA RICH?”

When I introduced meta decision theory, I used unconditional credences. But we could formulate both evidential and causal versions of meta decision theory. According to causal meta decision theory (CMDT), we should maximize causal meta expected value (CMEV) where:

\[ CMEV(A) = \sum_{i=1}^{n} C(A \Rightarrow D_i) D_i(A) \]

Again using ‘⇒’ to denote the counterfactual conditional. It should be clear that for all \( A, D \), \( \Pr(A \Rightarrow D) = \Pr(D) \). Acting one way rather than another cannot affect which decision theory is true, so nothing is lost by simply using unconditional credences:

\[ CMEV(A) = \sum_{i=1}^{n} C(D_i) D_i(A) \]

In contrast, according to evidential meta decision theory (EMDT), we should maximize evidential meta expected value (EMEV), where:

\[ EMEV(A) = \sum_{i=1}^{n} C(D_i|A) D_i(A) \]

These two views will almost never come apart: It is a very rare situation when acting one way or another gives you evidence for one decision theory rather than another. But it is not impossible for the two to come apart. And if we look at those admittedly rare cases, we can construct a counterexample to the “Why Ain’cha Rich?” argument.

According to the “Why Ain’cha Rich?” argument, the average return of one-boxing exceeds the average return of two-boxing. Moreover, everyone can see that the average return of one-boxing exceeds the average return of two-boxing, so one-boxing foreseeably gives us more of what we want than two-boxing does. And, so the argument goes, a decision theory cannot be correct if it recommends an option that foreseeably gives you less of what you want than some other option does. Therefore, CDT cannot be correct.

In response, the defender of CDT can say that Newcomb’s cases are unusual: These are cases in which a devious person has set things up to
reward irrational behavior. So it is not surprising that irrational people like those who act in accordance with EDT end up richer. However, to date, the defender of CDT has not been able to come up with a convincing case where one gets rewarded for not following EDT.\textsuperscript{34} And that seems problematic.

However, if we are comparing CDMT and EMDT, things are different. Once we allow decision-theoretic uncertainty into the picture, we can construct a case in which performing the action EMDT recommends foreseeably makes one poorer. So there is no longer an asymmetry between the causal and evidential approach, and the “Why Ain’cha Rich?” argument loses its force. Here is the case:

*The Meta Newcomb Problem*

Sophie faces two boxes, as follows:

<table>
<thead>
<tr>
<th>Wishes in both boxes</th>
<th>Wishes in one box only</th>
</tr>
</thead>
<tbody>
<tr>
<td>One box</td>
<td>2 million lives\textsuperscript{35}</td>
</tr>
<tr>
<td>Two box</td>
<td>3 million lives</td>
</tr>
</tbody>
</table>

Sophie’s beliefs are as follows. She has 51\% credence in EDT and 49\% credence in CDT. Before taking her action, she is almost certain that there are wishes in both boxes. However, *conditional* on her two-boxing, she is almost certain that there is a wish only in the transparent box.

Given these credences:

\[
V_{\text{EDT}}(\text{One Box}) = -1 \times 2 \text{ million lives} + 0 \times 0 \text{ lives} = -2 \text{ million lives}
\]

\[
V_{\text{EDT}}(\text{Two Box}) = 0 \times 3 \text{ million lives} + -1 \times 1 \text{ million lives} = -1 \text{ million lives}
\]

\[
V_{\text{CDT}}(\text{One Box}) = -1 \times 2 \text{ million lives} + 0 \times 0 \text{ lives} = -2 \text{ million lives}
\]

\[
V_{\text{CDT}}(\text{Two Box}) = -1 \times 3 \text{ million lives} + 0 \times 1 \text{ million lives} = -3 \text{ million lives}
\]

So the meta decision problem looks as follows:

<table>
<thead>
<tr>
<th></th>
<th>Value, given EDT</th>
<th>Value, given CDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>One box</td>
<td>~2 million</td>
<td>~2 million</td>
</tr>
<tr>
<td>Two box</td>
<td>~1 million</td>
<td>~3 million</td>
</tr>
</tbody>
</table>

\textsuperscript{34} David Lewis (“Why Ain’cha Rich?,” *Noûs*, xv, 3 (1981): 377–80) argues that it is impossible. In his introduction to decision theory, Weatherson summarizes the literature on this as follows: “It turns out to be very hard, perhaps impossible, to construct a problem of this sort for evidential decision theorists” (p. 89; available at brian.weatherson.org). Arntzenius (“No Regrets,” *op. cit.*.) has proposed an example, but it is debatable whether the example is successful or is even coherent. For discussion of that case, see Arif Ahmed and Huw Price, “Arntzenius on ‘Why Ain’cha Rich?’,” *Erkennnis*, s, 1 (2012): 15–30.

\textsuperscript{35} Here I use “lives saved” rather than dollars, because linear value over number of lives saved is more plausible than linear value over dollars.
However, Sophie places great weight, epistemically, on what people actually do in Newcomb cases (rather than on what people claim their intuitions are about what they would do in such cases). She thinks that the actions of typical human agents in Newcomb cases provide very good evidence in favor of CDT or EDT. And she believes she is a typical human agent. So how she acts will affect her credences in the two decision theories. If she one-boxes, she will update in favor of EDT and will come to have 52% credence in EDT and only 48% in CDT. If she two-boxes, she will significantly update in favor of CDT and come to have 60% credence in CDT and 40% credence in EDT.

What should Sophie do? To answer this, let us work out the expected values. We have:

\[
V_{CMDT}(\text{One Box}) = 0.51 \times 2 + 0.49 \times 2 = 2 \\
V_{CMDT}(\text{Two Box}) = 0.51 \times 1 + 0.49 \times 3 = 1.98
\]

So CMDT recommends one-boxing. And we have:

\[
V_{EMDT}(\text{One Box}) = 0.52 \times 2 + 0.48 \times 2 = 2 \\
V_{EMDT}(\text{Two Box}) = 0.4 \times 1 + 0.6 \times 3 = 2.2
\]

So EMDT recommends two-boxing.

Thus, if we take into account decision-theoretic uncertainty, then the causal theory can tell you to one-box while the evidential theory tells you to two-box. In the above case, if Sophie follows EMDT, she foreseeably ends up saving fewer lives than if she follows CMDT. So, in contrast with the debate between EDT and CDT, one cannot construct a “Why Ain’cha Rich?” argument in favor of EMDT over CMDT.

This makes it seem very plausible that the correct meta decision theory is causal. But we can go a bit further than that: It would seem odd for the correct meta decision theory to be causal while the correct first-order decision theory is evidential. It seems plausible that our views about which variety of first-order decision theory is correct and which variety of meta decision theory is correct should be at least roughly coherent. So evidence about which meta decision theory is true seems also to give evidence about which first-order decision theory is true. So, even if we cannot construct a counterexample to “Why Ain’cha Rich?” for EDT, the fact that we can construct such a counterexample for EMDT weakens, at least to some degree, the “Why Ain’cha Rich?” argument in favor of EDT.

As well as being able to provide an explanation of our divergent intuitions across cases, considerations relating to meta decision theory
allow us to generate novel arguments against EDT. Thus, meta decision theory (and the idea of metanormativism more generally) seems to be a powerful tool in the causal versus evidential debate.

VII. CONCLUSION

In this article, I have argued that meta decision theory has two important implications. First, it can explain the apparent divergence in our intuitions between the Standard Predictor, The Smoking Lesion, and The Psychopath Button. Second, it undermines both the intuitive argument in favor of EDT and, to some extent, the “Why Ain’cha Rich?” argument as well. Considerations of decision-theoretic uncertainty are therefore a powerful tool for use in debates between causal and evidential decision theory—a tool that gives the causal approach a significant new advantage.

WILLIAM MACASKILL

University of Oxford