The Halo of Uncertainty

Desire, Belief and Moral Newcomb Problems

Brian Weatherson

March 26, 2019

Contents

1 Lewis's Argument **5**
1.1 Introduction 5
1.2 Four Versions of Max 10
1.3 Plan for Book 13

2 Moral Uncertaintism **15**
2.1 Naming The Topic 15
2.2 Against Moral Uncertaintism 19
2.3 Guise of the Good 21

Chapter 1

Lewis's Argument

1.1 Introduction

In a pair of papers from the second half of his illustrious career, David Lewis argued against a thesis he called "Desire as Belief". The papers had the somewhat unimaginative names "Desire as Belief" (Lewis, 1988) and "Desire as Belief II" (Lewis, 1996). The arguments have proven to be rather perplexing. There is nothing like a consensus in the literature on how to formulate the argument, on what its strengths and weaknesses are, or even what the argument would show if it succeeded.

But let's start with a reasonably simple version of the argument. Assume Max is a rational agent, who can be accurately represented as having credence function C and value function V. That is, Max's preferences are such that he always prefers the option with the highest expected value for V, given his credence function C. Now if you know almost anything about philosophical work on decision theory over the last 50 years, you'll know that this assumption contains an ambiguity. When we say Max optimises expected utility, are we understanding that the way that evidential decision theorists do, or the way that causal decision theorists do, or perhaps some other way?[1] Lewis says in a parenthetical paragraph that how we resolve this ambiguity isn't going to matter.[2] Collins (2015) pointedly disagrees, and I am going to adopt Collins's view wholeheartedly in what follows. But set that aside for now, and just say that Max's utility function is V and credence function is C.

[1] For good treatments of these two ideas, albeit from partisans of the causal side of the debate, see Weirich (2012) and Joyce (1999).

[2] Include reference

Assume also, and this really is just for simplicity, that Max's utility function is binary in the following sense. For every maximally specific possibility, it either is Good, and gets value 1, or is Bad, and gets value 0. I will relax this assumption at times, but it really helps ease the presentation if we start with this simple case.

Now we make the big assumption about Max. His desires correlate with his beliefs about the good. Here's how we're going to represent that. For any factual proposition A, we'll assume that there is another proposition A°, to be read as *A is good*. (And I'll come back to what I mean by saying something is a factual proposition. For now, all that matters is that I am not assuming that the operator $^\circ$ can be applied iteratively. I don't assume that Max has attitudes towards, or even the conceptual resources to think about, a proposition we might write as $A^{\circ\circ}$, i.e., it is good that it is good that A.) In the first paper, Lewis had used rather than A°, but most fonts contain only a few letters with this kind of overring, so using a degree symbol after the letter is more convenient. In the second paper, Lewis calls this the halo function, which explains half the title of this book. Anyway, Max can think thoughts about these propositions A°. And what he thinks is that the amount that he values A exactly tracks how probable he thinks it is that A is good. In symbols, Max endorses:

Desire as Belief (DaB) $V(A) = C(A^\circ)$

I am going to have a lot to say about the different ways we might interpret this equation, but for now it's just an equation. Crucially, Max is stably disposed to conform to this equation. In particular, even after learning new information, Max will still satisfy this equation, for all A.

Sadly for Max, and especially sadly for those philosophers who think that ideal agents should be like Max and conform to this equation, it is irrational to consistently hold onto it. The proof I'm about to give of this is not the complicated argument that Lewis gives in (1988), but a slightly expansive version of the argument he gives in (1996). Lewis notes the argument is similar to one offered by Arló Costa et al. (1995). My presentation draws on the presentations in Russell and Hawthorne (2016) and Collins (2015). As well as using V and C for Max's value and credence functions, I'll use subscripts for updating. That is, I'll write V_A for Max's value function after he updates with the information that A, and C_A for his credence function after he updates with the information that A. Whenever I use a single uppercase letter for the name of a proposition (or similar

thing that Max has attitudes towards), that will be a factual proposition (in a sense of factual I'll come back to far below).

Finally, I'll use w to either denote a single 'world', or as a variable ranging over worlds. I'll say a lot more about worlds in what follows, but for now note that they are things that set the truth value of all propositions that are currently relevant. So they are more fine-grained in some senses, and less fine-grained in others, than Lewisian concreta. They are more fine-grained because we could have a pair of worlds, in this sense, that are alike in descriptive respects, but unlike in evaluative respects. I will need these worlds because I will try to model agents who are uncertain about evaluative claims, but whose actions are sensitive to their beliefs about the evaluative. They are less fine-grained because they only determine the truth values of things that are currently relevant. They are like the 'small worlds' in the sense of Savage (1954). They don't determine last night's baseball results, let alone the speed of each pitch in each of those games. Because they are small in this sense, we will assume that they are finitely many of them. This is an idealisation, but not much of one. It is rare, if not impossible, that we are interested in more than finitely many things at once. And even if one of those things could in principle take an infinity of values (because, e.g., it is a continuous function like the speed of a pitch), the difference between a continuous function and a finite approximation to it will typically be negligible.

With that formalism on the table, I can state the four assumptions quite easily. The first is a formalisation of one of the assumptions we've already made about Max; the other three are new.

Good-Bad All worlds are either GOOD or BAD. If w is GOOD, then $V(w) = 1$, and otherwise $V(w) = 0$
Invariance $V_A(w) = V(w)$
Additivity $V(A) = \sum_w V(w)C(w|A)$
Restricted Conditionalisation $C_A(B) = C(B|A)$

I'm calling the last *Restricted* Conditionalisation because I only mean it to apply in the case where A is a factual proposition. Lewis doesn't explicitly say what restriction he is putting on conditionalisation, but he does make clear in footnote 6 of (1996) that he doesn't need or want to endorse a general version.

Note that we really need Good-Bad to make **DaB** plausible. If we didn't have Good-Bad, then we wouldn't even have V and C on the same scale. But if values and probabilities are at least on the same scale, then

the thesis passes a minimal threshold of plausibility. Nevertheless, it is false. It is false because it leads to some absurd independence results. In particular, the following proof shows that A and A° are probabilistically independent according to C.

$C(A^\circ)$	$= V(A)$	
	$= \sum_w V(w)C(w\|A)$	(Additivity)
	$= \sum_w V_A(w)C(w\|A)$	(Invariance)
	$= \sum_w V_A(w)C_A(w)$	(Restricted Conditionalisation)
	$= V_A(A)$	(Additivity), applied after updating on A
	$= C_A(A^\circ)$	(DaB), applied after updating on A
	$= C(A^\circ\|A)$	(Restricted COnditionalisation)

That is, Max treats A and A° as independent. And this is absurd. Imagine that A is the proposition that Max's hero, Jean, performs a certain action. As it stands, Max is very unsure whether this would be the right thing for Jean to do. But he is sure that Jean will do the right thing. This seems like a coherent set of attitudes for Max, at least if it is coherent to be uncertain about what is best. But we've just shown that given **DaB** and these other assumptions, it is not coherent. If Max satisfies these assumptions, then the fact that Jean does something can be no evidence that it is right, no matter how much background evidence Max has that Jean is a moral role-model. That's absurd, so one of the assumptions must go.

The argument I've given here is taken more directly from Nissan-Rozen (2015) than from Lewis, though you can sort of see it as a variant on the argument Lewis gives. Lewis argues that Independence is absurd by considering what would happen if Max updated his credences with the information that $A \vee A^\circ$. At some level, the real lesson of Lewis's arguments here, and his related arguments about conditionals (Lewis, 1976, 1986), is that correlations involving credences are really hard to preserve across updates. And since anything you could say could in principle be learned - it's at least possible to meet an Oracle and have them tell you that p for any p you can say - that puts a real constraint on these correlation theses. And that's all I've done in the previous paragraph. Max has learned, via his observation of Jean's moral probity, that if Jean makes A true, then A° is true as well. If the correlation between $C(A^\circ)$ and $V(A)$ holds universally, it must hold in this special case. But it doesn't, so there is something wrong with the correlation.

But I'm not sure that focussing on the person who learns $A \vee A^\circ$ is the best way to see the problem. Lewis points out that if Max does give

credence 1 to $A \vee A^\circ$, and A and A° are independent, then he must given credence 1 to either A or A°. And then he tries to argue this is absurd. But while there are, I think, three different ways to finish this argument, none of them seem quite compelling.

First, we could note that no matter how we update, assuming that after update Max's credences are a probability function, and $C(A \vee A^\circ) = 1$, we can prove[3] after update either $C(A) = 1$ or $C(A^\circ) = 1$. And we could appeal to the intuitive implausibility of that. The problem with this appeal to intuition is that it's so hard to imagine what it would be like to learn $A \vee A^\circ$, and nothing else, that it's hard to know what such an update should intuitively be like. Unlike the story I told of Max and Jean, it's hard to see what everyday experiences would lead to learning that. And I, at least, don't have firm intuitions about how to update if Max learns that via an Oracle.

Second, we could try to argue that this learning should go via conditionalisation. That is, after learning $A \vee A^\circ$, then Max's new credence in any proposition X should be his old credence in it conditional on $A \vee A^\circ$. And then we can prove[4] that right now Max must have an extreme credence, either 1 or 0, in either A or A°. That is, even before meeting any Oracles or the like, Max cannot both be uncertain about whether A will happen, and whether it should happen. That's really absurd. But it is easy enough for the defender of **DaB** to simply deny that updating on morally loaded propositions like $A \vee A^\circ$ should go via conditionalisation.

Third, we could try to argue by elimination that all possible formal models for update will have problems just as bad as the problems mentioned in the last paragraph. And the problem with this is simply that it's too hard to exhaust the formal models for update, even if we impose some

[3] Assume for reductio that C is a probability function, that this function makes A, A° independent, that both $C(A)$ and $C(A^\circ)$ are less than 1, and $C(A \vee A^\circ) = 1$. Let $C(A) = x$, and $C(A^\circ) = y$. By independence, $C(A \wedge A^\circ) = xy$. Since $C(A) = C(A \wedge A^\circ) + C(A \wedge \neg A^\circ)$, it follows that $C(A \wedge \neg A^\circ) = x - xy$. Since $C(\neg A^\circ) = C(A \wedge \neg A^\circ) + C(\neg A \wedge \neg A^\circ)$, it follows that $1 - y = (x - xy) + C(\neg A \wedge \neg A^\circ)$. But $C(\neg A \wedge \neg A^\circ) = 0$, since $C(A \vee A^\circ) = 1$. So $1 - y = x - xy$. Since $y < 1$, we can divide both sides by $1 - y$, getting $1 = x$, contradicting the assumption that $x < 1$.

[4] Assume for reductio that C is a probability function, and that A, A° are indepedendent both unconditionally, and conditional on $A \vee A^\circ$. The proof of the previous footnote implies that at least one of $C(A|A \vee A^\circ)$ and $C(A^\circ|A \vee A^\circ)$ is 1. The situation is completely symmetric, so without loss of generality assume it is $C(A|A \vee A^\circ) = 1$. From this it follows that $C(\neg A \wedge A^\circ|A \vee A^\circ) = 0$, so $C(\neg A \wedge A^\circ) = 0$. So $C(A^\circ) = C(A \wedge A^\circ) + C(\neg A \wedge A^\circ) = C(A \wedge A^\circ)$. But by the independence of A, A°, we also have $C(A \wedge A^\circ) = C(A)C(A^\circ)$. Putting these two together, it follows that $C(A^\circ) = C(A)C(A^\circ)$, contradicting the assumption that $C(A) < 1$.

kind of minimal viability condition on the models. This will be a bit of a running theme of later chapters of this book.

So I don't think this particular argument will work. But I don't want to make too much of the difference. After all, my argument can just be rephrased as saying that if A and A° are independent, so are $\neg A$ and A°, and that is absurd in the case where Max learns $\neg A \vee A^\circ$. So the similarities between the argument I'm offering and the one Lewis had already offered are greater than the differences.

1.2 Four Versions of Max

So it's bad to be just like Max. But what follows from that philosophically? Lewis says that the argument raises a problem for "anti-Humean" views. But this obscures more than it clarifies, since there are a lot of views that Lewis took to be anti-Humean. I find it helpful to think through four ways we might understand both the halo operator, and Max, that generate the potential argument. The first three ways take A° to mean *A is good*. The fourth will be a bit more complex.

Max-1 does not have desires at all. Or, if he has them, they play no role in his action. So here's how a typical action of Max's proceeds. He walks to the bowl to get an orange. He doesn't do this because he wants an orange, and believes they are in the bowl. Rather, he believes that there are oranges in the bowl, and believes that it would be good to have an orange, and those two beliefs completely explain his walking to the bowl. We can model Max using the familiar C, V functions from decision theory, but this model is misleading; V is really measuring a second part of Max's credal state.

Max-2 has a more familiar belief-desire psychology. He walks to the bowl because of a desire for oranges, and a belief that that's where the oranges are. But in Max-2 there is a metaphysically necessary connection between desires and the good. It just isn't possible for him to have a desire for an orange without a belief that oranges are good, or vice versa. Now it is less misleading to use a standard decision-theoretic model on Max, since he really does have beliefs (represented by C), and desires (represented by V). But these two are prevented from freely recombining; some combinations of C and V are impossible for Max.

Max-3 is like Max-2, but without the restriction on recombination. What's distinctive about Max-3 is that he is governed by a norm connecting desires with beliefs about the good. Sometimes he desires that which

he does not believe to be good, and sometimes he believes that to be good which he does not desire. But these are failings; he shouldn't do that. I could subdivide this case further, depending on how to disambiguate this 'should', but that's unnecessary for now.

Max-4 takes a bit longer to state. No longer read A° as *A is good*. Instead, read it as saying *A is X*, where X is the feature that actually makes actions good. So if all there is to morality is conformity to the categorical imperative, then A° means *A conforms to the categorical imperative*. Now Max-4 is like Max-3. Metaphysically there is nothing impossible about C and V floating free from each other. But if Max is being good, there should be a connection. In particular, Max should only want to do, i.e., value as good, actions that conform to the categorical imperative. That is, Max should value A only if he believes that A°.

Now consider four philosophers, who I'll name Athena-1, 2, 3 and 4. Each Athena thinks the corresponding version of Max is a good model for real humans. So Athena-1 thinks that humans are (at least often) like Max-1; Athena-2 thinks they are like Max-2, and so on. And then we can think that the formal argument is meant to show that each of the Athenas is mistaken.

And this certainly feels like it would be a victory for any number of anti-Humean theses. Athena-1 thinks denies the Humean dictum that reasons are slaves of the passions. Rather, she thinks that (at least when things are going well), reasons are in the driver's seat. And Athena-2 seems to deny the Humean dictum that there are no necessary connections between distinct existences. She thinks that there is a metaphysical connection between the beliefs of (at least some) ordinary humans and their desires.

In Lewis's two papers the relationship between his argument against what he calls anti-Humean moral psychology (views that deny that reason is the slave of the passions), and what he calls anti-Humean modal metaphysics (views that accept metaphysical connections between distinct existences). Thinking through the possible ways to interpret Max suggests really the argument is an argument against both of these. I'm going to set both this suggestion, and more careful Lewis exegsis aside however, and think a bit more about Athena-3 and Athena-4.[5]

Athena-3 allows that people can have desires that come apart from their beliefs about the good. She just thinks that this is a failing. Simplifying only a little, she thinks that hypocricy is a vice. She thinks it is bad to do that which is bad by your own lights. Or, at least, she holds a

[5]Note to self: If I come back to a Lewis exegesis chapter, flag it here.

slightly more complicated probabilistic version of that principle.

On the face of it, it's hard to see how that is anti-Humean. If Athena-3 thinks that the direction of explanation goes from beliefs to values, then she's getting close to denying that reason ought be the slave of the passions. But she need not endorse this order of explanation in virtue of accepting the correlation. She may think that passions, i.e., desires, should be driving the ship, and have a moral epistemology where beliefs should follow feelings of value. Or, returning to the point I flagged about the ambiguity of 'should', she may think that not satisfying this correlation is a very specific kind of vice, not one that is always most important. So she may think that all things considered, reason should be the slave of the passions, but from the perspective of not being a hypocrite, reason and passion should correlate. And as long as we think norms can exist while being over-ridden, this position looks consistent with any kind of Humean moral psychology.

Athena-4 doesn't even think anything that substantive. She just thinks that there is some feature in virtue of which things are valuable, and it is a failing to not value things one takes to have that feature. Lewis sometimes writes that the argument is related to debates about "objective ethics", but I'm not assuming Athena-4 believes in any such thing. Perhaps she does, but perhaps she thinks that things are valuable in virtue of folks around here valuing them. She's mostly just committed to the idea that it is a failing to not value the valuable. And if that makes one anti-Humean, it's not clear why we'd want to be Humeans.

To be a bit more careful, Athena-4 is making a slightly more substantive claim than this. Recall that I've set up the argument so far using **Good-Bad** as a simplifying assumption, and that eventually that will have to be dropped. It will turn out when we drop that assumption in virtue of something more plausible, Athena-4 will have to make some non-trivial assumptions about the topolgy of value. In particular, she will have to think that values have something like the structure of (some subset of) the real numbers. To put in terms from present debates, she'll have to believe in a moral theory that can be "consequentialised". That's not to say she has to be a consequentialism in any ordinary sense, just that her theory has to be translatable into consequentialist language. Ever since Moore (1903) there has been a view in philosophy that this is a trivial requirement, so Athena-4 really isn't committing to anything here. I'm more sympathetic to the argument that Campbell Brown (2011) makes that not all theories survive such translation. But this is a really small side point. It can't be any

part of Humeanism in meta-ethics or metaphysics that our moral theory falls on one or other side of this divide. If Lewis's argument shows that Athena-4 is mistaken, it proves way more than that Humeans win one or other debate.

And this is a big reason for being very suspicious of the argument so far. It simply proves too much. That doesn't tell us where or how the argument fails, but it gives us very good reason to think it must indeed fail.

Goes with four versions of Athena

1. Guise of the good
2. Necessary connection
3. Enkrasia
4. Really really broad

Note that in 1988 paper Lewis seems to appreciate broadness, but just backs off for some unclear reason.

1.3 Plan for Book

- Write at the end

Chapter 2

Moral Uncertaintism

2.1 Naming The Topic

Moral uncertaintism, as I'll use the phrase here, is the view that we should treat moral uncertainty in the same way that we treat factual uncertainty, as far as that is possible. What do I mean by "how we treat factual uncertainty"? Well, consider this little vignette, only partially fictionalised.

Imagine that I am only interested in money, specifically in getting as much money as possible. And imagine that very near my office, there is a casino. Here's something that I could do that would result in me getting a lot more money than I currently have. I get my hands on as much cash as I can, walk down to that casino, and bet it all on the roulette wheel. In particular, I bet on the number that will actually win. In a sense, this is possible. After all, for any number, I can bet a lot of cash on that number, and one of them will win, so whichever one will win, is one I can bet a lot of cash on. I keep repeating this until they start looking suspicious, at which point I collect my winnings, and leave with a very large amount of money.

In reality, I don't do that, and not just because money is not my sole aim in life. Why not? The obvious, and mostly correct, answer is that I don't know which number will win. In situations of uncertainty, like my uncertainty about which number will win, I don't do the thing that produces the best return. Rather, I do the thing that produces the best *expected* returns. At least, that's what I do when I can reasonably assign probabilities to the various possible outcomes. When not even that is possible, I have to rely on more qualitative approaches.

There are two features of this vignette that I want to draw attention

to, because they'll become important in what follows. First, while I don't know what action will maximise my money, there is one action that I know will not maximise my money. That's the action of declining to bet. And yet that's what I do. Sometimes maximising expected returns guarantees not maximising actual returns. In fact, when there are casinos in the area, that is almost always the case. Second, we can imagine a very fictionalised version of the story where the casino offered more than fair bets on roulette. Imagine they paid out $50 for every $1 bet if you guess the correct one of the 37 (or 38) numbers that could win. Then I really should go to the casino and bet heavily. And that's true even though it is very improbable that I'll win.[^To clarify this example, imagine that this 50-1 offer is a one-time deal; you can only exercise it for one bet on one spin of the wheel. And imagine that the casinos, with their Benthamite levels of surveillence, know whether you are teaming up with other people to bet on all the numbers, and won't allow those kind of shenanigans.] Just how much I should bet turns out to be a tricky question, depending on unclear issues about the shape of the function from how much money I have to how much utility I get from that money. But very plausibly I should bet a lot, even though it is more than 97% likely that I'll lose.

The moral uncertaintist thinks something similar is true for moral uncertainty. They say that in circumstances where we don't know what the morally right thing to do is, we often shouldn't do that right thing. Doing the right thing, in situations of moral uncertainty, is like betting on the roulette number that actually wins. It's nice that your strategy worked on this occasion, but it was the wrong strategy to follow. Rather what you should do is, if possible, maximise the expected moral value of your action. In cases where even that is impossible, because the relevant probabilities are not defined, you should use some other qualitative heuristic that is sensitive to your moral uncertainty.

I am following Elizabeth Harman (2015) in calling this view moral uncertaintism. I'm also going to follow Harman in suggesting that it isn't a good view. Proponents of the view, and some opponents too, have started using a different name for the view. They call it 'moral hedging'.[1] I think this is a bad name, and it's a bad name for a philosophically interesting reason. So I'm going to start this short discussion of moral uncertaintism by saying a bit about why it is a bad name, and why I'm using Harman's terminology instead.

To see why the name 'moral hedging' has seemed appealing, consider

[1]Note to self: Include some examples here.

the case of Louise. (This kind of case has been used frequently in the literature.) Louise is trying to decide whether to have meat or vegetables for dinner. She's thought a bit about the ethics of meat eating, and she's decided that it is 90% likely that meat eating is morally acceptable. But she also thinks that if meat eating is morally unacceptable, it is an abomination. So she faces the following decision problem.

	Meat-eating is acceptable (90%)	Meat-eating is unacceptable (10%)
Eat meat	1	-100
Eat vegetables	0	0

The numbers in each cell refer to the moral value of her choices. Assume she prefers meat to vegetables, and has a weak duty to promote her own well being where permissible, so if meat is acceptable, it is slightly morally preferable to eat meat to vegetables. But as noted above, if meat-eating is unacceptable, it is really morally bad - that's what the -100 value represents. Then for Louise, the expected moral value of eating meat is $0.9 \times 1 + 0.1 \times -100 = -9.1$. And the expected (and certain) moral value of eating vegetables is 0. Since $0 > -9.1$, the best option is to eat vegetables. That's what she should do.

And note that we've concluded she should do this without either saying anything substantive about the ethics of meat-eating, or without saying that she regards ethical vegetarianism is particularly likely. But what we have said is that, from Louise's perspective, meat-eating is morally risky. Eating vegetables, on the other hand, is morally safe. And the strategy of maximising expected moral value recommends this safe option. This is why, I think, the label 'moral hedging' has become popular. Louise should hedge her moral bets, play it safe, and settle for the vegetables.

But it wasn't just the theory that one should maximise expected moral value that led to this dietary recommendation. It also required some substantive assumptions about what Louise's moral views were, and what options were taken to be available to her. If we change those assumptions, we can easily make the view that we should maximise expected moral return look considerably less safe.

So consider Antoine, who in many respects is like Louise. He is also thinking about what to have for dinner. And he thinks it is 90% likely that meat-eating is morally acceptable. But conditional on meat-eating being unacceptable, he has slightly different views to Louise. She thought that if meat-eating is unacceptable, then one should settle for vegetables for din-

ner. Antoine mostly thinks that too. Conditional on meat-eating being unacceptable, he thinks it is 90% likely that one should have a vegetarian diet, and maybe be a touch sanctimonious about it around carnivores, but otherwise live life unchanged. But the other 10% of his conditional credence goes to the view that if meat-eating is wrong, then it's kind of like murder, and one has affirmative duties to protect those at risk of murder. In particular, that last 1% of Antoine's moral worldview goes to the theory that killing humans to prevent them killing animals is morally good, and declining to do so is a bad form of moral cowardice. That's especially true in cases where killing a human would lead in the long run to fewer animals being killed. And around where Antoine lives, cattle ranching is a rather unpopular line of work. Every cattle rancher you kill probably means one fewer person employed to raise cattle for beef.

So here is the decision table Antoine faces. The middle column is the possibility that ordinary ethical vegetarianism is the correct view; the last column is the view that ethical vegetarianism requires the morally righteous to butcher the butchers. And Antoine doesn't really think that's right - he only gives it 1% credence - but there really are a lot of cows you could save this way, probably about 200 for every rancher you kill. So here's how the table looks for him.

	Meat-eating is acceptable (90%)	Meat-eating is unacceptable (9%)	Meat-eating must be stopped (1%)
Eat meat	1	-100	-200
Eat vegetables	0	0	-100
Kill ranchers	-100	-100	20000

Now we just have to run the numbers. The expected moral value of eating meat is $0.9 \times 1 + 0.09 \times -100 + 0.01 \times -200 = -10.1$. The expected moral value of eating vegetables is $0.9 \times 0 + 0.09 \times 0 + 0.01 \times -100 = -1$. And the expected moral value of killing ranchers is $0.9 \times -100 + 0.09 \times -100 + 0.01 \times 20000 = 101$. By far, the best option is to turn into a rancher killing vigilante.

I don't mean this example to be a knock-down refutation of moral uncertaintism. I think moral uncertainists have at least two good responses to this example. (Though as we'll see in a bit, neither of them is cost-free.)

First, they could say that what matters for decision making under moral uncertainty is not what credences a thinker actually assigns to various moral theories, but how probable those theories really are, given that thinker's evidence. And while Antoine actually gives some credence to the view that ranchers much be killed, he shouldn't.

Second, they could say that maximising expected moral value is only one moral desideratum among many. This kind of view has some important historical precedents. Peter Abelard argued that we have moral obligations both to do what is actually right, and do do what we think is right. You can think of the obligation to maximise expected moral value as a modern, formal version of Abelard's view that we should do what we think is right. Abelard thought that people with false moral beliefs faced moral dilemmas; whatever they do will violate one of their moral obligations. And it isn't absurd to say that Antoine faces a dilemma of sorts as well.

So I don't think Antoine provides an immediately compelling argument that moral uncertaintism is false. But he does provide an immediately compelling argument that it's a very bad idea to call this view moral hedging. It is really misleading to say that the hedging, careful, play it safe view is to murder ranchers because there is a 1% chance that this is a really compelling moral obligation. Sometimes maximising expected moral value will lead one to play it safe, and sometimes it will lead to acting with reckless abandon. Of course, maximising expected financial value, which really is constitutive of a certain kind of financial prudence, has the same characteristics. So better to stick with the idea that what's definitive of the theory is that moral and factual uncertainty are treated the same way. That's an interesting enough view to discuss. And it's especially interesting for the purposes of this book because it looks like it will be committed to the kind of thesis that Lewis argued against.

2.2 Against Moral Uncertaintism

In my *Normative Externalism* (2019) I argued against moral uncertaintism at some length. I won't repeat all those arguments here, but just to have some details about the view on the table, I will run over two quick arguments. I argued that moral uncertaintism is under-specified, and under-motivated. I'll take those in order first.

To see the under-specification problem, consider Camille, who is mostly like Antoine. That is, he has the same moral credences as Antoine, and so by his lights it maximises expected moral value to kill ranchers. But Camille also has a strong desire to not kill people unless he is pretty confident that they are wrong-doers. This is over and above his desire to do the right thing. And so relative to all the things Camille values (i.e., both doing the right thing and not killing the probably innocent),

it maximises expected value to eat vegetables rather than kill ranchers. Let's assume that Camille does that, and that it would have been morally wrong to kill the ranchers.

The moral uncertaintist says Camille shouldn't have done that; he should have killed the ranchers. But what's the sense of 'should' here? It can't be that he morally should kill the ranchers. By hypothesis, morally he should not do this. It can't be that be rationally should kill the ranchers, at least if rationality is concerned with doing well by one's own lights. By hypothesis, it would be irrational by his lights to kill the ranchers. The most you can say is that if he had the same beliefs (i.e., the credal distribution from Antoine's table), but only valued doing the right thing as such, then it would produce the most (expected) value to kill the ranchers. But why should we even care about that conditional? Why think it corresponds to any kind of norm we antecedently care about? It would be actually immoral to kill the ranchers, and bad, i.e., irrational, by Camille's own lights. The moral uncertaintist is aiming for some odd hybrid norm, and it isn't one I think we should give any weight to.

To see the under-motivation problem, it helps to work through the various motivations for moral uncertaintism. I won't go through these in great detail, but I've already said enough here to see the outline of the problem.

One argument for moral uncertaintism is that it guards against an unhelpful form of moral recklessness. Just doing the right thing, whatever it is, might be excessively morally risky when one doesn't know what the right thing to do is. But if the aim is to reduce moral risk, a theory that recommends Antoine kills ranchers doesn't have a lot going for it.

Another pair of motivations are that it is unfair to ask people to do the right thing when they don't know what it is, and that it is useless advice to tell people to just do the right thing when they don't know what it is. But it's not so easy to maximise expected moral value either. There are two technical challenges here. One is figuring out how to translate different moral theories onto a common scale. This has become known as the problem of inter-theoretic value comparisons, and it seems just as hard as any other problem in first-order ethics (?). A second is figuring out the probability of different moral theories. Just how probable is it that Kant was right about the murderer at the door? If it's unfair to expect me to figure out whether Kant was right before acting, and it's bad advice to tell me in that situation to do the right thing, then it's also unfair to expect me to figure out the probability that Kant was right, and it's bad advice to

tell me to react to the actual probability that he was right.[2] When we use probabilistic reasoning in everyday life, whether it be buying insurance or not buying casino chips, we can usually offer compelling reasons for using one probability function rather than a radically different one. In the moral case, Keynesian uncertainty is the normal case, not an aberration. Hypothetical examples where we just know that this theory has probability 0.9 and that theory has probability 0.1 are even more absurd than cases where we know the moral truth. In realistic cases, *do thing right thing* is no more unfair a standard, and no worse advice, than *maximise expected moral value*.

Finally, one might try to motivate moral uncertaintism by appeal to a quite general intuition that all uncertainty should be treated alike. This is a much better argument than the previous ones, I think, and in *Normative Externalism* I spend two long chapters replying to it. Here I'll just note that if we want to treat all uncertainty alike, then we have to treat *all* uncertainty alike. We have to treat epistemic, logical and decision-theoretic uncertainty the same way we treat factual and moral uncertainty. And, I argue in that book, if we do that then we can't accept any of the existing theories of how to handle moral uncertainty, and we can't avail ourselves of either of the plausible responses to Antoine's case.

2.3 Guise of the Good

[2]Current philosophers are sometimes perplexed about why we should worry about a hypothetical case where someone turns up at the door planning to kill someone you know, and yet you nevertheless have non-trivial moral obligations towards that unwelcome guest. It helps to recall just when it was that Kant and Constant had the exchange this example originated in, and which killers were turning up at which doors in those days. For much more on this, see Rousselière (2018).

Bibliography

Arló Costa, Horatio, Collins, Jessica, and Levi, Isaac. 1995. "Desire-as-belief implies opinionation or indifference." *Analysis* 55: 2-5, doi:10.1093/analys/55.1.2.

Brown, Campbell. 2011. "Consequentialize This." *Ethics* 121: 749-771, doi:10.1086/660696.

Collins, Jessica. 2015. "Decision Theory After Lewis." In Barry Loewer and Jonathan Schaffer (eds.), *A Companion to David Lewis*, 446–458. John Wiley and Sons.

Harman, Elizabeth. 2015. "The Irrelevance of Moral Uncertainty." *Oxford Studies in Metaethics* 10: 53-79, doi:10.1093/acprof:oso/9780198738695.003.0003.

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

Lewis, David. 1976. "Probabilities of Conditionals and Conditional Probabilities." *Philosophical Review* 85: 297-315, doi:10.2307/2184045. Reprinted in *Philosophical Papers*, Volume II, pp. 133-152.

—. 1986. "Probabilities of Conditionals and Conditional Probabilities II." *Philosophical Review* 95: 581-589, doi:10.2307/2185051. Reprinted in *Papers in Philosophical Logic*, pp. 57-65.

—. 1988. "Desire as Belief." *Mind* 97: 323-332, doi:10.1093/mind/xcvii.387.323.

—. 1996. "Desire as Belief II." *Mind* 105: 303-313, doi:10.1093/mind/105.418.303.

Moore, G. E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

Nissan-Rozen, Ittay. 2015. "A Triviality Result for the "Desire by Necessity"Thesis." *Synthese* 192: 2535–2556.

Rousselière, Geneviève. 2018. "On political responsibility in post-revolutionary times: Kant and Constant's debate on lying." *European Journal of Political Theory* 17: 214-232, doi:10.1177/1474885115588100.

Russell, Jeffrey Sanford and Hawthorne, John. 2016. "General Dynamic Triviality Theorems." *Philosophical Review* 125: 307-339, doi:10.1215/00318108-3516936.

Savage, Leonard. 1954. *The Foundations of Statistics*. New York: John Wiley.

Weatherson, Brian. 2019. *Normative Externalism*. Oxford: Oxford University Press.

Weirich, Paul. 2012. "Causal Decision Theory." In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2012 edition.