Moral Uncertainty and the Desire as Belief Thesis

Brian Weatherson

In recent years, many philosophers have defended what Elizabeth ? calls 'moral uncertaintism'. This view says that in cases of moral ignorance, one should be guided by moral probabilities, in much the same way that one should be guided by factual probabilities in cases of factual ignorance. There have been many discussions of the details of moral uncertaintism in the recent literature, and here I want to focus on a big-picture structural feature of the view.

Moral uncertaintism is a version of an anti-Humean theory of moral motivation. It says that, at least when things go well, one's actions are responsive to a certain class of one's beliefs. It holds that moral attitudes are in a certain sense 'besires'; they behave like both beliefs and desires. They have the formal structure of beliefs - they satisfy the probability calculus and (presumably) are responsive to evidence in the ways characteristic of beliefs. But they have, at least when things go well, the motivational force of desires. So it is interesting to look back to the active debate in the 1980s and 1990s about the viability of besires as a way to evaluate the viability of moral uncertaintism.

I'm going to focus on one particular prominent objection to anti-Humean theories of motivation: David Lewis's argument in "Desire as Belief" and "Desire as Belief II" that anti-Humean theories are incompatible with some platitudes about how beliefs are updated. If correct, Lewis's arguments would be fatal for moral uncertaintism. I will argue, however, that Lewis's arguments are not correct. The various parts of my response to Lewis are going to be fairly familiar, but I will be putting those parts together in a novel way. I'm going to argue that Lewis's argument starts with an assumption that evidential decision theorists should reject, then makes a move in the middle that causal decision theorists should reject, and between these two moves there are few (if any) theorists who should feel worried about the argument.

? have argued that Lewis's argument fails at a different spot, in the so-called Invariance assumption. I think that's wrong. Moral uncertainists, and anti-Humeans more generally, should accept Invariance. But their arguments against it bring out two really interesting points. One is that the common equation of value in evidential decision theory with 'news value' is at best really misleading. We should rather think of it something like contentedness with the news. The other is that there are some hard puzzles in

how to model the state space that moral probabilities are defined over. They can't be possible worlds, because the supervenience of the normative on the descriptive means there is no space there for distinctively moral uncertainty. A natural move is to think of them as pairs of something that determines the physical facts (perhaps a Lewisian concreta) and something that determines the moral facts. This idea of treating the state space as physical-moral pairs is not at all new; Bradley (2009) uses this to offer another kind of response to Lewis's argument. But it turns out there is a puzzle here. Should the second item in the pair determine the value of just the physical world in that very pair, or of all worlds? Both options have their downsides, and neither is obviously compatible with moral uncertaintism. I'll argue that the second option is less bad, but the issues here are tricky.

Finally, I'll discuss a recent strengthening of Lewis's argument due to ?. They argue that Lewis's argument goes through with just some very weak structural constraints on principles of updating beliefs. In particular, they argue that the argument can be made to work with just the assumption that the updating rule satisfies Monotonicity. That's the constraint that if one updates on A, and then performs other updates, the resulting state is still one where A is taken to be settled as true, as long as the other updates were consistent with A. Using just this assumption, plus Invariance, we can generate a problem for anti-Humean theories, including moral uncertaintism. They further argue that the only way to save these theories will be to ditch Monotonicity, and I am going to set out and endorse this argument. Finally, they suggest that the only way to ditch Monotonicity will be to have a non-truth-conditional theory of moral claims. That would be a problem for moral uncertaintism, at least on the assumption that the contents of beliefs are truth-conditional. I'll close with some tentative suggestions for how to avoid this outcome, while noting that none of the suggestions are cost-free for moral uncertaintists.

1 Moral Uncertaintism

Compare the following two little vignettes.

> Tweedledee is driving down a country road. By the side of the road is a young child distractedly chasing a butterfly. Tweedledee knows that if the butterfly swerves into the road, the child will follow it. And he knows butterflies do occasionally swerve. Nevertheless, he takes no action to slow down or move away from the child, and drives quickly down the road.
>
> Tweedledum has already driven down that road, and is settling

> into a country pub for lunch. The vegetarian options are a bit dire, as they tend to be in pubs in that part of the world. But Tweedledum knows that there are good arguments that meat-eating is morally rephrensible, and if those arguments are close to correct, the wrong of meat-eating is orders of magnitude more serious than the downside of having a sub-par lunch. Nevertheless, he orders the steak pie, and enjoys it.

Traditionally, most moral theorists treated these cases quite differently. What Tweedledee does is seriously wrong, even if the child does not run into the road. By not slowing down, Tweedledee is doing something that has a substantial chance of resulting in a terrible, tragic outcome, and doing so for a trifling gain. That's a really bad case of reckless action, even if no one actually gets hurt. On the other hand, we need to actually conclude our moral debates about meat-eating in order to judge Tweedledum's actions. If meat-eating turns out to be morally acceptable, then Tweedledum ran a moral risk, but got lucky, and is not subject to any criticism.

Moral uncertainists[1] think the cases are much more analogous than this analysis suggests. In each case, we can judge the people from the inside, as they are now, before the facts are settled. Before we know whether the child will run into the road, it is seriously wrong to not take care to avoid them. And before we know whether meat-eating is immoral killing, it is seriously wrong in more or less the same way to keep eating meat. It might turn out that meat-eating is morally acceptable. But not hedging against that moral risk is wrong for Tweedledum, just like not hedging against the vehicular risk is wrong for Tweedledee.

There is a related argument for moral uncertaintism that will become important below. Good people, say the uncertaintists, care about doing the right thing. But the advice *Do the right thing* is not very helpful to people like Tweedledum. They don't know what is right. While Tweedledum can't be guided by the moral facts, perhaps he can be guided by the moral probabilities.[2]

Harman doesn't make just this distinction, but it will be helpful to our discussion to distinguish a special class of theories that I'll call *strong moral uncertaintism*. These theories say that Tweedledee's and Tweedledum's cases aren't just similar, they are similar in all philosophically significant respects. We should treat moral uncertainty, like what confronts Tweedledum, exactly

[1] Include citations

[2] More citations here

the same way that we treat factual uncertainty, like what confronts Tweedledee. Most moral uncertaintists qualify, at least to some extent, the analogy here, though it is natural to read Jacob Ross (2006) and Michael Smith (2009) as defending strong moral uncertaintism.

Alongside the rise of interest in moral uncertaintism, there have been a number of criticisms of the view. There are difficult technical issues in getting the details of uncertaintism right, and Brian ? has argued that these pose an insuperable difficulty to the view. Brian ? has argued that moral uncertaintism is committed to being on the wrong side of debates about moral motivation. And Elizabeth ? has argued that moral uncertaintism is committed to being on the wrong side of debates about blame and moral ignorance. All of these are controversial; it isn't obvious either that moral uncertaintism has these commitments, not that the side in question is the wrong side. Rather than dig into those debates, I want to take a step back, and look at how uncertaintism connects to some broader debates.

One issue for moral uncertaintism concerns the kind of criticisability it posits. What kind of norm does Tweedledum violate when he orders the steak pie. On the one hand, it can't really be moral criticism, on pain of contradiction. After all, it is a wide open question whether morality forbids meat-eaiting. The most popular move at this point is to say that moral uncertaintism concerns rational norms; it would be irrational to do the thing that may be seriously morally wrong. I'll mostly work with this assumption, but note that it is already a slight qualification to strong moral uncertaintism. After all, the norm Tweedledee violates is a straightforwardly moral norm.

Another issue concerns what to say about people who are not just morally uncertain, but straightforwardly morally mistaken. Moral uncertaintists have a hard time accommodating the points Nomy ? makes about cases of 'inadvertent virtue', i.e., right-doing by people who think they are acting wrongly, and 'misguided conscience', i.e., wrong-doing by people who think they are acting rightly. This issue will mostly stay under the surface here, but is important to keep in mind.

2 The Lewisian Objection

In order to avoid technical complications about the formulation of moral uncertaintism, I'm going to focus on an exceedingly simple case. The case has the downside of being completely unrealistic, and the upside of sidestepping those technical challenges that ? argues are fatal for moral uncertaintism. So we'll assume that X is facing a choice between some options, and X doesn't know a bunch of moral facts. But X does know that their choice, whatever

it is, will be either Good or Bad. And X also knows that all Good choices are equally right and equally good, and all Bad choices are equally wrong and equally bad. In other words, we can usefully model the moral value of X's choice as being either 1 or 0, depending on what X chooses, and on what the moral facts are. Given this extremely idealised set of assumptions, every extant approach to moral uncertaintism says that X should choose the option that has the highest probability of being Good. That will option will maximise expected moral value, and will be the one that is most probably morally acceptable, so lots of moral uncertainists should like it.

In fact, it is plausible that moral uncertaintists should like the following equation, where A is an arbitrary factual proposition, the 'halo' maps a factual proposition A onto the proposition that A is good, V is the value (or desirability) of a proposition being true, and Pr is the relevant probability function.[3]

$$V(\text{A}) = \text{Pr}(\mathring{\text{A}})$$

This is what Lewis calls the desire as belief thesis. And, he argues, it must be false. To see why, we just need one bit of theory.

Say that a *world* is a specification of all the things that matter to fixing a value for an outcome. This is more or less what Savage calls a 'small world'. A world in this sense is more detailed in some ways that a Lewisian concreta, and much less detailed in others. Imagine that we are tasked with deciding what will happen to the captured enemies of the revolution. A small world will fix what we do, and will fix the moral value of that, but will not fix much more. It won't determine, for instance, who won last night's baseball game. But it might fix things in a way that is both metaphysically impossible, and a priori incoherent. Assume (plausibly enough) that it is wrong to execute these enemies. Then there may be no possible world where we execute them, and this is Good. And, if some form of Kantianism is correct, it may even be a priori knowable that there is no such world. But we'll include in our model a world where the executions are carried out, and this is Good. And this is because even if one could in principle know that no such world obtains, a particular actor at a particular time in history may not know this, and may be unsure what morality demands of them in the middle of a revolution. And we want to model how things go for such an ignorant agent.

[3]Note that I'm not taking a stance here on whether Pr measure something epistemic, like an evidential probability function, or something doxastic, like the agent's credence. The arguments to follow go through either way.

In many cases, especially when we know the result will be Good or Bad and not anything in between, there will be finitely many worlds to consider. Each of these worlds could be filled out in more detail, but we're assuming the filling out won't matter, at least to the relative value of the worlds being considerd. Given this, we'll assume the following equation holds for values.

$$V(\mathrm{A}) = \sum_{w \in \mathrm{A}} V(w) \Pr(w|\mathrm{A})$$

That is, the value of a proposition is the weighted average of the value of the worlds where the proposition is true, where the weighting is given by the probability of each world being the one that makes the proposition true.

Lewis attributes this view about value to Richard ?, who in turn credits XXX with describing it as the 'news-value' of a proposition. This last description, while widely repeated, turns out to be misleading in an important way; it's really more like one's contentedness with the news that A. This will matter a bit in what follows, as will the fact that this really looks like the kind of value an evidential, as opposed to a causal, decision theorist will care about.

Now here is the big theoretical claim that we need. We will use subscripts to denote the value of a function after updating on a piece of evidence. So $V_{\mathrm{B}}(\mathrm{A})$ is the value of A after updating on B, and $\Pr_{\mathrm{B}}(\mathrm{A})$ is the probability of A after updating on B.

$$V_{\mathrm{B}}(w) = V(w)$$

That is, the value of a world does not change when one learns something new. And that's for the simple reason that the worlds, as defined, fix all the information relevant to determining value. If that's so, then learning something new can't change the value of the world. (We'll come back to this argument in section 4.) Given that principle, we can infer the following equation:

$$V(\mathrm{AB}) = V_{\mathrm{B}}(\mathrm{A})$$

Proof:

$$\begin{aligned} V(AB) &= \sum_{w \in A} V(w) \Pr(w|AB) \\ &= \sum_{w \in A} V_B(w) Pr(w|AB) \\ &= \sum_{w \in A} V_B(w) Pr_B(w|A) \\ &= V_B(A) \end{aligned}$$

Substituting A for B, and using the fact that AA = A, we get

$$V(\mathrm{A}) = V_{\mathrm{A}}(\mathrm{A})$$

Now the desire as belief equation $V(\mathrm{A}) = \Pr(\mathring{\mathrm{A}})$ is meant to hold for all (rational) subjects at all times. So it should hold before and after conditionalising on A. So we have $\Pr(\mathring{\mathrm{A}}) = V(\mathrm{A}) = V_{\mathrm{A}}(\mathrm{A}) = \Pr_{\mathrm{A}}(\mathring{\mathrm{A}})$. That is, $\Pr(\mathring{\mathrm{A}}) = \Pr_{\mathrm{A}}(\mathring{\mathrm{A}})$. That is, A and Å are probabilistically independent. And we have proven this on the basis of just considerations of rationality, so it must hold for all agents at all times. But this is absurd. Let A be the proposition that a person you have a lot of moral respect for took a particular decision in a tough moral situation. Then A should be evidence for Å, but the independence thesis says it is not.

We can turn the independence claim into even more absurd results if we assume it holds after various updates. Imagine, says Lewis, that an agent has just learned $\mathrm{A} \vee \mathring{\mathrm{A}}$. (It seems like anyone could learn this via testimony, at least given the permissive view we are taking towards moral epistemology that allows that rational people can have moral beliefs that are a priori false.) It is incoherent to be sure that $\mathrm{A} \vee \mathring{\mathrm{A}}$, to have a probability for each of A and for Å that is strictly between 0 and 1, and have A and Å be probabilistically independent. Since the agent is sure of $\mathrm{A} \vee \mathring{\mathrm{A}}$, they just learned it, and the independence claim has been proven to be a constraint on all agents. So this implies that on learning $\mathrm{A} \vee \mathring{\mathrm{A}}$, the agent must be sure of one of A, $\neg$A, Å or $\neg$Å. That's already absurd (assuming moral uncertainty is rational), but we can arguably make it worse. Assuming that learning goes by conditionalisation, the only way to guarantee that one of these four will be certain after conditionalising on $\mathrm{A} \vee \mathring{\mathrm{A}}$ is that prior to conditionalisation, the agent must have been certain of one of the four of them. (At least, assuming A and Å were independent prior to learning, which we've also proven is a constraint.)

This is all absurd, and Lewis concludes that the villain of the piece is the initial equation, $V(\text{A}) = \Pr(\text{Å})$. And if that equation falls, then it seems moral uncertaintism must fall with it, since the moral uncertaintist endorsed the idea that the value of an action is tied to the probability that it is Good. So we have a new, and seemingly systematic, objection to uncertaintism. The problem is, it doesn't work. The next section will outline why.

3 Replying to Lewis

In this section I'll look at a number of ways of interpreting the formalism in Lewis's argument. First, I'll consider the interpretation that Lewis himself put forward, that V measures something like desirability. Second, I'll consider what happens if V measures not desirability, but choice-worthiness, and our theory of choice is evidential decision theory. Third, I'll consider what happens if V measures not desirability, but choice-worthiness, and our theory of choice is causal decision theory. On each of the first two interpretations, I'll argue that the uncertaintist is not committed to the desire as belief equation. Rather, they are committed to the slightly more complicated equation $V(\text{A}) = \Pr(\text{Å} \mid \text{A})$. This is what Lewis calls Desire as Conditional Belief, and he notes (correctly) that he has no distinctive argument against it. On the third interpretation, I'll note the now familiar arguments that $V(\text{A}) = \sum_{w \in \text{A}} V(w) \Pr(w|\text{A})$ is false, so the argument doesn't get off the ground.

So let's start with the first interpretation, which is the one that Lewis focussed on. This says V measures the desirability of something being true. To make sure that we're really focussed on desirability, and not choice, we'll consider a case where a third party, call them X, is making a choice. And X has to choose between three options, A, B and C. We don't know which of these is morally best, because each of them has a flaw. A is arguably wrong because insufficiently benevolent, B is arguably wrong because it would be an unjustified promise-breaking, and C is arguably wrong because it would violate a duty of friendship. What should we hope X will do, if we are moral uncertaintists? (Remember we are still assuming we know that whatever X does, it will be Good or Bad, and not anything better, worse, or in between.)

The simple answer is that we should hope X does the thing that is most probably Good. But the simple answer is too simple; it gets things wrong in some cases. Imagine we know a little about X. We know that X is not a terrible person, and the fact that they choose something is some evidence that it is Good and not Bad. But we know a bit more than that. We know that X sometimes messes up when it comes to benevolence and friendship, but X

does not mess up when it comes to promises. We are sure that X would never wrongly break a promise. They might break a promise when this is justified, e.g., when the only way to save a drowning child involves breaking a promise to meet a friend for coffee. But they never wrongly break a promise.

Given that, we should hope that X takes option B. If X does take option B, we will be sure they have done something Good. And that's the best case scenario. If they do A or C, we will still think think there is some chance they are doing what is Bad. So if we care about getting to the Good, we should be hoping that its option B that gets chosen. That's not to say that option B is best to choose - it may be terrible and not be chosen because it is terrible. But it is the one we should hope is chosen.

The general lesson of this case is that we should hope that X chooses something that, given it is chosen, is probably good. In symbols, we have $V(\text{A}) = \Pr(\text{Å} \mid \text{A})$. And that's the thesis that Lewis calls Desire as Conditional Belief, and which he doesn't have a distinctive argument against. That's what the moral uncertaintist should adopt if they want to identify V with something like desirability or newsworthiness.

What if we identify V with something more like choice-worthiness? This is, after all, a more common focus for moral uncertaintists; they want to develop principles for morally uncertain agents to make choices. Here we need to make a choice between something like evidential decision theory, and something like causal decision theory. To see the difference, consider this rather improbable situation. Our hero has to choose between two options, A and U. And she knows the following things:

- Å is 1% more probable than Ů.
- If she chooses A, then $\Pr(\text{Å}) = 0.1$, and $\Pr(\text{Ů}) = 0.09$.
- If she chooses U, then $\Pr(\text{Å}) = 0.91$, and $\Pr(\text{Ů}) = 0.9$.

As far as I can tell, extant versions of moral uncertaintism do not take a stand on what to do in such a situation. And, as far as I can tell, this is a perfectly reasonable state of affairs, because Newcomb-like problems for moral uncertainty are not exactly thick on the ground. But still, for completeness, we might wonder how those theories would extend to such a case. And it isn't, I think, entirely obvious what to do. Let's say that evidential moral uncertaintism says that in this case, the hero should do U, because doing U makes it 90% likely that she'll do the right thing, while doing A makes this only 10% likely. And let's say that causal moral uncertaintism says that in this case, the hero should do A, because no matter what, Å is 1% more probable than Ů, and she wants to do what is probably right. I really don't have views

on which of these is correct. What I do want to say is that neither view is vulnerable to Lewis's argument, but the views differ on just where they think Lewis's argument goes wrong.

The evidential theory says that Desire as Belief does not represent their view; rather, their view is more like Desire as Conditional Belief. So they are perfectly content to see Lewis refute V(A) = Pr(Å), since they think value goes with probability of goodness conditional on the act being performed. After all, V(A) = Pr(Å) would imply that in this case it is better to do A than U, and the evidential moral uncertaintist denies this.

The causal theory says that Desire is Belief is right. But they deny what Lewis calls the addition rule for value, namely $V(\mathrm{A}) = \sum_{w \in \mathrm{A}} V(w) \Pr(w|\mathrm{A})$. Rather, they endorse something more like the theory of value put forward by that great causal decision theorist, David (?). Here's how they might do it, where the variable H ranges over hypotheses about the way things might be morally.

$$V(\mathrm{A}) = \sum_{H} V(\mathrm{A}H) Pr(H)$$

And given that account of value, we can't infer $V_{\mathrm{A}}(\mathrm{A}) = V(\mathrm{A})$, since we don't have $\Pr \mathrm{A}(H) = \Pr(H)$. And so a crucial step in the derivation doesn't go through.

Now it is possible that there are some antecedently plausible ways of developing moral uncertaintism so that Lewis's argument raises a problem. But I think most forms of moral uncertaintism will be untouched by it. Forms that look more or less like evidential decision theory will be, for independent reasons, committed to Desire as Conditional Belief, not Desire as Belief. And forms that look more or less like causal decision theory will reject the picture of value that lets us derive $V_{\mathrm{A}}(\mathrm{A}) = \mathrm{V}(\mathrm{A})$, which is essential to Lewis's argument. Causal decision theory is not the negation of evidential decision theory, so it is possible there are other options which are threatened. But the most obvious ways to develop moral uncertaintism are not.

References

Bradley, Ben. 2009. *Well-Being and Death*. Oxford: Oxford University Press. (2)

Ross, Jacob. 2006. "Rejecting Ethical Deflationism." *Ethics* 116:742–768, doi:10.1086/505234. (4)

Smith, Michael. 2009. "Consequentialism and the Nearest and Dearest Objection." In Ian Ravenscroft (ed.), *Minds, Ethics, and Conditionals: Themes From the Philosophy of Frank Jackson*, 237–266. Oxford: Oxford. (4)