Moral Uncertainty and Desire as Belief

Brian Weatherson

University of Michigan, Ann Arbor

June, 2018

Background

- Good people act in ways that is sensitive to uncertainty about the physical facts.
- Do they also act in ways that are sensitive to uncertainty about the moral facts?

Background

- Good people act in ways that is sensitive to uncertainty about the physical facts.
- Do they also act in ways that are sensitive to uncertainty about the moral facts?
- I say **no**, following inter alia Ittay and Liz. Good people are sensitive to true morality, not to their (reasonable) guesses about what morality might be.

Background

- Good people act in ways that is sensitive to uncertainty about the physical facts.
- Do they also act in ways that are sensitive to uncertainty about the moral facts?
- I say **no**, following inter alia Ittay and Liz. Good people are sensitive to true morality, not to their (reasonable) guesses about what morality might be. For many more details see my *Normative Externalism*, (OUP, sometime).
- There are a number of ways to answer **yes**.
- I'll call **moral uncertaintism** a strong form of the yes answer that says an overriding duty is to maximize what one (reasonably) believes is good.
- I thought Lewis's desire as belief arguments showed that moral uncertaintism was incoherent.

Background

- Good people act in ways that is sensitive to uncertainty about the physical facts.
- Do they also act in ways that are sensitive to uncertainty about the moral facts?
- I say **no**, following inter alia Ittay and Liz. Good people are sensitive to true morality, not to their (reasonable) guesses about what morality might be. For many more details see my *Normative Externalism*, (OUP, sometime).
- There are a number of ways to answer **yes**.
- I'll call **moral uncertaintism** a strong form of the yes answer that says an overriding duty is to maximize what one (reasonably) believes is good.
- I thought Lewis's desire as belief arguments showed that moral uncertaintism was incoherent. I was wrong.

Overview

1. Discuss Lewis's argument that desires and beliefs must be distinct, suggest it is a prima facie problem for moral uncertaintism.
2. Set out the difference between 'evidentialist' and 'causal' versions of moral uncertaintism, and note a plausible case where they come apart.
3. Show that Lewis's argument relies on being a causalist at one point, and an evidentialist at another point, and so isn't persuasive.
4. Describe two models for 'worlds' in the moral uncertaintist framework, and discuss the strengths and weaknesses of each.

Plan

Lewis's Argument

Two Kinds of Moral Uncertaintism

Responding To Lewis

What are Worlds

Lewis's Target

Lewis really had two targets that he didn't distinguish very carefully.

- There is a single state, e.g., a judgment that X is good, that is both a belief and a desire. This violates the Humean principle: *No necessary connection between distinct existences*.

Lewis's Target

Lewis really had two targets that he didn't distinguish very carefully.

- There is a single state, e.g., a judgment that X is good, that is both a belief and a desire. This violates the Humean principle: *No necessary connection between distinct existences*.
- Having some belief, e.g., a belief that X is good, makes it rationally mandatory to have some desire, e.g., a desire to do X. This violates the Humean principle: *Reason is the slave of the passions*.

Lewis's Target

Lewis really had two targets that he didn't distinguish very carefully.

- There is a single state, e.g., a judgment that X is good, that is both a belief and a desire. This violates the Humean principle: *No necessary connection between distinct existences*.
- Having some belief, e.g., a belief that X is good, makes it rationally mandatory to have some desire, e.g., a desire to do X. This violates the Humean principle: *Reason is the slave of the passions*.

I'm primarily interested in the second.

The Equation

- Assume we have a class of factual descriptive propositions.
- For any factual proposition A, let A° be the proposition that A is good.

The Equation

- Assume we have a class of factual descriptive propositions.
- For any factual proposition A, let A° be the proposition that A is good.
- Assume for now that we know everything is either Good or Bad, and all Good things are equally good, and all Bad things are equally bad. (Obviously a simplifying assumption.)
- So we can set the value of Good things to 1, and the value of Bad things to 0.

The Equation

- Assume we have a class of factual descriptive propositions.
- For any factual proposition A, let A° be the proposition that A is good.
- Assume for now that we know everything is either Good or Bad, and all Good things are equally good, and all Bad things are equally bad. (Obviously a simplifying assumption.)
- So we can set the value of Good things to 1, and the value of Bad things to 0.

This makes the equation plausible.

$$V(A) = \Pr(A^\circ)$$

Worlds

- A world *w* specifies the truth value of any truth-apt claim that is relevant to a current decision.
- Assume in a given decision there are finitely many of these. This is a bit idealising, but actually plausible.
- And assume that claims about goodness are truth-apt, as the moral uncertaintist sort of needs.
- So worlds will contain specification of whether things are Good or Bad.
- So half of the worlds will be metaphysically impossible, but that's ok.

Assumptions

Restricted Invariance $V_A(w) = V(w)$

Additivity $V(A) = \sum_w V(w)\,\mathrm{Pr}(w|A)$

Restricted Conditionalisation $\mathrm{Pr}_A(B) = \mathrm{Pr}(B|A)$

Independence Proof

$$
\begin{aligned}
\Pr(A^\circ) &= V(A) && \\
&= \sum_w V(w)\Pr(w|A) && \text{(Additivity)} \\
&= \sum_w V_A(w)\Pr(w|A) && \text{(Restricted Invariance)} \\
&= \sum_w V_A(w)\Pr_A(w|A) && \text{(Restricted Conditionalisation)} \\
&= V_A(A) && \text{(Additivity), applied to updated values} \\
&= \Pr_A(A^\circ) && \\
&= \Pr(A^\circ|A) && \text{(Restricted Conditionalisation)}
\end{aligned}
$$

Absurdity

- Lewis makes a further assumption to show that this ‘trivialises’ the view, a less restricted version of Conditionalisation.
- I think that further assumptions is implausible.
- But the independence result is already absurd.
- If A is that a person we have a high moral opinion of takes a particular decision, then A and A° are evidence for each other.

Plan

Lewis's Argument

Two Kinds of Moral Uncertaintism

Responding To Lewis

What are Worlds

A Puzzle Case

- Hero faces a choice between A, B and some less attractive options.
- Right now, we think A° has probability 0.5, and B° has probability 0.9.
- But we know Hero is very good at making A-type actions. If she does A, we will be certain it is Good. That is $\Pr(A^\circ|A) = 1$. But we don't think she's any kind of expert about B-type actions. So $\Pr(B^\circ|B) = \Pr(B^\circ) = 0.9$.
- What should we hope Hero does?
- Separately, if Hero knows all this, what would we advise her to do, and what, from an uncertaintist perspective, should she do?

Option A and Hope

- If Hero does A, then we'll be sure that she does something Good.
- That's a nice feature of her action to have.
- Indeed, it's the best case scenario.
- So I think it's what we should hope happens.

Option A and Hope

- If Hero does A, then we'll be sure that she does something Good.
- That's a nice feature of her action to have.
- Indeed, it's the best case scenario.
- So I think it's what we should hope happens.
- Of course, I'm speaking for the uncertaintist here; I think what we should hope depends on what's really Good.

Option B and Deliberation

- Hero starts out thinking B is more likely Good.
- It would be very weird to choose A on the grounds that her choosing it would be evidence that it is Good.
- After all, if she chooses it on those grounds, then it is hard to see how she is any kind of expert.
- And if she's not an expert, she shouldn't change her credence in A°.
- So maybe there is a case here for option B.

Newcomb's Problem

- This feels to me like a Newcomb's problem.
- We should hope Hero does A - like we should hope our friend takes one box.

Newcomb's Problem

- This feels to me like a Newcomb's problem.
- We should hope Hero does A - like we should hope our friend takes one box.
- But the reasons we should hope this are not necessarily reasons that can be used in deliberation.
- Arguably from the deliberative perspective, our friend should take both boxes.

Newcomb's Problem

- This feels to me like a Newcomb's problem.
- We should hope Hero does A - like we should hope our friend takes one box.
- But the reasons we should hope this are not necessarily reasons that can be used in deliberation.
- Arguably from the deliberative perspective, our friend should take both boxes.
- You can say all that and still think it is hard philosophical question about what should be done. The evaluative perspective is distinct from the perspective of hope, and the deliberative perspective.

Two Options

Evidential Moral Uncertaintism Hero should choose option A. In general, people should maximise $\Pr(A^\circ|A)$.

Causal Moral Uncertaintism Hero should choose option B. In general, people should maximise $\Pr(A^\circ)$.

Two Options

Evidential Moral Uncertaintism Hero should choose option A. In general, people should maximise $\Pr(A^\circ|A)$.

Causal Moral Uncertaintism Hero should choose option B. In general, people should maximise $\Pr(A^\circ)$.

I think the evidential version is better, but I'm not an uncertaintist, so I doubt my intuitions count for much here.

Two Options

Evidential Moral Uncertaintism Hero should choose option A. In general, people should maximise $\Pr(A^\circ|A)$.

Causal Moral Uncertaintism Hero should choose option B. In general, people should maximise $\Pr(A^\circ)$.

I think the evidential version is better, but I'm not an uncertaintist, so I doubt my intuitions count for much here. Also, whether this is exactly the right way to formulate the causal view turns on some tricky questions about the way to think about utilitarianism under moral uncertainties. Maybe we can talk about this in questions.

An Argument I Reject

- You could try to argue this way.
- Both forms of uncertaintism are implausible for one reason or another.
- So uncertaintism fails.
- That is really not my aim here.
- I think it's just kind of interesting to see a new choice point in developing a (false) theory.

Plan

Lewis's Argument

Two Kinds of Moral Uncertaintism

Responding To Lewis

What are Worlds

Quick Version

- Evidential versions of moral uncertaintism reject $V(A) = \Pr(A^\circ)$. Instead they accept $V(A) = \Pr(A^\circ|A)$. So the argument is a reductio of a position they do not hold.

Quick Version

- Evidential versions of moral uncertaintism reject $V(A) = \Pr(A^\circ)$. Instead they accept $V(A) = \Pr(A^\circ | A)$. So the argument is a reductio of a position they do not hold.
- Causal versions of moral uncertaintism reject the addition postulate. It's the rule, as Lewis himself says, for evidential decision theory.

Quick Version

- Evidential versions of moral uncertaintism reject $V(A) = \Pr(A^\circ)$. Instead they accept $V(A) = \Pr(A^\circ|A)$. So the argument is a reductio of a position they do not hold.
- Causal versions of moral uncertaintism reject the addition postulate. It's the rule, as Lewis himself says, for evidential decision theory.
- So really no one accepts the argument.

Evidential Version

- This is actually really easy to see.
- $V(A) = \Pr(A^\circ)$ implies that option B is better than option A in the worked example.
- But the evidential theorist doesn't want B over A.
- So Lewis's argument is a reductio of an equation they have independent reason to reject.

Causal Version

- This is a little trickier to see, because it depends on precisely how we understand A° in a causal model.
- And to be honest, I haven't worked out a good way to do that yet.
- But however you do it, if you multiply world values by the probability of that world conditional on an act, you'll get that conditional probabilities of goodness, not unconditional probabilities of goodness, matter.
- And that's not what the causalist wants.

Plan

Lewis's Argument

Two Kinds of Moral Uncertaintism

Responding To Lewis

What are Worlds

Overview

- I'm going to work through a puzzle for the evidentialist version of moral uncertaintism.
- It's a puzzle - that's not a coy way of saying it's an objection or a refutation.
- If we have time, I'll come back at the end to why the causalist faces a different kind of puzzle.

Worlds

- Worlds in this context are nothing like Lewisian concreta.
- They are what determine the truth value of relevant truth-apt claims, and they are what rational credences are defined over.
- They are more coarse-grained than Lewisian concreta in that they don't determine the truth-value of irrelevant claims.
- And they are more fine-grained than Lewisian concreta in that some of them, the ones involving false moral theories, are metaphysically impossible.

A Minimal Requirement

At the very least, worlds should do two things:

1. Set the truth value of those descriptive propositions that are relevant.
2. Set the moral value of that set of descriptive truths.

Minimal Worlds

So first hypothesis.

- Worlds are ordered pairs.
- The first member of the pair is a set d of descriptive facts.
- The second member is a number (either 0 or 1 in the simple context we're discussing) that sets the value of d.
- So a world is $\langle d, m \rangle$, and $V(\langle d, m \rangle) = m$.

A Nice Feature

If classical evidential decision theory with all values bounded is consistent, so is this model. We can prove this by turning a 'minimal worlds' model into a classical model.

- Let $G = \{\langle d, m \rangle : m = 1\}$, and $A^\circ = A \supset G$.
- So in the new model we have $V(A) = \Pr(A^\circ | A)$. (I won't prove this.)
- Set $\Pr_C(w) = \Pr(\langle w, 1 \rangle) + \Pr(\langle d, 0 \rangle)$.
- And set $V_C(w) = \frac{\Pr(\langle w,1 \rangle)}{\Pr(\langle w,1 \rangle) + \Pr(\langle w,0 \rangle)}$.

The classical model will agree with the minimal worlds model on anything they both take a view on, and this will be preserved under conditonalisation on factual propositions. And we can more or less do the reverse trick, turning any classical model with bounded utilities into a minimal worlds model.

A Simple Example

- So we can be sure the minimal worlds model is consistent and not trivial. But it has a weird feature. I will show this wil a simple example.

A Simple Example

- So we can be sure the minimal worlds model is consistent and not trivial. But it has a weird feature. I will show this wil a simple example.
- There is just one descriptive proposition that we care about.
- So the description will be that that proposition is true or false.
- Notate these as **T** and **F**.
- So there are four worlds: $\langle \mathbf{T}, 1\rangle$, $\langle \mathbf{T}, 0\rangle$, $\langle \mathbf{F}, 1\rangle$, and $\langle \mathbf{F}, 0\rangle$.
- Let's assume to start that each of these are equally likely, i.e., our Hero has credence $\frac{1}{4}$ in each.

Learning

- Then our Hero hears a little argument by analogy that convinces them that it would be Good if the proposition in question were True.
- That is, they rule out $\langle \mathbf{T}, 0 \rangle$.
- What should happen next?

Conditionalisation

- If they update by conditionalisation, then they will change their credence in $\langle \mathbf{T}, 0 \rangle$ to 0, and their credences in the other three worlds to $\frac{1}{3}$.
- So their credence that the proposition is True will fall from 0.5 to $\frac{1}{3}$.
- That doesn't seem right.

Against Conditionalisation

- One response is to say that learning moral propositions does not involve updating by conditionalisation.
- This somewhat undermines the idea that the distribution over $\langle d, m \rangle$ is really a belief.
- Remember we could construct it out of a belief-desire pair, and maybe the fact that it doesn't update by conditionalisation is evidence it's really a hybrid, not actual beliefs.
- But this is a weak reason; we don't update de se attitudes by conditionalisation either, and we think they are beliefs.

Complicated Worlds

- Another possible response is to say that worlds are not minimal.
- Take the m in $\langle d, m \rangle$ to not be a constant, but a function from possible values of d to possible moral values.
- So m says whether each d is Good or Bad.
- Now there will be a *lot* of worlds.

A Bonus, and A Cost

- The upside is that we can once again update by conditionalisation.
- The downside is that our representation includes (somewhat essentially) differences in mental representation that make no difference to behavioural dispositions.
- This might upset some people (like me) with functionalist leanings.
- Another potential cost, though this turns on questions that are left open, is that there is no extant way to model this theory in the classical theory in any dynamically consistent way.

Two Choices

There are other options, but I think these are the most natural.

- Worlds, on the uncertaintist picture, say how things are, and how good things are.
- The ‘how good things are’ can either tell us just how good the things are in that very world, or in all worlds.
- If we just say that it is things in that world, then we have to abandon conditionalisation.
- If we say it is things in all worlds, then we have to abandon functionalism.
- This is an objection to evidential versions of uncertaintism iff you are committed to conditionalisation and functionalism, but that’s a very strong pair of commitments.

The End

If we have time, I'll run through on the board a different problem for causal versions of moral uncertaintism, but I suspect we won't have time. Instead, I'll remind you of the two choice points in the developmeant of uncertaintism.

1. Should we maximize the probability of being Good right now, or the act that has the highest conditional probability of being Good conditional on being performed?
2. Should we have a simple model, losing conditionalisation, or a complex model, losing functionalism, or some third model not yet built?