Anti-Anti-Desire-As-Belief

Anon

2024-07-29

David Lewis put forward a decision theoretic argument against there being a tight connection between desires and beliefs about the good. I argue that his argument fails twice over. It makes inconsistent background assumptions about his opponents' views, and it over-generates so broadly that if it worked, it would also rule out some standard economic models. I end with a puzzle that arises from the response to Lewis. If one responds to moral uncertainty by saying one should maximise expected moral value, how does one treat cases where one's action is evidence for or against the goodness of different actions?

A particular anti-Humean, call her Auntie, believes there is a tight connection between wanting something and believing that it is good. David Lewis (1988, 1996) has a famous argument that Auntie's view is incoherent. The point of this note is to respond on Auntie's behalf.

This is not because I agree with Auntie. On the broader question I think Lewis is right and Auntie is wrong. But Lewis's argument doesn't show that Auntie is wrong, and it's useful to see why it does not.

The point of this paper is also not to stick up for Auntie when no one has stuck up for her before. There are lots of replies to Lewis on Auntie's behalf from all sorts of directions. What's new here is that I show how two criticisms in particular, one by Huw Price (1989) and one by Jessica Collins (2015), fit together. Both of them say that Auntie should reject one of the assumptions that Lewis attributes to her. My primary contribution is to show that the two assumptions they reject are inconsistent. That is, Lewis's argument must fail because it requires attributing inconsistent background assumptions to Auntie, and that's certainly unfair.

1 The Ludovician Argument

I'll start with a presentation of Lewis's argument, shorn of what seem to me to be extraneous details.

Assume that we have a finite set of worlds. We will use w as a variable over worlds. A world, in this sense, is a specification of the truth value of all the truth-apt things that are relevant to a particular decision. The worlds in this sense are more coarse grained than Ludovician concreta in that they only specify truth values of relevant propositions, not of all propositions. That's why we can assume that there are finitely many of them. But these worlds are more fine grained than Ludovician concreta in a different sense. They will be used to represent moral uncertainty. So there can be pairs of them that are descriptively alike but evaluatively distinct. Given the supervenience of the evaluative on the descriptive, this is impossible for Ludovician worlds.

For any descriptive proposition A, assume there is a distinct proposition Å, meaning that A is good. Let V be an agent's value function, and Pr their credence function, with subscripts representing what those functions are like after updating. So V_A and Pr_A are the values of the value and credence functions after updating on A. Strictly speaking given how I've set this up, it is sets of worlds not individual worlds that get values. But I'll sometimes write $V(w)$ when strictly it should be $V(\{w\})$; I don't think this can lead to any confusion. (Later I'll also write $Pr(w)$ for the probability of $Pr(\{w\})$; again it shouldn't result in confusion.)

Lewis's argument against Auntie uses five assumptions. In these assumptions B is an arbitrary proposition, and A is an arbitrary *descriptive* proposition.

Equation The way to represent Auntie's anti-Humean view is $V(A) = Pr(\text{Å})$.

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

Additivity $V(A) = \Sigma_w V(w)Pr(w \mid A)$

Restricted Conditionalisation $Pr_A(B) = Pr(B \mid A)$

Good-Bad All worlds are either GOOD or BAD. If w is GOOD, then $V(w) = 1$, and otherwise $V(w) = 0$.

The last assumption is obviously absurd, but it is useful for setting out the argument. In any case, if the first four assumptions are true, then they should be consistent with **Good-Bad**. Given those assumptions, here is Lewis's argument.

$$\begin{aligned}
\mathrm{Pr}(\text{Å}) &= \mathrm{V}(\mathrm{A}) \\
&= \Sigma_w \mathrm{V}(w)\ \mathrm{Pr}(w \mid \mathrm{A}) && (\textbf{Additivity}) \\
&= \Sigma_w \mathrm{V}_\mathrm{A}(w)\ \mathrm{Pr}(w \mid \mathrm{A}) && (\textbf{Invariance}) \\
&= \Sigma_w \mathrm{V}_\mathrm{A}(w)\ \mathrm{Pr}_\mathrm{A}(w \mid \mathrm{A}) && (\textbf{Restricted Conditionalisation}) \\
&= \mathrm{V}_\mathrm{A}(\mathrm{A}) && (\textbf{Additivity}\text{, applied to updated values}) \\
&= \mathrm{Pr}_\mathrm{A}(\text{Å}) && (\textbf{Equation}\text{, again after updating}) \\
&= \mathrm{Pr}(\text{Å} \mid \mathrm{A}) && (\textbf{Restriced Conditionalisation})
\end{aligned}$$

But it is absurd that A and Å are independent. At least, it's absurd if evaluative uncertainty is coherent. The following situation seems perfectly possible. The agent knows someone, call them Peter, who they greatly admire. Peter faces a difficult decision; let A be that he takes one option, and B that he takes the other salient option. Right now agent thinks it is 60% likely that A is good, and 40% likely that B is good. But agent *really* admires Peter. They are sure that whatever Peter does, it will be good. So conditional on A, their credence in Å is 100%. This all seems coherent, so the conclusion of Lewis's argument must be mistaken. Lewis himself argues that independence leads to incoherence, so the last line of the argument is a reductio.

Now the argument I've given might not exactly look like the argument Lewis gives. He spends a lot of time in each paper spelling out a different reason that the independence conclusion is absurd. But despite the amount of attention Lewis gives to this issue, whether it is plausible to say A and Å are always independent is not at issue. All of Auntie's defenders in the literature (at least all the ones I've read) agree that it is not plausible. They don't try to accept the conclusion of this reductio; rather they reject one of the premises. The premises I've presented here are all ones that Lewis makes at some point or other in setting out the argument for independence. So I think the version I've given here, following Collins and Ittay-Rozen, is faithful enough to Lewis.

2 Questioning the Assumptions

Let's think a bit harder about what Auntie says about this agent who is waiting to see what admirable Peter will do. Given Auntie's view, should agent prefer that Peter makes A true, or should they be indifferent about whether Peter makes A or B true? Both options have some plausibility. On the one hand, right now the agent thinks that A is a little likelier to be good. On the other hand, whatever Peter does, the agent will be completely happy with it, because they will then think that Peter has done the right thing.

I'm not going to resolve this question for Auntie; at the end I'll come back to the question and place it in a broader philosophical context. What I do want to stress is that precisifying Auntie's view requires saying what she thinks about this question. And the assumptions one attributes to Auntie should be consistent with her answer. Lewis's argument does not satisfy this constraint, whatever Auntie says about agent and Peter.

Assume, first, that agent wants Peter to do A, because it's right now what is thought to be better. Then agent has to reject **Additivity**. **Additivity** is the rule for people who evaluate choices conditional on those things being done, not for people who evaluate choices given their current values. As Collins points out, it's the rule for one-boxing in Newcomb's Problem, and it is weird that a two-boxer like Lewis should appeal to it.

So assume, alternatively, that Auntie thinks the agent should be indifferent between Peter's choices. Then Auntie will reject **Equation**. According to **Equation**, the agent should value propositions according to their current evaluations of the goodness of the proposition. But on this assumption, the agent evaluates propositions like A and B according to how good they are thought to be conditional on obtaining. That is, Auntie's view is not **Equation**, but $V(A) = Pr(\mathring{A} \mid A)$. This is exactly what Price recommended Auntie adopt immediately after Lewis's first paper came out.

In the second paper Lewis has a response to Price's suggestion, but as Hàjek (2015) observes, it is very hard to understand what the response really is. Lewis states the kind of view Price endorses, makes a couple of observations about it, and then ends as if the question is settled. If it's meant to be a reductio, it's really not clear what the implausible conclusion is. Hàjek speculates that a paragraph or more simply went missing; the text is puzzling enough to take such speculations seriously.[1]

Whatever Auntie says about agent and Peter, she has grounds to reject one of the as-

[1] Since Hàjek's paper was published, we've had two volumes of correspondance by Lewis published (Beebee and Fisher 2020a, 2020b). But unfortunately nothing in them sheds light on this interpretative question.

sumptions Lewis attributes to her. If she says agent prefers Peter to make A true, the **Additivity** assumption should be rejected; as indeed Collins rejects it. If she says agent is indifferent between Peter's actions, the **Equation** should be rejected; as indeed Price rejects it. Attributing both **Additivity** and **Equation** to Auntie implies that Auntie inconsistently holds that agent both prefers and does not prefer that Peter makes A true. Auntie certainly has grounds to reject this attribution of inconsistent assumptions.

That is to say, while Lewis did succeed in deriving an implausible result from Auntie's view plus some auxiliary hypotheses, it is perfectly reasonable to say that the auxiliary hypotheses are to blame rather than Auntie's view. Once Auntie decides what to say about Peter and his admirer, she has a conclusive reason to reject one or other of these auxiliary hypothesis. Whatever other flaws Auntie's view has, it isn't to blame for the implausible conclusion Lewis derives from it.

3 Auntie the Capitalist

There is another assumption which Auntie should obviously reject: **Good-Bad**. Lewis acknowledges that this is a simplifying assumption, but says that we can restate Auntie's view without it. The real assumption is that there for any w, there is a numerical value for w, which measures how good it is. Let g be the function from worlds to goodness, so $\mathrm{g}(w) = x$ means that w has x units of goodness. The assumption that g is a function into the reals isn't completely trivial, but let's assume Auntie is happy to live with it. Then really what Lewis needs is **Corrected Equation**.

Corrected Equation The way to represent Auntie's anti-Humean view is $V(A) = \Sigma_w g(w) \Pr(w)$. That is, agent values A according to its expected goodness.

Lewis shows, using the same assumptions as before, that given this understanding of Auntie's view, it also leads to absurdity. And as before, I think his argument requires attributing views to Auntie that she would surely reject as soon as she makes her mind up about the agent who admires Peter. But let's say that I'm wrong about that. There is something else problematic about Lewis's argument at this point.

Nothing else in Lewis's argument turns on the fact that g is a measure of goodness. The argument goes through just as well (or just as badly) for any numerical function that g could be. That function could be a measure of anything. It could, for instance, be a measure of how much profit agent makes in *w*. In that case, **Corrected Equation** says that agent values propositions according to their expected profitability. That's just the standard theory of the firm from basic economics. If Lewis's argument shows that Auntie's view is inconsistent, it also shows that the standard theory of the firm is inconsistent.

To be sure, there is a lot wrong with the 'standard theory of the firm' as a theory of either real or idealised firms. But it's rather implausible that it's trivial, and particularly implausible that it could be shown to be trivial by some simple decision theory. That would be particularly ironic given how much of decision theory was developed to explain decision making by idealised firms.

The lesson here is that Lewis's argument over-generates. If it shows anything, it shows that having one's valuation track the expected value of any numerical measure is inconsistent. But that can't be right. Having no priority in the world other than maximising expected profits might be morally abhorrent, but it isn't inconsistent with decision theory. So Lewis's argument must be wrong.

4 A Puzzle

That completes my objection to Lewis's argument. I'll end with a puzzle that arises from the discussion here.

Go back to agent the agent who thinks that whatever Peter does will be correct. Change the case so that agent is in fact Peter. That is, in this version of the case, Peter isn't sure what's right, and isn't sure what he'll do, but is sure that whatever he does will be good. Assume that Peter only cares about maximising the good, even when he doesn't know what is in fact good.[2] Question: What should Peter want to have happen, assuming all this?

I can think of at least four coherent responses to this question.

First, one might think that since Peter thinks A is more likely to be good, and he wants to do good, he should make A true.

[2] Perhaps Peter was convinced by the arguments from MacAskill, Bykvist, and Ord (2020) that this is what he should do.

Second, one might think that since Peter will be sure he does the good thing whatever he does, he should be indifferent between making A true and making B true.

Third, one might think that this is really just a special case of Newcomb's Problem, where maximising expected utility according to unconditional probabilities (over states causally independenrt of one's action) gives a different recommendation to maximising expected utility according to conditional probability. This answer says that whatever you say about Newcomb's Problem, whether you say conditional probabilities are to be used (as most one-boxers say), or unconditional probabilities are to be used (as some two-boxers say), the same goes here.

The third answer isn't inconsistent with the previous two. But it is a distinct answer. It is an answer that people who disagree about Newcomb's Problem can agree is correct. But it's also a substantive claim, since there is a coherent way to deny it. In particular, it is coherent to adopt the version of causal decision theory that Lewis (1981) defends for descriptive uncertainty, and something that looks like evidential decision theory for moral uncertainty.

Here is how that might go. Loosely following Bradley and List (2009), let worlds be ordered pairs $\langle d, v \rangle$, such that d settles the (relevant) descriptive facts, and v is a numerical measure[3] of the goodness of the world.[4] In the terminology used earlier $g((d, v)) = v$; the second term is how good the world is.

Let K be a partition of the worlds such that whatever the agent does makes no causal difference to which member of the partition is actual. Intuitively, the true element of K is the conjunction of all the facts that are outside the causal control of the chooser. Crucially, K settles the *facts* outside the agent's causal control; it does not settle anything evaluative.[5] For any proposition A, the value to the agent of A is given by this equation:

$$V(A) = \Sigma_{k \in K} \Pr(k)\ \Sigma_{\langle d,\, v \rangle \in k}\ v\Pr(\langle d, v \rangle \mid A \wedge k)$$

The agent should then make true the proposition with the highest value that it is within their power to make true. The inner sum in this equation looks like preferred definition of decision-theoretic value for Evidential Decision Theorists. In this respect I'm following Lewis closely. As he says, "Within a single dependency hypothesis, so to speak, V-maximising [i.e., Evidential Decision Theory] is right." (Lewis 1981, 7). The idea here is that if Lewis could be right about this claim, and all moral uncertainty takes place within dependency hypotheses, then the puzzle here will not be just like Newcomb's Problem.

[3] At this point I differ from Bradley and List. They take the v part of $\langle d, v \rangle$ to be something like an evaluative theory, something that tells you how different things are to be valued. I'm taking it just to be a number, saying how good things are. This is why I think **Invariance** is plausible for worlds thus understood. The way I'm setting things up, **Invariance** is just the claim that the value of a terminal node doesn't change depending on where you are in a decision tree.

[4] This way of thinking about worlds helps explain some terminology that I left undefined earlier. A descriptive proposition is a proposition p such that for any d, v, and v', if $\langle d, v \rangle$ and $\langle d, v' \rangle$ are worlds, then $\langle d, v \rangle \in p$ iff $\langle d, v' \rangle \in p$.

[5] That is, all the cells of the partition are descriptive propositions.

Finally, one might think this version of the Peter example is incoherent. Couldn't one simply think about what to do, and then having made a decision, learn what the admirable person thinks is right, and hence what is right? Well, maybe it's not so simple. Maybe one thinks that one always acts in the right way, even if one's thoughts are not always right. Perhaps one has an inner voice, a la Socrates, that prevents one from *acting* the wrong way, but which only kicks in at the moment of action.

I'm inclined to think that last possibility, where one is somewhat confident that one will somehow find oneself unable to act wrongly, is just conceivable enough for the example to be coherent. Just like with Newcomb's Problem, all that's needed to get the problem going is that the action is some evidence of some underlying fact. In Newcomb's Problem we can get a difference between Evidential Decision Theory and Causal Decision Theory even with an imperfect demon, as long as their predictions are known to be better than chance. In this case, we can get a difference between maximising conditional expected goodness and unconditional expected goodness as long as the decider thinks their action is some evidence that they did the right thing. Is that a coherent assumption to make? I think it probably is, and if so, it raises an interesting question about the details of views on moral uncertainty.

References

Beebee, Helen, and A. R. J. Fisher, eds. 2020a. *Philosophical Letters of David k. Lewis*. Vol. 1. Oxford: Oxford University Press.

———, eds. 2020b. *Philosophical Letters of David k. Lewis*. Vol. 2. Oxford: Oxford University Press.

Bradley, Richard, and Christian List. 2009. "Desire-as-Belief Revisited." *Analysis* 69 (1): 31–37. https://doi.org/10.1093/analys/ann005.

Collins, Jessica. 2015. "Decision Theory After Lewis." In *A Companion to David Lewis*, edited by Barry Loewer and Jonathan Schaffer, 446–58. John Wiley; Sons.

Hàjek, Alan. 2015. "On the Plurality of Lewis's Triviality Results." In *A Companion to David Lewis*, edited by Barry Loewer and Jonathan Schaffer, 425–45. John Wiley; Sons.

Lewis, David. 1981. "Causal Decision Theory." *Australasian Journal of Philosophy* 59 (1): 5–30. https://doi.org/10.1080/00048408112340011.

———. 1988. "Desire as Belief." *Mind* 97 (387): 323–32. https://doi.org/10.1093/mind/xcvii.387.323.

———. 1996. "Desire as Belief II." *Mind* 105 (418): 303–13. https://doi.org/10.1093/mind/105.418.303.

MacAskill, William, Krister Bykvist, and Toby Ord. 2020. *Moral Uncertainty*. Oxford: Oxford University Press.

Price, Huw. 1989. "Defending Desire-as-Belief." *Mind* 98 (389): 119–27. https://doi.

org/10.1093/mind/XCVIII.389.119.