Anti-Anti-Desire-As-Belief

Anon

2024-11-01

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?

1 The Cast

A particular kind of 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. The response will be in an important respect not at all novel. What I'm going to do is essentially put together the responses offered by Huw Price (1989) and Jessica Collins (2015), and show that between them, they offer a response that any

Auntie can accept, and so Lewis's argument shouldn't make any of them lose a moment's sleep.

I'm going to distinguish two kinds of Aunties, I'll call them Auntie C and Auntie E, and argue that both of them can satisfy the description *believes there is a tight connection between wanting something and believing that it is good*, and that they each have a principled reason to reject one of the premises in Lewis's anti-Auntie argument. I'll then argue that there is no one who should have accepted all the premises of Lewis's argument in the first place.

To do this, we need one last character. Auntie is of course an Auntie of someone; call the someone Peter. Naturally enough, Auntie thinks the world of Peter. Both Auntie E and Auntie C (who are both aunts of Peter), both think that if Peter does something, it's very likely to be the right thing to do. Indeed, they are fairly deferential to Peter in this respect; if Peter does something they thought was wrong, they take that as some (strong but inconclusive) reason to change their belief about the morality of the action.

Peter is currently contemplating some action *a*. Let A be the proposition that Peter does *a*, and Å be the proposition that doing *a* is good. Unless stated otherwise, I'll follow Lewis in making the simplifying assumption that all actions are good or bad. This simplifies a lot, though eventually we'll have to drop it because it simplifies too much. In this case, I'll also make the similar simplifying assumption that it's common knowledge that not doing *a* is good iff doing *a* is not good, so ~Å can be read as either of the epistemically

equivalent propositions *a is not good* and *not a is good*.

Before Peter acts, both Auntie's have the same credal distribution, satisfying these constraints.

- C(Å | A) = 0.8
- C(~Å | ~A) = 0.9
- C(A) = 0.7

?@tbl-credence shows the credence each Auntie has in each of the four possibilities from crossing A with Å.

	Å	¬Å
A	0.56	0.14
¬A	0.03	0.27

Auntie's credence that Peter will do A, and that it will be right. {#tbl-credence}

Now you might think at this point that I've said enough to tell you what each Auntie hopes Peter will do. After all, I've told you everything relevant about each Auntie's credence in Å, and I've told you that their credences in propositions about goodness determine their values. But I haven't told you one thing extra - I haven't told you what decision theory the two Aunties follow.

Auntie C endorses a version of causal decision theory, in particular something like the version supported by David Lewis (1981).[1] In particular, Auntie's values are given by **Auntie C's Value**. In it, *x* is an arbitrary action, X is the proposition that *x* is performed, and G is the proposition that a good action was performed. In the formula, C_x be the result of *imaging* the credence function C on the proposition *x* is performed. Auntie C believes changing the moral facts is a bigger change to the world than changing any descriptive facts, so imaging always moves credences up or down in **?@tbl-credence**, never left or right.[2]

Auntie C's Value V(x) = C_X_(G)

This resembles equation (11) in "Causal Decision Theory". Indeed, after the first character, it just is the special case of that equation where the only possible values are 1 and 0. But the first character matters. Lewis is presenting a theory of usefulness, not of value. His formula is meant to measure the thing that a rational actor maximises. It is not measuring the thing an altruistic friend hopes is maximised. We'll come back to this point in **?@sec-open-question**. For now, I just want to note the similarities to Lewis's own theory.

Using this formula, Auntie C hopes that Peter does a iff C(Å) > C(¬Å). Since

[1] Here I'm following Collins, who notes that it is odd that Lewis attributes to Auntie a form of evidential decision theory, which Lewis himself does not endorse.

[2] At this point you might worry that talk of changing the moral facts is incoherent, given the supervenience of the normative on the descriptive. But we're trying to model Auntie's mental states here, and she's morally uncertain. So the worlds have to be epistemic possibilities, not metaphysical possibilities. For more on this, see (**?**).

C(Å) = 0.59, and C(¬Å) = 0.41, that means she does hope that Peter does A.

Auntie E's view is easier to state. She is an evidential decision theorist. For her, the value of an arbitrary action x is given by this formula.

$$V(x) = C(G \mid X)$$

That is, she looks at Peter's options, and hopes that he does the one that she is most confident is good, conditional on Peter doing it. That means she hopes Peter does not do a, since then she'll have credence 0.9 that Peter has done the right thing. If Peter does a, she'll only have credence 0.8 that he'll have done the right thing, which isn't as good.

That puts our players on the stage. It's now time to introduce Lewis's argument.

2 The Ludovician Argument

I'll present here a somewhat simplified version of Lewis's argument, following the presentation in Collins (2015) and (**?**). This version is somewhat simpler than Lewis's, but it's easy to find parts in Lewis's text where he endorses every premise here, and the premises are sufficient to get the result. So I think it's fair to attribute the argument to him.

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.[3]

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 C their credence function, with superscripts representing what those functions are like after updating. So V^A and C^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 $C(w)$ for the probability of $C(\{w\})$; 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 $V(A) = C(Å)$.

Invariance $V^A(w) = V(w)$

[3]That we're using metaphysically impossible worlds should not cause any concern here. As (**?**) argues, we want the quantifiers here to range over epistemically possible worlds. For people who are unsure of the moral facts, as happens here, that means that there are some metaphysically impossible worlds that are epistemically possible.

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

Restricted Conditionalisation $C^A(B) = C(B \mid A)$

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.

$C(Å)$	$= V(A)$	
	$= \Sigma_w V(w)\, C(w \mid A)$	(**Additivity**)
	$= \Sigma_w V^A(w)\, C(w \mid A)$	(**Invariance**)
	$= \Sigma_w V^A(w)\, C^A(w \mid A)$	(**Restricted Conditionalisation**)
	$= V^A(A)$	(**Additivity**, applied to updated values)
	$= C^A(Å)$	(**Equation**, again after updating)
	$= C(Å \mid A)$	(**Restriced Conditionalisation**)

But it is absurd that A and Å are independent. At least, it's absurd if evaluative uncertainty is coherent. It's possible to be like Auntie and not know whether A is true, but think that it being true is evidence that it is good.

At this point Lewis does not rely on the intuition that Auntie's case is possible, but instead gives an argument that A and Å cannot always be independent. It goes roughly like

this.

1. It is possible that someone could have positive credence in all four cells of **?@tbl-credence**.
2. It is further possible that such a person could learn A ∨ Å and nothing else.
3. If such a person learned A ∨ Å and nothing else, they should update by conditionalisation on A ∨ Å.
4. From 1-3, plus some algebra, it follows that A and Å will not be independent post-conditionalisation.
5. But the earlier argument shows that Desire-as-Belief implies A and Å will be independent after any possible learning experience.

The problem is that step 3 is not particularly plausible. If moral beliefs have centered world contents, the kind of content described in (**?**), then they probably should not be updated by conditionalisation. On most views of learning what time it is, the learner does not update by conditionalisation. And moral beliefs plausibly do have centered world contents; that is a natural interpretation of the meta-ethical view that Lewis himself endorses (**?**).

I don't mean to endorse the objection to premise 3 in the last paragraph; I think the questions here are rather tricky. I do mean to endorse two other claims. One is that just thinking about cases like Peter and his Aunties gives us a stronger reason to reject independence than Lewis's argument from conditionalisation. Second, if one does think that

independence is independently implausible, Lewis only needs **Restricted Conditionalisation** for the rest of the argument to work. And **Restricted Conditionalisation** is plausible, even if one rejects conditionalisation for centered worlds contents.

3 Aunties' Responses

With the argument in front of us, it's easy enough to see how the Aunties can respond. Auntie E rejects **Equation**, and Auntie C rejects **Additivity**. The main point of this section is to argue that each Auntie has a principled reason to reject the premise in Lewis relies on. A secondary point is to argue that this means any desire-as-belief theorist has a reason to reject Lewis's argument.

Let's start with Auntie E. She objects to Lewis's argument in the same way Huw Price (1989) does - by rejecting **Equation**. According to **Equation**, Auntie should value propositions according to her current evaluations of the goodness of the proposition. But Auntie E thinks that gets Peter's case wrong, and for a simple reason. How valuable a proposition is not a function of how likely it is to be good, but how likely it is that it would be good, were it true. That is, Auntie's view is not **Equation**, but $V(A) = Pr(Å \mid A)$. This is exactly what Price recommended Auntie adopt immediately after Lewis's first paper on desire-as-belief came out.

Lewis's response to Auntie E is rather hard to follow.[4] As he says in a letter to Michael

[4] Hàjek (2015) speculates that a paragraph or more simply went missing. That's not particularly plausible,

McDermott (Lewis [1993] 2020), he doesn't think Auntie E has a desire-as-belief theory. To see why, it helps to say more about Auntie E's view.

Following Gibbard (1990), she takes beliefs to be sets, and the members of those sets to be pairs. In particular, they are factual-normative pairs, and an arbitrary one is denoted by $\langle f, n \rangle$. Gibbard uses $\langle w, n \rangle$, but Auntie E worries that this leads to confusion about what a *world* is. After all, she thinks, worlds set the value of everything the decision maker needs, and the w part of Gibbard's pairs does not. Using f has the added advantage that if one is suspicious about a fact-value distinction here, you can let it stand for *fusion*; the first element of each pair is a fusion of things. It's the kind of thing Lewis (1986) thinks is a possible world.

If we adopt the simplifying assumption that everything is GOOD or BAD, then n will be either 1 (for GOOD), or 0 (for BAD). Then there is a special proposition, which Lewis calls G. It is $\{\langle f, n \rangle : n = 1\}$. As Lewis notes, Auntie E's theory is that by necessity, the agent will have a desire for G. This isn't reduced to belief because it is, as he says in the letter to McDermott, "independent of belief". (In the 1996 paper, Lewis)

- Need to make them triples of facts, centers, and norms, not doubles
- Need to say that the necessity here is

Now let's consider Auntie C. She thinks that **Additivity** is wrong because it gets Peter's case wrong. She thinks it gets the case wrong for a simple reason: it's the way evidential

especially for a paper that was reprinted twice in Lewis's lifetime, but it does speak to the oddness of the argument.

decision theorists value gambles, and she's a causal decision theorist, not an evidential decision theorist.[5]

Again, Lewis has a response. He says that this is to confuse choice-worthiness and desirability.

> A famous difficulty need not concern us here. Suppose a certain action would serve as an effective means to your ends. Yet at the same time it would constitute evidence—evidence available to you in no other way—that you are predestined inescapably to some dreadful misfortune. Should you perform that action?—Yes; your destiny is not a consideration, since that is outside your control. Do you desire to perform it?—No; you want good news, not bad. Since our topic here is not choiceworthiness but desire, and since the two diverge, we adopt an "evidential" conception of expected value, on which the value of the useful action that brings bad news is low. Choiceworthiness is governed by a different, "causal", conception of expected value. (Lewis 1996, 304)

Auntie finds this very puzzling. Imagine a person, call them Rosencrantz, is playing Newcomb's Problem. It sounds like Lewis is saying that Rosencrantz should (a) take both boxes, and (b) desire that they take one box. But that means that their actions won't flow from their desires in the way that Humeans, like Lewis, say that they should. Indeed, one

[5] So far Auntie C is following closely the argument in Collins (2015), though she will ultimately disagree with Collins on a point of detail: Auntie accepts **Restricted Conditionalisation**.

of the motivations of the first Desire as Belief paper, right on the first page, was preserving the link between desire and action.

So Auntie C is unmoved. She wants what is best, and she doesn't want people to do things just so she'll have evidence that things are better than she thought.

Let's sum up. Lewis showed that it is impossible to hold onto the four named principles, Auntie's anti-Humean view, and probabilistic dependence between A and Å. The argument that this combination is untenable seems sound. And I think (though some disagree) that we should hold on to **Invariance**, **Restricted Conditionalisation** and the possibility of probabilistic dependence. The question then is which of these three to give up:

1. Anti-Humeanism
2. **Equation**
3. **Additivity**

My argument is that Lewis has not given us a compelling reason to reject 1, rather than rejecting 2 or 3. What's tricky about the case is that it's hard to say which of 2 or 3 we should reject. One kind of anti-Humean, Auntie E, rejects 2, since she thinks that desirability tracks the conditional probability of an act being good. Another kind of anti-Humean, Auntie C, rejects 3, since she thinks desirability goes with causal choice-worthiness. Both these rejections seem principled, so either anti-Humean has a principled reason to reject Lewis's argument.

4 A Decision Problem

To back this point up, imagine Peter is not making a morally loaded decision. Instead he's playing a version of the problem (**?**) calls Nice Demon. The rules are simple.

1. Peter will choose Up or Down.
2. Demon will try to predict Peter's choice.
3. If Demon's prediction is correct, Peter gets $1. Otherwise he gets nothing.
4. Demon is 0.9 likely to correctly predict Down (if Peter chooses it), and 0.8 likely to correctly predict Up (if Peter chooses it).

Both Aunties start deliberation with credence 0.7 that Peter will choose Up. Question: What does Auntie hope that Peter will do? Answer: They disagree.

Auntie C hopes that Peter will choose Up. She has credence 0.59 that Demon has predicted Up, and she wants Peter to get the money. This case is, for her, another counterexample to **Additivity**.

Auntie E hopes that Peter will choose Down. She agrees with Lewis on what is desirable, i.e., she thinks desire goes with evidential choice-worthiness. This case is, for her, another reason to reject **Equation**.

Who should accept both **Addivity** and **Equation**? Only someone who thinks Peter should choose both Up and Down. That's implausible, so

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.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.

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. "Are We Free to Break the Laws?" *Theoria* 47 (3): 113–21. https://doi.org/10.1111/j.1755-2567.1981.tb00473.x.

———. 1986. *On the Plurality of Worlds*. Oxford: Blackwell Publishers.

———. 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.

———. (1993) 2020. "Letter to Michael McDermott, 6 December 1993." In *Philosophical Letters of David K. Lewis*, edited by Helen Beebee and A. R. J. Fisher, 2:508. 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.