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 Lewis's Argument

David Lewis (1988, 1996) has an argument against the view that desires can be reduced to beliefs. I'm going to respond on behalf of the Desire-as-Belief (hereafter, DAB) thesis. In particular, I'm going to offer two responses. The responses themselves will not be particularly original; one is due to Huw Price (1989), and the other to Jessica Collins (2015). What will I hope be original is showing how they fit together. I'll argue that

given the other assumptions Lewis makes, the DAB theorist has independent reason to reject one or other of the premises Lewis offers.

I'll start with a presentation of Lewis's argument, shorn of what seem to me to be extraneous details. This will look a little different to Lewis's own presentation, but all the premises here are ones he uses, and its gets to the conclusion he wants, so it seems to be a fair representation.[1]

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. They are 'small worlds' in the sense of Savage (1954), not possible worlds in the sense of Lewis (1986). For the most part they are more coarse-grained than the concrete worlds of *Plurality*, but in one respect they are more fine-grained: they might differ from each other purely in normative features.[2]

For any descriptive proposition A, assume there is a distinct proposition Å, meaning that A is desirable.[3] 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

[1] I'm drawing here on presentations of the argument by Collins (2015), Hàjek (2015), and Nissan-Rozen (2015).

[2] This is a consequence of the point Chalmers (2011) makes that it is epistemic, not metaphysical, possibility that matters here, plus the assumption that our characters might be normatively uncertain.

[3] It's more common here to let Å be that A is good, but I want to separate questions about the desire-belief relationship from questions about moral motivation, so my characters will be motivated to do something just to the extent they believe it desirable, leaving it as a further question whether all and only good things are desirable.

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 DAB uses six assumptions. In these assumptions B is an arbitrary proposition, and A is an arbitrary *descriptive* proposition.

Binary Desirability All worlds are either desirable or undesirable. If w is desirable, then $V(w) = 1$, and otherwise $V(w) = 0$.

Equation Given **Binary Desirability**, the correct way to represent DAB is $V(A) = Pr(Å)$.

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

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

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

Possible Independence For some Å, $Pr(Å) \neq Pr(Å \mid A)$

The first assumption is obviously absurd, but it is useful for setting out the argument. In any case, if the last five assumptions are true, then they should be consistent with **Binary Desirability**. Given those assumptions, here is Lewis's argument. By **Possible Independence** there is an Å such that $Pr(Å) \neq Pr(Å \mid A)$. But given the other assumptions we can reason as follows.

$Pr(Å) = V(A)$ (**Binary Desirability** + **Equation**)

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

And the last line contradicts our assumption. So not all six of these assumptions can be correct, if DAB is true. Since Lewis thinks they are all correct, he concludes DAB is false.

As I said, this presentation doesn't look a lot like Lewis's argument. Most notably, **Possible Independence** isn't an assumption for Lewis, it is something he derives from yet further premises. Much of the 1988 paper is devoted to spelling out the absurd consequences of denying **Possible Independence**. These arguments haven't convinced everyone. They assume that conditionalisation is the right way to update on any new information, descriptive or normative. If normative propositions are centered worlds, as the picture in Lewis (1989) suggests, that seems like the wrong way to update. If the picture of self-locating belief that Lewis (1979) offers is correct, we can't update our beliefs about the time by conditionalisation when the alarm clock goes off.[4]

[4] The point that conditionalisation isn't the right way to update beliefs with centered worlds contents, and this raises a problem for Lewis, is made by Graham Oddie (1994).

But **Possible Independence** is surely true. Assume that A is a proposition about someone, call him Peter, might do. And assume that we desire that the morally good thing is done, that we don't know whether A is good or not, but we are very confidence in Peter's moral judgment. If Peter does A, that's good evidence A is good. That all seems coherent, which is enough to support **Possible Independence**.

While unrestricted conditionalisation is questionable, **Restricted Conditionalisation** seems fairly secure. In some presentations of the argument, Lewis uses a version of invariance that says the value of any proposition does not change on learning A. This is questionable, but **Worldly Invariance** seems fairly secure. It's just the view that in a decision tree, the value of a terminal node doesn't depend on where we are in the tree. That's normally taken for granted in formal models of dynamic choice, and I think rightly so.

If we treat **Binary Desirability** as a harmless simplification, that means the only substantive assumptions left are **Equation** and **Additivity**. Given those, the argument against DAB goes through. I'm going to argue that anyone sympathetic to DAB has independent reason to reject one or other of those claims. Part of the argument that this is an *independent* reason is that different sympathisers will reject one rather than the other. To show this, I'll start with a short story.

2 Auntie and Auntie

Our heroes are two anti-Humeans, called Auntie E and Auntie C, who both endorse a version of DAB. Both of them are aunts of Peter, the moral exemplar from Section 1. Both Aunties E and Auntie C 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. They are both moralists, and think something is desirable iff it is good.

Let A be the Proposition that Peter does some action *a*, and Å that it is deisrable/good. Like Lewis does with **Binary Desirability**, I'll make a simplifying assumption: it's common knowledge that not doing *a* is good iff doing *a* is not good. That means ~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

Table 2 shows the credence each Auntie has in each of the four possibilities from crossing

A with Å.

Table 2: Auntie's credence that Peter will do A, and that it will be right.

	Å	¬Å
A	0.56	0.14
¬A	0.03	0.27

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 E is an evidential decision theorist. For her, the value of an arbitrary action x is given by **Auntie E's Value**. In this formula, where X is the proposition that x is performed, and D is the propositions that things are desirable.

Auntie E's Value $V(x) = C(D \mid X)$

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.

Auntie C endorses a version of causal decision theory, in particular something like the version supported by David Lewis (1981).[5] In particular, Auntie's values are given by **Auntie C's Value**. 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 Table 2, never left or right.[6]

Auntie C's Value $V(x) = C_X(D)$

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 Section 4. 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 $C(Å) > C(\neg Å)$. Since $C(Å) = 0.59$, and $C(\neg Å) = 0.41$, that means she does hope that Peter does A.

The next two steps are to see why the Auntie's reject Lewis's argument.

[5] 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.

[6] Recall here that the worlds are epistemic possibilities, not metaphysical ones, so it makes sense to talk about merely changing the moral facts.

3 DAB and EDT

The easier case is Auntie E. She rejects **Equation**. The relationship between desire and belief is not V(A) = Pr(Å), but V(A) = Pr(Å | A). Lewis is aware of this response, it's developed by Price (1989), and his response is that this isn't a form of desire as *belief*. His thought, and this comes out a little more clearly in a recently published letter than in the papers (Lewis [1993] 2020), is that belief isn't load-bearing in Auntie E's view.

Everyone agrees that beliefs play a role in instrumental desires. If desires to take a pill because one believes it will cure one's disease, that desire will go away if one loses either the belief that it's a cure, or the belief that one is diseased. Lewis notes that Auntie E is committed to the existence of a proposition D consisting of all and only the desirable worlds, and to the claim that by necessity any agent with desires will desire it. Moreover, he argues, on Auntie E's view this is the only non-instrumental desire an agent has. Beliefs don't affect non-instrumental desires on this view, since everyone has the same non-instrumental desires. They just affect instrumental desires. But Humeans and anti-Humeans agree about the connection between belief and non-instrumental desires.

I'll offer three replies on Auntie E's behalf. I'm not defending her view in general; I'm not an evidential decision theorist. I'm just defending the claim that this is a kind of desire as beleif.

First, the simplifying assumption **Binary Desirability** looks less benign here. The construction of the necessarily desired proposition D requires this assumption. If some

worlds are more desirable than others there are still preferences that Auntie E thinks are necessary (i.e., that a more desirable world is preferred to a less desirable one), but no desires.

Second, the necessity claim Auntie E seems committed to looks less worrying once remember what kinds of things worlds are in this context. They are the elements of the contents of attitudes. That is, on Lewis's view, they are centered worlds. Auntie E's view is really that by necessity an agent desires that one of the centered worlds found desirable by the current center, i.e., that very agent, is actualised. That is, agents desire what they believe desirable. That's not a surprising necessity claim; it's the essence of the view Lewis is arguing against.

Third, this is a very odd complaint from *David Lewis*. It's agreed on all sides that on Auntie E's view, what an agent desires supervenes on what they believe. Moreover, the reverse supervenience, of belief on desire, probably doesn't hold. Lewis's complaint here is that Auntie E hasn't made desire depend on belief in the right way, even though she has ensured these supervenience claims hold. The only way to make this complaint work is to understand dependence using some hyperintensional notion like grounding. But there's no sign of such a notion in Lewis's work, and there are good reasons to think it couldn't be added to his view (MacBride and Janssen-Lauret 2022).

So I think EDT offers a way out of Lewis's argument. That's not totally surprising; it's not Lewis's view. Things are trickier with Auntie C.

4 DAB and CDT

Auntie C says that the misstep in Lewis's argument is **Additivity**. Following Collins, she says that this is something that only an adherent of EDT should accept. She's mostly mystified about why Lewis, famously an enemy of EDT, would have accepted it in the first place.

In the second DAB paper, Lewis has a response to this. Oddly, it's in a parenthetical paragraph. After describing an action that's essentially two-boxing in Newcomb's problem, he writes

> Should you take that actions?—Yes ... Do you desire to perform it?—No ... [O]ur topic here is not choiceworthiness but desire ... [so] we adopt an "evidential" conception of expected value ... Choiceworthiness is governed by a different "causal" conception. (Lewis 1996, 303)

So Lewis thinks that Auntie C is confusing the two notions, and that she only rejects **Aditivity** because of this confusion. But there's something important that can be said on Auntie C's behalf.

In the first DAB paper, Lewis sets out the notion of desire that he thinks is at issue. There, he is most concerned to distinguish it from a notion that is intuitively used in actions taken from duty. We might intuitively say that we did X from duty, though we desired to do Y. Lewis resists; he thinks that in any such case we also desire X. As he puts it,

> We are within our rights to construe 'desire' inclusively, to cover the entire range of states that move us. (Lewis 1988, 323)

This isn't an optional move on Lewis's part. If he doesn't make this move, the Humean picture of action he wants to defend fails immediately.

Now Auntie C says that if we're modelling "the entire range of states that move us", and we are, as Lewis recommends, moved to take two boxes in Newcomb's Problem, we really better not insist that our model satisfies **Additivity**. We'd be left saying that the two-boxer acts against all desire, merely motivated by the duty to do what's rational. This would be a violation of Lewis's Humeanism. It would also be phenomenologically implausible. The two-boxer isn't moved by duty, but by the desire for the extra $1000. So for both theoretical and phenomenological reasons, we need a notion of desire that violates **Additivity**.

Indeed, she insists that it isn't just in first-personal cases like Newcomb's Problem that **Additivity** fails. Imagine that Peter faces the choice in Table 3, where the values in the box now represent the moral value of the choice, and Auntie C thinks Peter choosing Up is excellent evidence for Left, while his choosing Down is excellent evidence for Right.

Table 3: Peter's second choice

	Left	Right
Up	3	0

	Left	Right
Down	5	1

The first quote of Lewis's suggests that both Auntie's should desire that Peter choose Up, since that will be evidence that an outcome of value 3 will obtain. But it seems coherent to hope that he chooses Down. Several of the arguments for two-boxing seem replicable here. It's weird to hope that Peter chooses Down conditional on Left, and hope he chooses Down conditional on Right, and unconditionally hope that he chooses Up.

This is not to deny that there is a theoretical role for something like news-value. Joyce (1999) makes extensive use of that notion in developing a causal decision theory. It's just that in the context of defending Humeanism about action, two-boxers like Lewis can't equate news-value with desirability. So **Addition** has to fail, since as Lewis agrees, **Addition** entails that desirability goes with news value.

5 Conclusion

Any defender of DAB has to pick whether to take Auntie E's side or Auntie C's side in the original question about Peter. Whichever side they pick, they will have an independent reason to reject one of the premises Lewis puts forward in his argument against their view. So they have independent reason to reject Lewis's argument.

I've been stressing *independent* here because it's not news that any defender of DAB has to reject some premise of Lewis's argument. That follows from the fact that his argument is valid! For the response to be more than table-thumping, the proponent of DAB has to say which premise they reject, and why it is reasonable to reject it. The point of the examples involving Peter is to identify the premise in question, which will differ between different proponents, and say what that reason is. In the second DAB paper, Lewis offers responses to each of these reasons, and in the last two sections I've gone over why those responses don't work. It's somewhat ironic that in each case my analysis has rested in part on the view that Lewis's response is inconsistent with what he says elsewhere: about the role of possible worlds in philosophy in Section 3, and about the role of desire in producing action in Section 4. I'll leave to another day whether someone who was willing to abandon large parts of the Lewisian framework might be able to rescue his argument against DAB.

Chalmers, David. 2011. "Frege's Puzzle and the Objects of Credence." *Mind* 120 (479): 587–635. https://doi.org/10.1093/mind/fzr046.

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.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge: Cam-

bridge University Press.

Lewis, David. 1979. "Attitudes *de Dicto* and *de Se*." *Philosophical Review* 88 (4): 513–43. https://doi.org/10.2307/2184646.

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

———. 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.

———. 1989. "Dispositional Theories of Value." *Aristotelian Society Supplementary Volume* 63 (1): 113–37. https://doi.org/10.1093/aristoteliansupp/63.1.89.

———. 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.

MacBride, Fraser, and Frederique Janssen-Lauret. 2022. "Why Lewis Would Have Rejected Grounding." In *Perspectives on the Philosophy of David K. Lewis*, edited by Helen Beebee and A. R. J. Fisher, 66–91. Oxford University Press. https://doi.org/10.1093/oso/9780192845443.003.0005.

Nissan-Rozen, Ittay. 2015. "A Triviality Result for the "Desire by Necessity"thesis." *Synthese* 192 (8): 2535–56.

Oddie, Graham. 1994. “Harmony, Purity, Truth.” *Mind* 103 (412): 451–72. https://doi.org/10.1093/mind/103.412.451.

Price, Huw. 1989. “Defending Desire-as-Belief.” *Mind* 98 (389): 119–27. https://doi.org/10.1093/mind/XCVIII.389.119.

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