Anti-Anti-Desire-As-Belief

Anon

2026-01-20

David Lewis put forward a decision-theoretic argument against the Desire-as-Belief (DAB) thesis. His argument relies on several assumptions, including what I'll call Equation ($V(A) = Pr(Å)$) and Additivity ($V(A) = \Sigma_w V(w)Pr(w \mid A)$). I argue that any defender of DAB has available a response to Lewis's argument, though which response depends on their broader decision-theoretic commitments. Those who favor evidential decision theory have reasons to reject Equation in favor of a conditional formulation ($V(A) = Pr(Å \mid A)$), while those who favor causal decision theory have reasons to reject Additivity. Lewis's argument against DAB only works against a theorist who, like Lewis himself, has a view of decision theory which is somewhat causal and somewhat evidential. Without a much stronger argument that this is the unique correct theory, the DAB thesis seems safe.

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 it 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 *On the Plurality of Worlds*, 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

[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.

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 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 Dependence 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 Dependence** there is an Å such that $Pr(Å) \neq Pr(Å \mid A)$. But given the other assumptions we can reason as follows.

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

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 Dependence** 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 Dependence**. 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 Dependence** 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 Dependence**.

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.

- $C(Å \mid A) = 0.8$
- $C(¬Å \mid ¬A) = 0.9$
- $C(A) = 0.7$

Table 1 shows the credence each Auntie has in each of the four possibilities from crossing A with Å.

Table 1: 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

At this point there is a puzzle we can generate using just **Equation** and **Additivity**. I'll assume that $V(A \wedge Å) = V(\neg A \wedge \neg Å) = 1$, while $V(A \wedge \neg Å) = V(\neg A \wedge Å) = 0$, and that these four combinations are worlds for the purpose of **Additivity**. By **Equation**, $V(A) = 0.59 > 0.41 = V(\neg A)$. On the other hand, **Additivity** gives the opposite ranking. (In this little proof I'll suppress conjunctions, so AÅ is short for A∧Å.)

Spelling out **Addition**, we get $V(A) = V(AÅ)Pr(AÅ|A) + V(A\neg Å)Pr(A\neg Å|A) + V(\neg AÅ)Pr(\neg AÅ|A) + V(\neg A\neg Å)Pr(\neg A\neg Å|A)$. The last three terms all contain a 0, since $V(A\neg Å)$ and $Pr(\neg A|A)$ are 0. Since $V(AÅ) = 1$, and $Pr(AÅ|A) = Pr(Å|A)$, this reduces to $V(A) = Pr(Å|A)$. A similar argument shows that **Addition** entails that $V(\neg A) = Pr(\neg Å|\neg A)$. So we have $V(A) = Pr(Å|A) = 0.8 < 0.9 = Pr(\neg Å|\neg A) = V(\neg A)$. In other words, given our assumptions about Aunties' credences, we have both $V(A) > V(\neg A)$ and vice versa.

One could take this to be a simpler argument than Lewis's for the same conclusion. DAB leads to contradiction, even without **Worldly Invariance** or **Restricted Conditionalisation**. As long as the Peter case is possible, and it certainly looks possible, **Equation** and

Additivity suffice for a contradiction. If the DAB theorist is committed to **Equation** and **Additivity**, they are refuted, unless they can somehow show the Peter case to be incoherent.

I'm going to argue that would be too quick. This case is not an argument against DAB; it is a reason to reject the conjunction of **Equation** and **Additivity**. The argument is that there are two coherent ways for someone broadly sympathetic to DAB to think about the Peter example. One of them rejects **Equation**; the other rejects **Additivity**.

Auntie E represents the first strategy. She is a thorough-going 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 Gx is the proposition that doing x is good.

Auntie E's Value $V(x) = Pr(Gx \mid X)$

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. This means that she has to (and does) give up **Equation**, since it implies that it is better for Peter to do *a*.

Auntie C endorses a version of causal decision theory, though as we'll see a more radical

one than the one supported by David Lewis (1981).[5] In particular, Auntie's values are given by **Auntie C's Value**. In the formula, $\Pr^X$ denotes the result of *imaging* the probability function Pr 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 1, never left or right.[6]

Auntie C's Value $V(x) = \Pr^X(Gx)$

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 onlooker hopes is maximised.[7] We'll come back to this point in Section 4. For now, I just want to note the formal similarities to Lewis's own theory.

Using this formula, Auntie C hopes that Peter does a. So she has to, and does, give up **Additivity**, since it implies that is better for Peter to not do a.

[5] Here I'm following Collins, who notes that it is odd that Lewis takes **Additivity** for granted, since it goes more naturally with evidential decision theory.

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

[7] Melissa Fusco (2026) notes that there is no Dutch Book argument against updating on a proposition by imaging iff that proposition is about one's own actions. If the probabilities relevant to desire are always post-update probabilities, that would be a problem, since X in this case is about Peter's actions not Auntie C's. So she's forced to insist that it is pre-update probabilities not post-update probabilities, that enter into the formula for desirability. This doesn't look incoherent, though it's an important complication.

The next step is to argue that both Aunties are really DAB theorists, so it isn't true that all DAB theorists have to accept **Equation** and **Additivity**.[8]

3 DAB and EDT

Let's start with 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.

[8] I actually hold something stronger, namely that no DAB theorist should accept both principles, but here I'll just argue that it isn't compulsory to accept both.

I'll offer some 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 DAB.

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

Lewis (1996) argues that this view is unHumean in an important way. Hume says that any desire can be rational; that's what we take from the quip that it is not contrary to reason to, for example, prefer the destruction of the world to a scratch? On Auntie E's view one cannot but desire D. This doesn't sound like Auntie E is respecting Hume's dictum. But note that Hume's example concerns desires concerning descriptive propositions. Auntie E's view puts no constraints whatsoever on descriptive desires. If she wants to desire

the destruction of the world she is free to; she just has to come to believe this is good. This might not be much of an imposition; having that belief might be constituted by the functional states that Lewis associates with the desire for the destruction.[9]

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 thinks **Additivity** obviously fails because in Peter's case; it makes sense to hope that he does the thing that's got a 59% chance of being good, not the thing that's got a 41% chance of being good. Isn't this, she thinks, just what a causal decision theorist like Lewis should have thought all along?

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

[9]There is much more to say here of course; the point here is to flag the possibility that Lewis's rejoinder to Price presupposes that Lewis and Price have a common theory of belief, and differ on how desire relates to it. Developing this defence of Auntie E might require rejecting this presupposition, but going further down this line would require a much more thorough engagement with Price's broader philosophy.

> 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 **Additivity** because of this confusion. To respond on Auntie C's behalf, let's break up four things Lewis might mean here.

1. There are multiple distinct concepts relating to pro-attitudes, including choice-worthiness, what we might call wishworthiness (like I wish I were a one-boxer), preference, and many others.
2. At least one of these notions is measured by evidential expected value.
3. *Desire* picks out one of those notions so measured.
4. The anti-Humean only succeeds if they show that this notion, *desire* goes with belief.

Clearly, 1 is true. Lewis can define 'Humean' how he likes, but it seems to me that there is a good sense in which 4 is just as clearly false. If any pro-attitude is conceptually tied to belief, that seems like an anti-Humean view.[10] Still, let's go along with Lewis for now; we can just take him to stipulate what 'Humean' means here.

[10] Note that Hume's famous quote about finger scratching concerns neither desire nor choiceworthiness, but preference.

Auntie C does not need to reject 2. She might reject it; I think there is some interest in a 'mad dog' causalism that rejects all use of evidential value. But arguing for that would take us far afield.[11] So let's assume that 2 is true, and that some evidential notion that is important in philosophical psychology. Still, Auntie C is justified in rejecting 3.

The reason is that desire seems closely tied to intention via principles of means-end coherence. In Newcomb's Problem, Lewis holds that V(one-box) is high but I should two-box. If V measures desire, this generates a puzzle about means-end coherence. I desire that I one-box; I believe I'll one-box just in case I choose to one-box; yet I don't choose to one-box. This seems incoherent; it violates the fundamental principle of instrumental rationality. Now maybe *desire* is ambiguous, and Lewis is right that there's a disambiguation that follows V, and which hence isn't relevant to means-end coherence. But all Auntie C needs is that some plausible notion of desire is relevant to means-end coherence. Once that's true, and she's a two-boxer, she has reason to think that **Additivity** doesn't apply to it.

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

[11] In particular, we'd have to respond to the many ways in which Jim Joyce (1999, 2026), no friend of evidential *decision* theory, has made use of evidential value.

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. I'll leave to another day which of the Auntie's has a better response; the point of this paper is that anyone who was sympathetic to DAB will have been sympathetic to one of the Auntie's, and hence will have a response to Lewis.

References

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.

Fusco, Melissa. 2026. "Imaging and the Diachronic Dutch Book." *Philosophy and Phenomenological Research* 112 (1): 243–75. https://doi.org/10.1111/phpr.70065.

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: Cambridge University Press.

———. 2026. “Choosing Among Unratifiable Acts.” Presented at the 2026 American Philosophical Association, Eastern Division.

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.

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.